Preface

This work is a study on the implementation in submicron and deep-submicron CMOS technologies of analog integrated circuits for the read-out of silicon sensors.

In the past few years detection techniques based on the use of silicon sensors have been playing an increasing role both in fundamental research and in applications. The photodiodes employed in the receivers of optical communication systems are the most common example, but silicon detectors have found a widespread use also in data acquisition systems for nuclear and high energy physics and as radiation detectors in general. The development of active pixel sensors (APD) is opening the perspective of building fully integrated imaging systems in standard CMOS processes, while micromachining techniques, introducing also mechanical capabilities, enlarge the domain of silicon sensors to further applications, like pressure measuring devices and accelerometers. Moreover, the fabrication of sensors in the same starting material used to build the processing electronics provides the opportunity of reaching a high level of integration, which is very important for portable or implantable systems.

The striking evolution of digital microelectronics has determined a progressive transfer of the information processing from the analog to the digital domain, moving the interface between the digital and the analog units as close as possible to the sensor side. The analog building blocks are therefore confined mainly to the periphery of the system, where they provide signal amplification, basic filtering and analog to digital conversion.

One of the main characteristics of CMOS technology, that presently dominates the market of digital integrated circuits, is its capability of building excellent and compact switches; on the other hand, very linear capacitors can be easily produced by adding few process steps. As a result, CMOS has gained more and more popularity also in the world of analog signal processing, where the switched capacitor technique has been playing an overwhelming role. However the evolution of CMOS technologies is dictated only by the needs of digital circuits, whose performance is improved by reducing the minimum feature size of the transistors. Present state-of-the-art CMOS processes have minimum gate length of 0.18 - 0.25 μm; at the same time, non submicron processes are being gradually phased-out.

The continuous scaling of the technology has two contrasting effects on the performance of analog circuits. In fact the quality of some parameters that are crucial for analog design (like threshold voltage dispersion and transconductance) improves; at the same time maximum power supply is reduced while the threshold voltages are not scaled in the same proportion, thereby squeezing the allowable signal swing and worsening the signal to noise ratio. This effect can be partially alleviated by using alternative circuit topologies, more suitable for low-voltage operation. Unfortunately, these architectures often require more area and dissipate more power than conventional ones. The use of conventional structures would give undoubtedly added value to the system and their limitations in scaled CMOS processes must be deeply investigated.

In this work we have focused our attention on two particular applications, namely:

- data acquisition systems for particle detectors used in high energy and nuclear physics
- front-end for very high-sensitivity photodetectors

We think however that the problems encountered in these areas are quite representative of the issues which arise designing sensors read-out electronics in submicron CMOS technologies.

The studies on the particle detector front-end have been carried out in collaboration with the Microelectronics group of the European Laboratory for Particle Physics (CERN) and have provided the first steps towards the design of a complex front-end architecture that will be used in one of the four big experiment presently under construction at CERN.

This thesis is divided into six chapters, which are shortly described hereafter.

Chapter1
This chapter provides an analysis of the major effects of the scaling of the MOS transistor on the design of analog integrated circuits. The considerations are carried out at the device level and are supported by experimental measurements on individual transistors implemented in two different quarter micron technologies. These measurements have been performed in the Microelectronics Group in CERN.

Chapter 2
After the preliminary work, we have aimed at verifying the results obtained at the device level with circuits designed for specific applications. This chapter describes one of the two applications (the front-end system for the Silicon Drift Detectors of the ALICE experiment) and details the system level considerations which lead to the choice of the architecture. The proposed solution requires the design of a full custom chip with 32 channels, each performing the amplification and the analog to digital conversion of the detector signals.

Chapter 3
The key component of the front-end system for the SDDs is a successive approximation converter. This chapter reviews this topology and outline the consequences of its implementation in a submicron CMOS technology.

Chapter 4
The high level of modularity of the SDD front-end chip requires the integration of many analog to digital converters on the same die. This chapter describes a first prototype, designed in a 0.7μm process and implementing sixteen converter on the same substrate.

Chapter 5

The analysis presented in chapter 1 suggest that the implementation of an analog circuit in a deep-submicron technology can result in better performance. This aspect has been investigated by designing and testing a 10 bit low power ADC, which is described in chapter 5. The choice of this circuit as a demonstrator stems of course from the fact that this converter would be a very suitable building block for the SDD system.

Chapter 6
This last chapter presents the design of the front-end for the photodetector, discussing in particular the possibility of a full integration of very high gain transimpedance amplifier.

Contents

1 Introduction: the scaling of the MOS transistor and its impact on analog design

The characteristics of CMOS technologies are tuned to optimise the performance of digital circuits. As a consequence, the minimum dimensions of the devices are continuously scaled, in order to increase the integration density and the speed and to reduce the power consumption.

The main side effect of the scaling is that the maximum allowed power supply is reduced, in order to avoid destructive electric fields inside the device, while the threshold voltages are not scaled in proportion to prevent off-state leakage. The available signal swing gets therefore smaller and smaller and this trend may seriously affect the performance of analog circuits. Some authors [1] have argued that in the near future analog and digital blocks could hardly be produced using the same process.

On the other hand, analog designs could benefit from the improvement of important figures of merit of the MOS transistors which are typical of the new submicron and deep submicron technologies. Additionally, alternative operating conditions of the MOS device, usually exploited only for niche applications, may have an increasing role and partially alleviate the problems arising from the squeezed power supplies.

This chapter presents an introductory analysis to the above aspects, providing the basic material to understand the design choices discussed in the following of this work.

1.1 MOS transistor scaling

As we have already anticipated, the reduction of the dimensions of the MOS transistor required also an adequate scaling of other physical parameters of the device. In fact, the source and drain are separated from the bulk by a reverse-biased *pn* junction; since the doping of the bulk is much lower than the doping of the electrodes the associated depletion region extends mainly underneath the gate. If the dimensions of the transistor are reduced, the two depletions regions become closer and closer and the short channel effects are exacerbated up to a point in which the device becomes unusable. A first remedy is then to reduce the maximum applicable voltage on the drain; the extension of the depletion region can be in fact approximated as[1]

$$d \simeq \sqrt{\frac{2\varepsilon_{Si}V}{qN_A}} \tag{1.1}$$

where

- ε_{Si} is the dielectric constant of the silicon
- q is the charge of the electron
- N_A is doping concentration in the bulk
- V is the reverse bias voltage between drain and bulk

[1] Whenever is necessary to specify the nature of the transistor we refer to NMOS transistor. However, the extension of the considerations of this chapter to the complementary device is immediate

Suppose now that both the minimum channel length and width of the device are scaled by a factor $\frac{1}{S}$; from eq. 1 we see that if the voltage V is scaled also by $\frac{1}{S}$ while the doping concentration is multiplied by a factor S, the size of the depletion region scales as $\frac{1}{S}$ as well, so that the ratio between the depletion width and the channel width is the same. The increase of the channel doping concentration rises the threshold voltage and to compensate for this drawback the gate oxide thickness must be reduced. This in turn, reduces the maximum voltage applicable to the gate.

In an ideal scaling procedure, also the threshold voltage should be scaled exactly by $\frac{1}{S}$; in this way, the *ratio* between the physical dimensions of the device stay the same and the transistor is just a down-sized copy of a bigger one and can therefore be described by the same characteristic equations. Since both the voltages and the physical dimensions are reduced in the same proportion, the electric fields inside the device don't scale and this approach is called "constant field scaling" [2].

The constant field scaling presents a couple of problems, that are particularly relevant for digital applications. According to the long channel theory, the drain-source current for a MOS transistor[2] is defined by

$$I_{DS} = \frac{KW}{2L}(V_{GS} - V_{TH})^2 \tag{1.2}$$

where V_{TH} is the threshold voltage and $K = \mu C_{ox}$, (with μ mobility of the carrier and C_{ox} gate capacitance per unit area). Eq. 2 suggests that the device turns off abruptly when $V_{GS} = V_{TH}$, but this is not the case and the current goes to zero smoothly while $V_{GS} - V_{TH}$ becomes more and more negative.

In the *subthreshold* region, the drain-source current is more realistically described by

$$I_{DS} = I_{D0}\frac{W}{L}e^{\frac{V_{GS}-V_M}{nV_t}} \tag{1.3}$$

where V_t is the thermal voltage $\frac{kT}{q}$ and n is parameter, defined by

$$n = 1 + \frac{\gamma}{2\sqrt{2\phi_F + V_{SB}}} \tag{1.4}$$

with $\gamma = \frac{\sqrt{2\varepsilon_{Si}N_A}}{C_{ox}}$ and ϕ_F the Fermi level. I_{D0} is a parameter depending on the technology and V_M is the upper limit of the weak inversion region. From eq. 3 we see that the bigger n, the more difficult is to switch off the transistor. The value of n can be assumed to be between 1.3 and 1.4 in many different technologies [3].

In a constant field scaling, both N_A and C_{ox} are scaled by S and hence γ scales by $\frac{1}{\sqrt{S}}$. Since all voltages are supposed to scale by $\frac{1}{S}$, n does not change. As a consequence, the weak inversion region is *relatively* larger and a big fraction of the available voltage swing must be used just to turn on and off the transistor, thereby reducing the noise immunity in digital circuits. Another important aspect is that the scaling of the power supplies makes difficult the compatibility between technologies of subsequent generations. Therefore, for digital design, a *constant* voltage scaling would be preferable; since this is not feasible, what practically results is a compromise between a constant voltage scaling and a constant field scaling, so that the physical dimensions and the oxide thickness are scaled by $\frac{1}{S}$, whereas the threshold voltages and the other parameters are scaled by $\frac{1}{S'}$, with $1 < S' < S$.

[2] For simplicity and without loss of generality we assume that the device works in the saturation region

1.2 Impact of transistors scaling on analog design

1.2.1 Transconductance

The transconductance is the most important parameter of the transistor when it is used for analog applications. The transconductance g_m of a long channel MOS device is defined by

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} = \frac{KW}{L}(V_{GS} - V_{TH}) = \sqrt{2K\frac{W}{L}I_{DS}} \tag{1.5}$$

The above expressions outline the fact that g_m depends on a *bias* parameter (the drain-source current or, equivalently, the overdrive voltage $V_{GS} - V_{TH}$), on a *design* parameter ($\frac{W}{L}$) and on a *technological* parameter $K = \mu C_{ox}$.

Actually, it must be observed that the eq. 5 is over-simplified, since it neglects the bulk effect which, increasing the threshold voltage, reduces the overdrive voltage and hence the g_m. A more accurate expression is [4]

$$g_m = \sqrt{\frac{2KW}{Ln}I_{DS}} \tag{1.6}$$

where n has been previously defined. Since $C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$, (where t_{ox} is the gate oxide thickness), the parameter K gets bigger with the scaling of the process. As a comparison, table 1 reports the values of the parameter K for different technologies from 1.2 μm down to 0.25 μm

Table 1.1: Technological parameter K for different CMOS technologies (NMOS transistors)

$L_{min}(\mu m)$	$t_{ox}(nm)$	$K(\frac{\mu A}{V^2})$
1.2	24	68
0.8	14	90
0.5	10	134
0.25	5	280

As it is seen from this table, the technological parameters improves significantly by reducing the oxide thickness; therefore a better g_m is obtained for the *same* design parameters and the *the same* bias conditions, implying that better performances can be attained with less power. It is necessary to note that, if the transistor is biased in strong inversion, the scaling of the L is not so useful to increase the g_m; in fact, to keep the transitor is saturation, which is the normal operation region for analog applications, a voltage $V_{DS} > V_{GS} - V_{TH}$ must be applied between the drain and source terminal of the device. If the length is reduced, the electric field is increased and so does the carrier velocity up to a point in which it reaches a saturation value v_{sat} [2]. Therefore, the increase in the g_m is less than what predicted by eq. 5 and for short channels the transconductance is better described by

$$g_m = WC_{ox}v_{sat} \tag{1.7}$$

and does not improve anymore by reducing the gate length.

An interesting parameter to discuss is the ratio between the transconductance and the current which is used to generate it, since it is a measure of the power efficiency of the device. In strong inversion this parameter is

$$\frac{g_m}{I_{DS}} = \sqrt{\frac{2KW}{nLI_{DS}}} \tag{1.8}$$

while for the weak inversion the following relation holds

$$\frac{g_m}{I_{DS}} = \frac{1}{nVt} \tag{1.9}$$

Therefore, the transconductance-to-current ratio is constant in weak inversion, while it is inversely proportional to $\sqrt{I_{DS}}$ in strong inversion. In the weak inversion region the MOS transistor exhibits a "bipolar-like" behaviour, which makes the g_m proportional to the current and maximises the power efficiency.

The boundary between strong and weak inversion is easily found imposing a continuity between the two regions, which if of course physically justified. Equating eq. 8 and eq. 9 yields

$$I_{DSWS} = 2nK\frac{W}{L}V_t{}^2 \tag{1.10}$$

Since eq. 10 depends *linearly* on K, we can conclude that scaling the technology makes it easier to work in the weak inversion region, which should be preferred for analog designs. In fact, we have just seen that in weak inversion the g_m for a given current is bigger; to complete our discussion we have to mention that eq. 3 holds in saturation, while in the liner region it should be modified as

$$I_{DS} = I_{D0}e^{\frac{V_{GS}-V_M}{nV_t}}(1-e^{-\frac{V_{DS}}{nV_t}}) \tag{1.11}$$

As in strong inversion, the saturation occurs when the current modulation due to the drain source voltage becomes negligible; however, in this case a V_{DS} of 200 mV and *independent* from the current flowing in the device is sufficient to make the term containing V_{DS} negligible and the saturation condition is easily achieved. This implies that even with low power supply is still possible to stack the transistors in cascode configurations, which are extremely useful in analog design. As an example of the expansion of the weak inversion region in submicron technologies, fig. 1 show a measured $\frac{g_m}{I_{DS}}$ curve on a transistor implemented in a 0.25 μm process[3] [5].

1.2.2 Noise

It is well known that the channel of a MOS transistor is a source of thermal noise [6], whose spectral density can be referred to the input of the device in the form

$$S_n = \frac{4nkT}{gm}F\Gamma \tag{1.12}$$

[3]For completeness, we have to say that this particular device was a big device, (W/L=2000/0.78) conceived for noise measurements. However it must be noted that the strong inversion region is reached only for bias current of the order of the mA

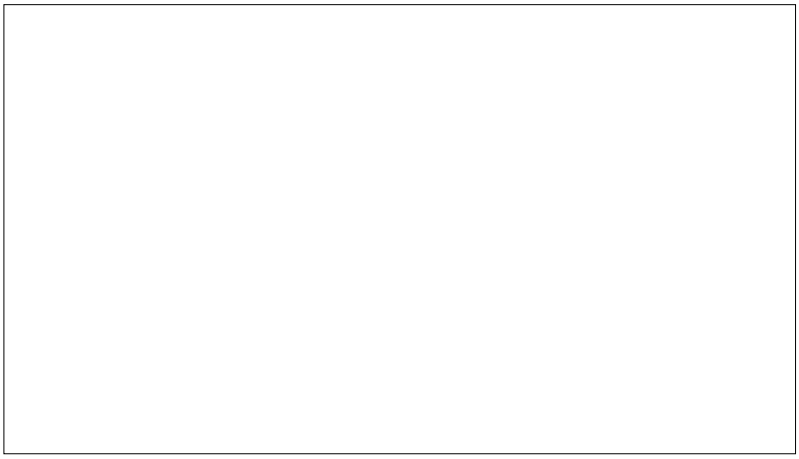

Figure 1.1: Typical $\frac{g_m}{I_{DS}}$ curve of a deep submicron device

where F is a coefficient taking into account the inversion condition of the device and is $\frac{1}{2}$ in weak inversion and $\frac{2}{3}$ in strong inversion. Γ, called "the excess noise factor" is an empirical term introduced to correct for the discrepancies between the elementary theory and the measurements; of course, the closer Γ is to 1, the better is the device. In non submicron technologies Γ is normally between 1 and 1.5 [6].

Moving to submicron technologies, the g_m tends to increase and it becomes easier to work in weak inversion, which, for the *same* current, gives intrinsically a smaller noise. Therefore, a reduction of the thermal noise is expected.

The second noise source found in MOS devices is the flicker noise or $\frac{1}{f}$ noise which has an input-referred spectral density defined by

$$S_n = \frac{K_f}{{C_{ox}}^2 WL}\frac{1}{f} \tag{1.13}$$

where K_f is a constant dependent on the technology. We see from this expression that for the *same* silicon area, the flicker noise decreases due to the increase in the gate oxide capacitance.

Deep submicron technologies offer therefore th opportunity of a better noise figure; however the evolution of the excess noise factor and of the flicker noise coefficient is an issue, because a worsening in these parameters could waste the improvements due to the technological aspects discussed above. For any reliable low-noise design data on these parameters are necessary and if they are not available from the factory, a preliminary qualification should be done by the user.

Table 2 shows the results of the measurements of the excess noise factors done in CERN on two quarter micron technologies.

Table 1.2: Excess white noise factors

	tech_a	tech_b
W. .I	$\Gamma = 1.1$	$\Gamma = 1.1$
M. I.	$\gamma = 0.85$	$\gamma = 0.65$
S. I.	$4 < \Gamma < 5.5$	$2 < \Gamma < 3.4$

For the moderate inversion we have use the expression $S_n = \frac{4nkT\gamma}{g_m}$, since in this region the inversion coefficient F is not well defined[4] and it is customary to include it in the excess noise factor.

We see from table 2 that the excess noise factor is very close to its ideal value ($\Gamma = 1$) in weak and moderate inversion, whereas it is bigger in strong inversion. These data are consistent with other previously reported in the literature for deep submicron technologies; the bigger values of the excess noise factors in strong inversion are explained [2] with typical short channels effects, like hot carriers degradation and weak avalanche phenomena. Actually, we could measured only devices with a channel length ranging from the minimum to a maximum of 1 μm.

The highest value of the flicker noise coefficient we have measured is 6 x 10^{-28} $\frac{C^2}{m^2}$ for PMOS transistors and 3 x 10^{-27} $\frac{C^2}{m^2}$ for NMOS transistors, which are not significantly different (though slightly lower) from the values found in the literature for non submicron technologies [6, 7]

1.2.3 Matching

Many analog circuits (like differential amplifiers, current mirrors and charge redistribution converters, to name only a few) rely on the fact that their are built with *identical* devices. Any deviation from the parameters of matched devices from their ideal ratio is than a source of error a limit the performance of the circuit.

The systematic mismatches can be partially compensated with suitable layout topologies (for instance common centroid geometries alleviate the impact of gradients along the circuit), while *random* mismatches are more difficult to deal with.

In MOS transistors random mismatches affect both the threshold voltage V_{TH} and the current gain factor $\beta = k\frac{W}{L}$ and can be studied with a statistical approach [8]. Therefore they are usually characterised with the standard deviation from their mean value. The threshold voltage mismatch is well described by

$$\sigma_{V_{TH}} = \frac{Bt_{ox}}{\sqrt{WL}} \tag{1.14}$$

where B is a parameter which is found to be 1 $mV\frac{\mu m}{nm}$ in many different CMOS processes. Hence an improvement in the threshold matching for the *same* silicon area is expected if the transistor is implemented in a submicron process.

The mismatch in the current gain factor is defined by

$$\sigma_\beta = \frac{A_\beta}{\sqrt{WL}} \tag{1.15}$$

Values found in the literature for A_β ranges from 1 to 3%μm

Measurements on matching are quite cumbersome, since many devices have to be measured in order to get a reliable statistics. In the early stage of this work in the Microelectronics group in CERN some studies have been carried out in a commercial quarte micron process; 100 chips were measured, each containing 5 couples of transistors we different gate area. The value found for B was 1 $mV\frac{\mu m}{nm}$, giving a $\sigma_{V_{TH}}$ of 5.5mV (the oxide thickness was 5.5 nm). For σ_β we found a value of 1.5%μm. The values found are consistent with the ones reported by other authors [1] and are fully compatible with the needs of precise analog design.

[4] In moderate inversion, F should be between 1/2 and 2/3, but the exact value is not easily predictable

1.3 Remarks on mixed-mode designs

So far we have been concerned with properties of single devices and we have considered only pure analog aspects. However, the success of CMOS technologies is due also to their capability of integrating both digital and analog functions on the same substrate.

One of the main problem on mixed-mode integrated circuits is due to the coupling of the digital noise with the analog parts. In submicron technologies the devices are usually built in a relatively thin (few microns) high resistivity layer that is epitaxially grown on highly conductive bulk. This bulk provides a medium by which the switching noise of the logic parts can easily propagate to the analog ones.

The best way to overcome this problem is to physically glue the backside of the chip to a ground plane. This solution is not alway practical, because the backside is always passivated an a thinning-metallisation procedure would be required. Therefore, this issue must be properly addressed at the design level and three main solutions have emerged over the years [9]:

1. Avoiding that digital and analog parts share any power and grounding bus; this is easier achieved in a submicron technology, since newer processes offer an higher number of layers for the interconnects (up to seven in a 0.25 μm process)

2. Avoiding logic design style which entail a large switching noise; as a consequence, for optimum performance, current steering logics have to be preferred to the more conventional CMOS families

3. Using as much as possible fully differential architectures in the analog parts.

The second solution is particularly cumbersome, because logic design is usually carried-out using standard cells available from the silicon foundry. The design of a customised library is a long process, since the cells have not only to be simulated and laid-out, but also integrated in a design environment suitable for the use of automatic synthesis tools. Therefore, this solution is not very practical and is normally used only when very high performances must be achieved.

1.4 Summary

The evolution of CMOS technology, though optimised for digital applications, presents some interesting aspects also for analog design. In the analog domain, however, the benefits do not come from the scaling, but from the fact that, while keeping constant the width and length of the transistors, important analog parameters like matching and, in some conditions, noise, improve. These improvements are traced mainly to the fact the the oxide thickness is scaled, affecting directly some basic characteristics (for instance the $\frac{1}{f}$ noise and threshold voltage mismatch of the devices).

Additionally, the limit between weak and strong inversion moves towards higher current densities and the weak inversion region, which in principle is preferable for analog design, can be better exploited. Moreover, the extensive use of this region of operation can preserve the use of standard architectures also with lower power supplies.

The analog design takes therefore an indirect advantage from the reduction of the minimum feature size, while the use of actual minimum geometry in analog circuits, exacerbating the small-channel effects, leads to a substantial degradation in performance.

2 Design of the front-end system for the Silicon Drift Detectors of the ALICE experiment

In the past fifteen years silicon sensors have found a very wide application in the domains of nuclear and particle physics. Their linear energy response, combined with their segmentability in almost any shape makes silicon detectors excellent for both charge and position measurements. Thanks to the advancements in VLSI technologies the electronics necessary for the signal processing can be implemented in a monolithic form, making possible the construction and the reliable operation of systems with millions of channels.

The design of such electronics is nevertheless a challenge and demands intensive trade-off between conflicting requirements. The optimisation procedure is usually driven by tight constraints on space and power consumption, which considerably reduce the degree of freedom in the choice of the architectures. Moreover, many experiments operate with levels of radiation that are bigger than what a standard process can afford. As in the design of any complex apparatus, a clear definition of the system specifications is of paramount importance, since both under-design and over-design likely result in final poor system performance.

In this chapter, the above items are outlined discussing the particular example of the front-end system for the Silicon Drift Detectors in the ALICE experiment. In the first part of the chapter a brief overview of the LHC project and of the ALICE experiment is given. The second part introduces the Silicon Drift Detectors and in the third part the architecture of the proposed read-out system is discussed.

2.1 CERN and the LHC program

CERN, the European Laboratory for Particle Physics, is an international organisation founded in 1954 and supported by 20 European countries. Its aim it to provide advanced research facilities to investigate the basic constituents of matter and the forces which rule their behaviour.

The study of nature at very small scales requires very high energy. This can be viewed as a consequence of the De Broglie relation ($\lambda = h/p$, where λ is the Planck constant), which states that the wavelength associated with a particle is inversely proportional to its momentum. In a typical high energy physics experiment, a beam of charged particles is brought to the required energy by mean of a particle accelerator and is made to collide with a fixed target or with another particle beam. A complex system of sensors and electronic modules (usually referred to as “detector”) is laid-out around the interaction point to study the products of the collisions. Accelerators and detectors are therefore the basic tools required by any experiment. Machines which can accelerate and collide two counter-rotating beams are usually preferred to fixed target accelerators, because the energy available for the collisions is higher.

CERN has a very large system of particle accelerators; the biggest machine presently operating in the laboratory is LEP, an electron-positron collider that is located in an underground tunnel with a circumference of 27 km. With this apparatus a total energy of 200 GeV can be attained. LEP has been operating since 1989 and during this decade has provided very pre-

cise tests of the standard model, which is widely accepted to interpret the phenomenology of elementary particles.

However, the accelerators working nowadays in CERN and in other similar laboratories around the world are not powerful enough to investigate important aspects of the standard model. Increasing collision energy (and hence probing smaller scales) opens moreover the possibility of new discoveries which could modify or even drastically change our present description of the fundamental laws of nature.

These motivations led CERN to the project of a new machine, called Large Hadron Collider (LHC), that will accelerate two beams of protons up to an energy of 7 TeV per particle. This new accelerator will start operating in 2005. A set of four detectors (ATLAS, CMS, ALICE and LHCB) is under construction to exploit the research opportunities offered by the LHC.

2.1.1 The ALICE experiment

The LHC machine will be able to accelerate not only protons, but also heavy nuclei to an energy of 5.5 TeV/nucleon. The collision between relativistic heavy nuclei is an excellent mean to investigate the behaviour of matter at extremely high densities, similar to the ones which should exist in nature just after the Big Bang and which may exist today in the core of collapsed astrophysical objects. This field attracts therefore the interest not only of particle physicists, but also of nuclear physicists and astrophysicists.

According to the standard model, the nucleons (protons and neutrons) are made of quarks, which interact by exchanging gluons. In the ordinary nuclear matter, quarks can not exist as free particles, but only in bound states called hadrons. The interactions between quarks are described by a theory called quantum cromodynamics (QCD). QCD has been built by analogy with the quantum electrodynamics (QED) which explains the weak and the electromagnetic forces. QCD is not tested to the same extent as QED; nevertheless it is in agreement with an enormous amount of experimental data and is not contradicted by any known experiment.

Theoretical calculations based on QCD indicate that at an energy density of 1.3 GeV/fm^3 the hadrons should "melt" to form a plasma of free quarks and gluons. By colliding beams of lead ions in the LHC, it will be possible to reach an energy density of 27 GeV/fm^3, which is well beyond the limit of the envisaged phase transition. Heavy ion collisions are therefore a mean of testing the predictions of QCD. Moreover, the large amount of particles involved in each event will make possible the study of strong interactions on a statistical basis (QCD thermodynamics). A detector dedicated to the study of heavy-ion collisions is therefore an important part of the LHC research program. This detector (called ALICE, acronymous for A Large Ion Collider Experiment) is presently being designed and should be ready at the start-up of the LHC. A pictorial view of ALICE is given in fig.1.

2.2 The Silicon Drift Detectors

In order to study the processes of interest, in a high energy physics experiment it is necessary to measure the properties of the particles produced in the collisions and to reconstruct their tracks. No single detector is optimal to perform all these measurements and a modern apparatus is composed of several sub-systems.

The ALICE experiment, as can be seen from fig.1, is formed by six sub-detectors. Among these, the Inner Tracking System (ITS) and the Time Projection Chamber (TPC) are used to

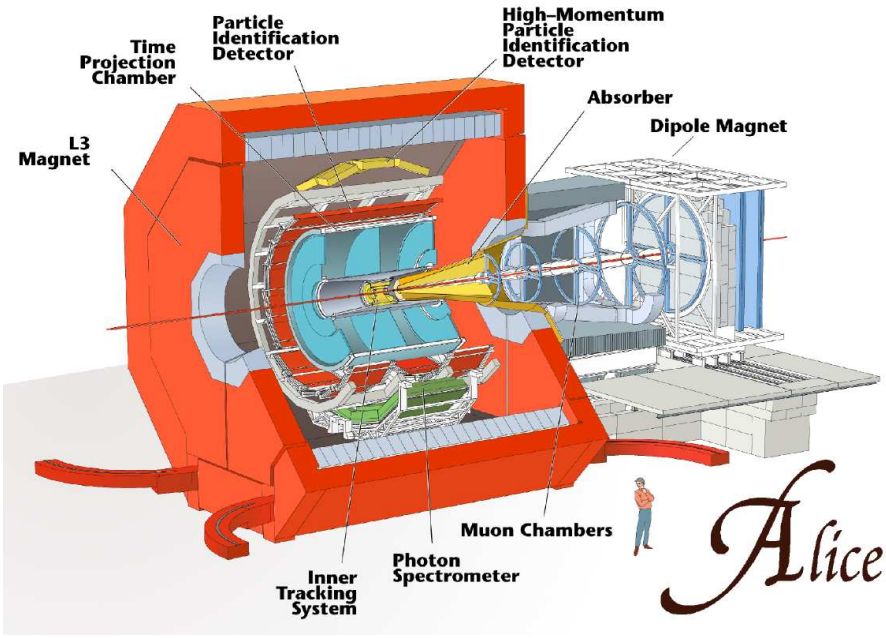

Figure 2.1: The ALICE detector

reconstruct the tracks of the particles and to measure their energies.

For the following of this discussion, it is sufficient to say that in heavy ion collisions, the number of produced particles is very high. Computer simulations indicate that a density of 90 particles /cm^{-2} can be reached in the inner parts of the ALICE detector. This large density requires in the ITS the use of sensors with very high granularity and good spatial resolution and the silicon detectors (pixels, microstrips, silicon drift chambers) provide the best performances from this respect. In silicon detectors, a charged particles is detected by the ionization it causes in a fully depleted pn junction. We concentrated here only on drift detectors, which are directly related to the work presented in this chapter.

A Silicon Drift Detector [10](SDD in the following) is obtained by implanting an array of p^+ strips on both side of a n^- doped bulk, as shown in fig.2. The pn junctions formed by the p^+ strips and the n^- bulk are reversed biased to fully deplete the detector from free carriers. The voltage is scaled along the p^+ arrays, so that a drift field parallel to the detector surface results. When a charged particle crosses the detector, produces electron-hole pairs by ionization. The holes are collected by the nearest p^+ strips and don't contribute to the signal. Since a negative bias is applied to both sides of the detector, the electrons are focused in the middle of the structure and drift towards the n^+ anodes, which are positively biased. The maximum drift path is typically of the order of the cm.

During the drift process, the electron cloud spreads because of diffusion and electrostatic repulsion and the charge is usually collected by more than one anode. In the rime domain the signals can be described as current pulses of Gaussian shape, whose amplitude and standard deviation depend on the impact point of the ionizing particle and whose integral over the interested anodes is the total charge released in the detector by the interaction. Therefore if the amplitude of the signals is recorded by performing analog read-out, the impact coordinate in the z direction can be determined by calculating the centroid of the charge distribution. The resulting accuracy is much higher than the distance between the anodes and with SDD having

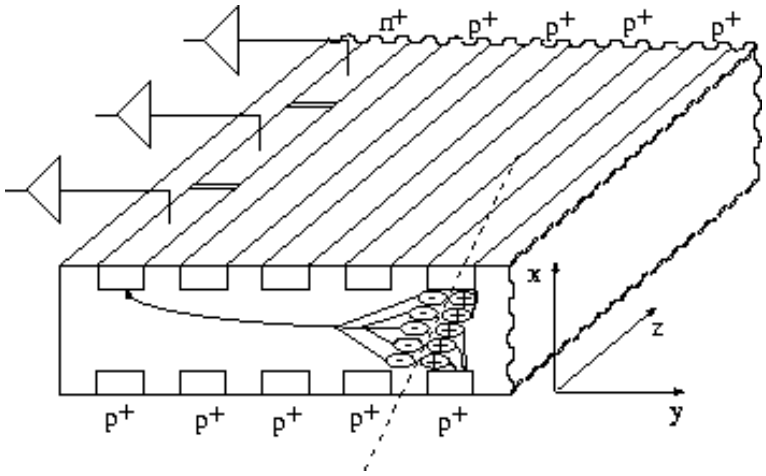

Figure 2.2: Schematic representation of a Silicon Drift Detector

an anode pitch of 200 μm a resolution of 30 μm has been obtained.

The coordinate in the y direction is deduced by measuring the drift time of the electrons. SDD are therefore bidimensional detectors. To measure the drift time it is necessary to provide a time reference. This is done by another detector in the system. In ALICE the time reference for the drift detectors will be derived from the trigger signal.

The velocity of the electrons in the semiconductor is related to the mobility μ and to the electric field E by the relation $v = \mu E$. A typical value for the electric field used in SDD is 500 V/cm, which yields a drift speed of about 8 μm/ns. Unfortunately, the mobility is a sensitive function of the temperature ($\mu \propto T^{-2.4}$) and to get a good position accuracy the temperature should be very well stabilised. A possible alternative is to calibrate the drift speed by injecting periodically a signal in a fixed point of the detector [11]. This can be accomplished by MOS charge injectors, as shown in fig. 3. The positive fixed charge at the interface between Si and SiO_2 creates a potential well in which electrons accumulate. Applying a negative voltage to a gate terminal, these electrons are injected in the drift region and are then collected at the anodes. The potential well is refilled by thermal generation processes in the vicinity of the MOS injectors. The refilling rate is small and therefore the injectors give a significant signal

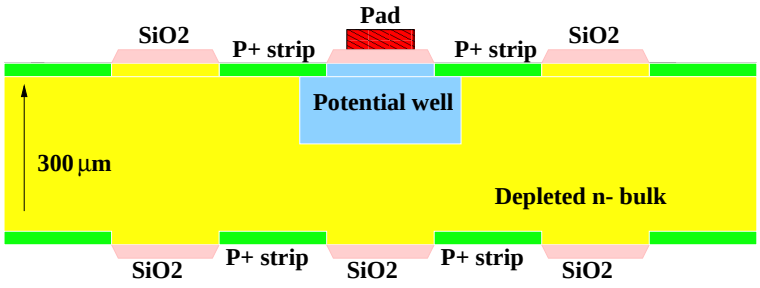

Figure 2.3: Scheme of the charge injection principle

only if the frequency of the applied negative pulse is slow (typically below 1 *kHz*). This is not a major limitation, since temperature variations are slow as well. The effectiveness of this method has been recently demonstrated. By calibrating the drift velocity with MOS injectors,

experimenters were able to get an accuracy position equivalent to a temperature stabilisation of 0.1 K ($\simeq$ 30 μm) [19, 20]. Silicon Drift Detectors can achieve in both coordinates a resolution of 30 μm, which is comparable to the resolutions of pixel detectors. Their main advantage with respect to pixels is the reduced number of electronic channel required to process the information from the detector. For instance, to read-out a SDD area of 7×7 cm^2 512 electronic channel are required. To read-out a pixel sensor with the same area with comparable resolution about 300.000 channels are needed.[1] The low number of electronic channels required make possible to use analog read-out. This, besides providing accurate position measurement, has another benefit because in silicon detectors the charge released is directly proportional to the energy lost by the particle in the medium. As we shall see in the next paragraph, the measurement of the specific energy loss of a ionizing particle is a powerful tool to identify the particle itself. Therefore, SDD can be used both for tracking and particle identification purposes.

The main drawback of SDD is the speed, because it is necessary to wait for the charge to drift. The drift time depends on the size of the detector and the applied field and typically ranges from 4 to 6 μs. Another disadvantage of SDD compared to strips or pixels is that a single defect in a specific point can compromise the functionality of the whole detector. As a consequence, silicon wafers of very good quality and very clean processing are required to produce SDD with a reasonable yield.

In order to minimise the connections between the detector and the outside world, which is very important in large systems, it is desirable to integrate on the detector also the high voltage divider which biases the cathode strips. This structure has to be very linear, in order to maintain a constant drift field inside the detector. The value of the resistors in the voltage divider must be chosen as a compromise between two opposite needs. On one hand, in fact, the power dissipation of the voltage divider should be minimised, in order to keep thermal gradients on the detector as small as possible. On the other hand, local defects may generate leakage current and the hole component of this current enters the voltage divider circuitry. If this current is comparable to the one flowing in the divider, it can alter the voltage at the cathode strips, introducing nonlinearities in the drift speed. Therefore, the value of the resistors can not be too high, so that the effect of the leakage current is negligible. A typical value used for the individual resistors in the high-voltage divider is 100 kΩ. Significant progresses have recently been reported in this domain [20].

Fig. 4 is a picture of a SDD detector similar to the one which will be used in ALICE. The high voltage is applied in the middle of the detector and two independent drift paths are created on both sides of the device. The charge is collected by 256 anodes on each side, with a pitch of 300 μm. The size of the detector (referred to the middle) is 87.6×77.9 mm^2. The hexagonal shape is dictated by the need of minimising dead zones when the detectors are placed next to each other to cover a large area.

In the following of this chapter, we will discussed the specifications for the front-end electronics of the SDD and will present the architecture proposed to meet these requirements.

[1] To be fair, the comparison should also consider the area and power required by the front-end electronics. To read -out a pixel cell with a state-of the art front-end requires an area of15000 μm^2 and a power of 60 μW. To read-out the same surface of a SDD is needed an average electronic area of 2000μm^2 and a power of 13 μW

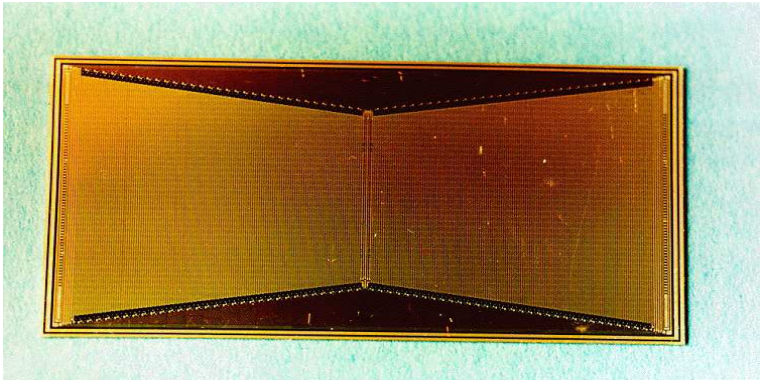

Figure 2.4: Picture of a Silicon Drift Detector

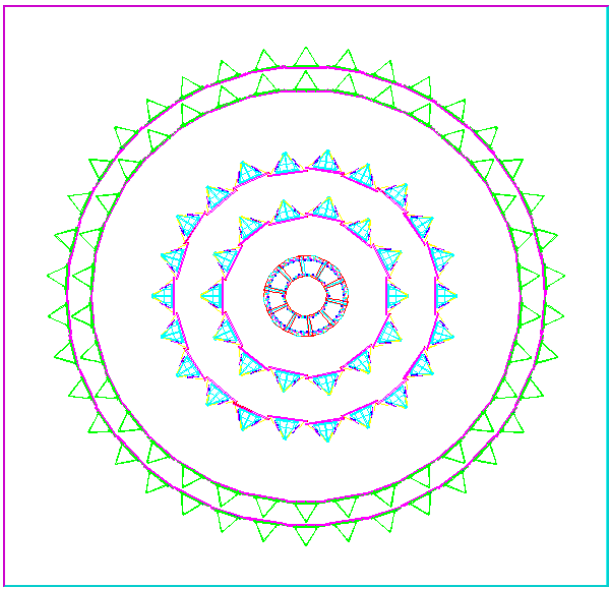

Figure 2.5: Front view of the Inner Tracking System (ITS)

2.3 Specifications of the front-end for the ALICE SDD

2.3.1 System overview

The design specifications of a detector are set by the physical goals that must be achieved. The main constraints can be understood with simple arguments, but the final requirements are the outcome of computer simulations which often need customised software packages. To optimise the performances of the SDD front-end, system-level studies have been carried out describing the detector and the signal processing units in a C^{++} code. The results of these simulations will be used in this paragraph to illustrate the specifications of the front-end electronics.

We have already seen that in the inner parts of ALICE the high density of particles demands the use of tracking detectors with very good spatial resolution. Six layers of silicon detectors (two of pixels, two of SDD and two of microstrips) will be arranged in a cylindrical structure, as shown in fig. 5 [21]. The radius of the innermost layer is 4 cm, the one of the outermost layer

is 43.6 cm and the total sensitive area is 6.74 m^2.

The ITS has a twofold purpose: to identify the particles with low momentum, which do not reach the outer detectors and to reconstruct the tracks of all the particles in the internal part of the apparatus [21].

In fig. 6 the specific energy loss of some typical particles is plotted versus the particles momenta. The y axis is in mip units. (One mip is the amount of charge released by a particle with minimum ionizing power crossing 300 μm of silicon and corresponds to 25000 electron-hole pairs). It is apparent from this graph that at low momenta, the energy loss is quite different from particle to particle and hence can be used for identification purposes.

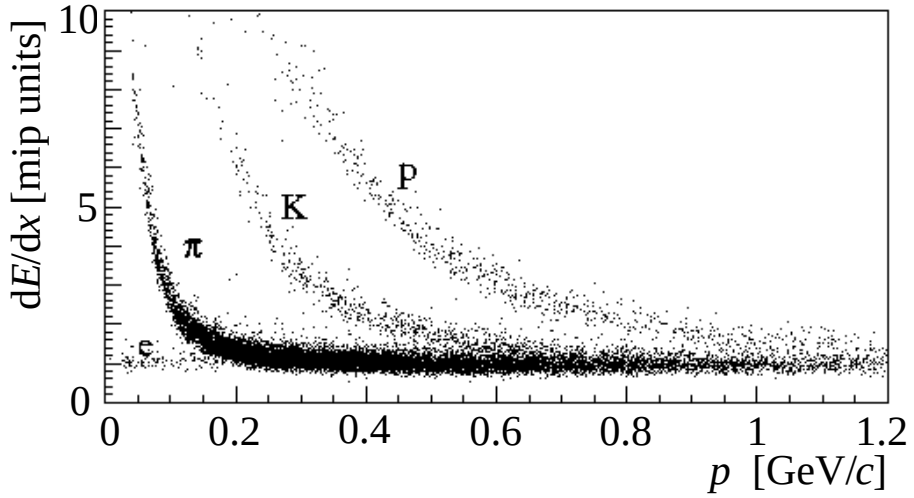

Figure 2.6: Example of specific energy loss as a function of the particles momenta

To measure the energy loss, the amount of charge released by the particles must be known and therefore it is necessary to perform analog read-out of the detectors. In the ITS this is done in the SDD and the strip layers.

Regarding the tracking reconstruction, it is important to remember that a particle traversing a medium can be scattered by the collision with the nuclei. This introduces a perturbation in the particle trajectory which limits the accuracy of the measure. In ALICE the same particle has to be observed by different detectors and is therefore important to cut down the scattering introduced by each layer of sensors. To reach this goal the amount of material in the sensitive volume must be reduced as much as possible, which implies that the electronics and the mechanical infrastructures near the detector have to be minimised. This has three major consequences on the design of the front-end electronics:

1. The space occupied by the front-end chips has to be small
2. The amount of cables necessary to connect the system to the outside world must be limited
3. The power dissipated by the electronics has to be low. The detector is in fact very sensitive to temperature variations and a low power consumption is mandatory to have an efficient thermal stabilisation with a minimum of refrigerant.

In the final set-up, the SDD will be arranged on linear frameworks called ladders, as shown in fig. 7. The ladders provide the mechanical support for the detectors and the electronics

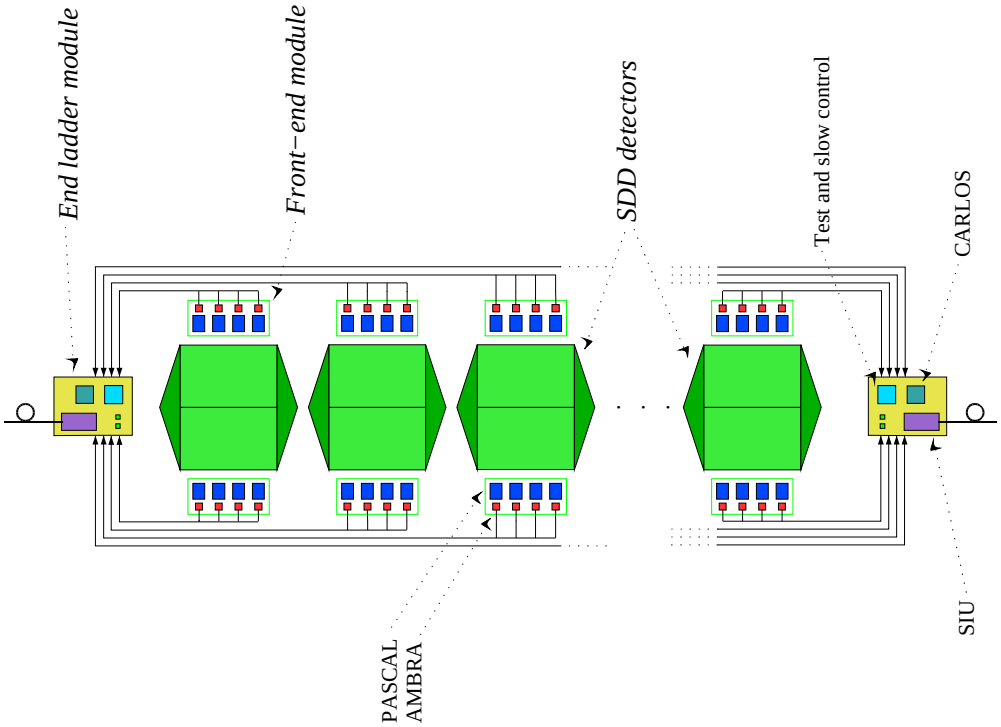

Figure 2.7: The SDD ladder architecture

board and they are placed side by side to form cylindrical structures (barrels). In the ITS there will be two barrels of SDD, hosting respectively 14 and 22 ladders (see fig. 8) The ladders in the internal barrel host 6 SDDs each, the ones in the external barrel 8. The SDDs have a bidirectional structure, with the high voltage applied in the middle and one array of 256 anodes on both sides. The electronics necessary for reading half a detector should fit in a board of 8×2 cm^2 and dissipate at most 5mW/channel. At the and of each ladder, another printed circuit accommodates the chips for the interface with the data acquisition system.

2.3.2 Signal dynamic range

The lower edge of the dynamic range is determined by the noise of the electronics, which depends on the parasitic capacitance seen from the input and the detector leakage current. The minimisation of the noise is constrained by the power budget allocated to the preamplifier and the required bandwidth. On the basis of system simulations, a target value of 250 e$^-$ rms has been determined. This value ensures a good resolution on the smallest signals of interest, occurring when a minimum ionizing particle hits the detector far from the anodes. The charge deposited is 4 fC, but since it is collected by more electrodes, a charge down to 1 fC must be measured with sufficient accuracy.

The high end of the dynamic range is fixed by the charge deposited near the anodes by the most ionizing particles of interest. These are defined as the particles whose tracks are reconstructed by the whole ALICE detector with at least 50% efficiency. As an example, fig.7 shows the combined tracking efficiency of the ITS and of the TPC for pions and protons. For the same kind of particles, in fig. 8 the limits of the dynamic range which allow 60 % and 90 %

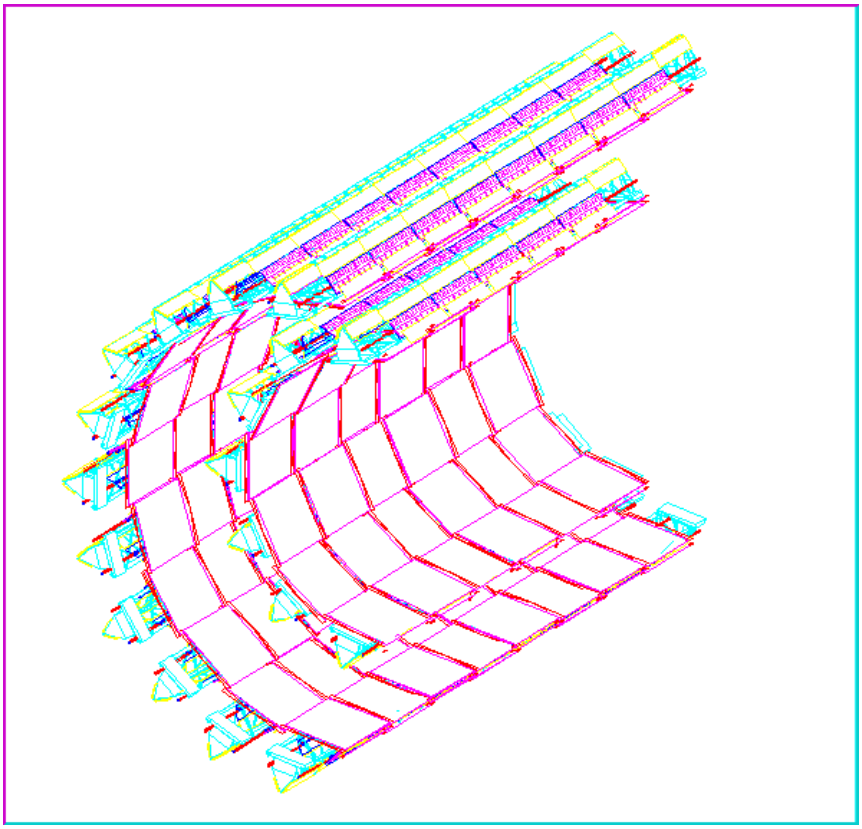

Figure 2.8: View of one half of the SDD barrels

of non saturated signals are plotted versus momentum.

From fig. 7 we see that the probability to reconstruct the tracks of pions is higher than 50% if they have momenta bigger than 160 MeV/c. For protons, a minimum momentum of 260 MeV/c is required for a 50% efficiency. From fig. 8 it is apparent that the maximum signal generated by pions having momenta higher than 160 MeV/c will lay under the dotted line. The maximum signal generated by protons with momenta of 260 MeV/c will be slightly above. In most cases, anyway, the charge will be collected by several anodes. The value of 200.000 e^- has therefore been chosen as the upper limit of the dynamic range.

In the time domain, the useful signals can be modelled as Gaussian current pulses with an amplitude between 13 nA and 1.3 μA and σ ranging from 10 ns to 30 ns and must be processed by the front-end electronics without saturating.

2.3.3 Amplifier bandwidth

The first element in a front-end system has to amplify the signal of the detector and to furnish it in appropriate form to the following stages. We suppose that the output of the amplifier is a voltage, since the signal processing we need can be easier carried out in the voltage domain. The amplifier has therefore to perform a current to voltage conversion, i.e. a transimpedance function.

The bandwidth of the amplifier is selected as a compromise between two opposite needs. In fact, on one side, the minimization of the noise calls for a small bandwidth. On the other, because of the high density there is a significant probability of having overlapping signals and

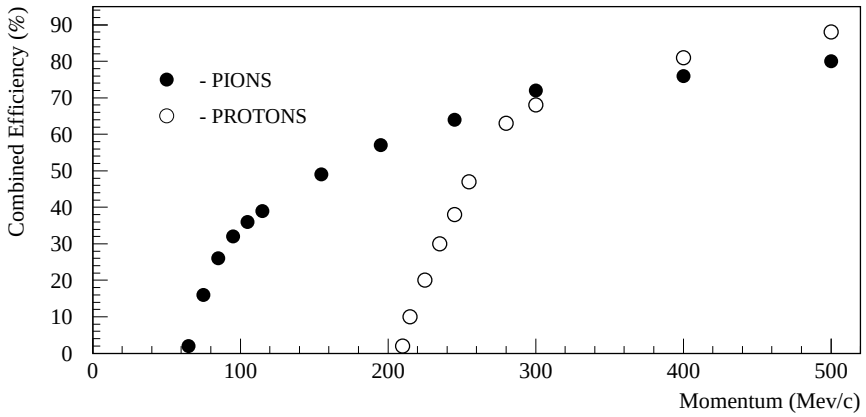

Figure 2.9: Combined ITS-TPC efficiency

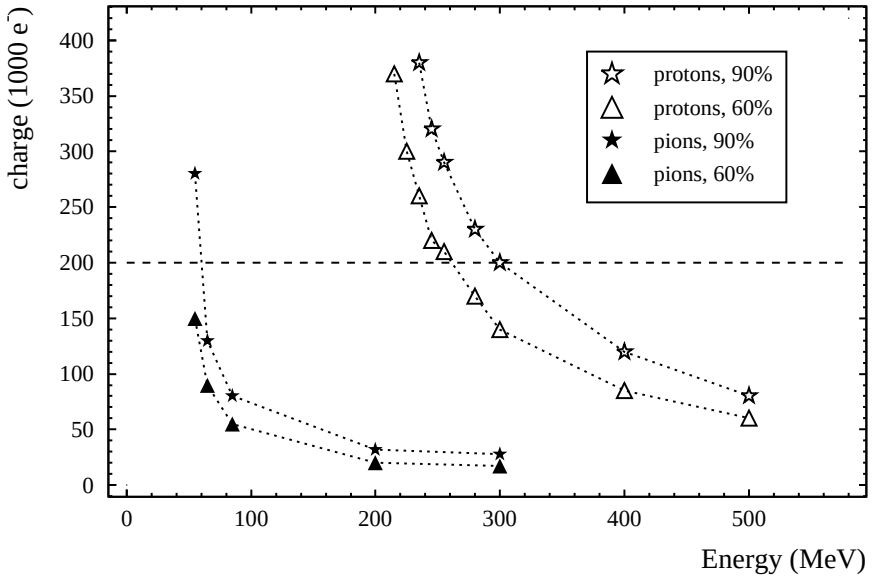

Figure 2.10: Dynamic range limits which allow 60% and 90% of non saturated signals for pions and protons, respectively

a slow system deteriorates the resolution on close-tracks.

For a first evaluation of this aspect, it is more convenient to discuss it in the time domain. If the amplifier is modelled with a linear, time-invariant network, its effect on close tracks separation can be quickly estimated with the following considerations:

- The signal at the output of the amplifier is the convolution of the detector signal with the pulse response of the amplifier
- Since the detector signal is Gaussian, the convolution can be easily evaluated if the pulse response is Gaussian as well [12].

The latter is only an approximation useful to simplify calculations, because a Gaussian filter

can not be implemented with a lumped electrical network. However, we shall see that these calculations are in good agreement with the computer simulations taking into account the actual transfer function of the amplifier. The convolution of two Gaussian signals is still a Gaussian signal, whose standard deviation is defined by the relation

$$\sigma_{out} = \sqrt{\sigma_{in}{}^2 + \sigma_{amp}{}^2} \tag{2.1}$$

where:

σ_{out} is the standard deviation of the output signal
σ_{in} is the standard deviation of the input signal
σ_{amp} is the standard deviation of the approximated amplifier pulse response.

From eq. 1 we can distinguish three cases, which are schematically illustrated in fig. 9

- $\sigma_{amp} \ll \sigma_{in}$: in this case the shape of the output signal will be a replica of the shape of the input one, amplified with the DC gain of the amplifier. If, for simplicity, we suppose the amplifier saturates at 1 V, the amplitude of the smallest signal of interest is 10 mV (= 1.3 μA/13 nA)

- $\sigma_{amp} \gg \sigma_{in}$: this situation is opposite to the previous one. The shape of the output signal will be that of the amplifier pulse response and for the signals of interest the amplifier acts like an integrator. Therefore, if 1 V is the value which corresponds to an input signal carrying a charge of 32 fC, the amplitude of the minimum signal is 31 mV (= 32 fC/1 fC)

- $\sigma_{amp} \approx \sigma_{in}$: we are in an intermediate situation and the output can be derived calculating expressly the convolution integral.

The effect of the convolution is to "stretch" the signals in the time axis, which in our case corresponds to the drift direction.

We suppose that two signals can be easily separated if the distance between their centroids is three times their standard deviation. If $\sigma_{amp} \ll \sigma_{in}$, the close-track resolution can be estimated as $3 \times \sigma_{in,max} \times v_d$ where $\sigma_{in,max}$ is the maximum standard deviation in the input signal and v_d is the drift speed. For the SDD in ALICE $\sigma_{in,max}$ is 30 ns and v_d is 6 μs, which yields an intrinsic a close-tracks resolution of 540 μm. The standard deviation of the amplifier pulse response should be short compared to the one of the signal, in order not to degrade the resolution of the detector. However, due to the quadratic nature of eq. 1, an amplifier with σ_{amp}=15 ns would degrade the close- track resolution of only 10 %. In the time domain a useful metric of the amplifier speed is the peaking time of its response to a δ-like input which, in the Gaussian approximation, can be assumed to be three times σ_{amp}. In the computer simulations the amplifier has been represented with a second order low-pass filter. From a theoretical point of view this is not the optimal filter for a SDD, [10] but it is a good solution when also silicon area and power consumption are taking into account.

The detailed simulation led to a choice of a peaking time of 50 ns. If we suppose again that we saturate at 1 V, a signal of 1 fC coming far from the anodes produces an output of 26 mV.

2.3.4 Analog to digital conversion

After the amplification, the signals should be converted as soon as possible in a digital form. In fact, any further processing should be done in the digital domain, which is much more flexible and robust than the analog one.

2.3.4.1 Sampling frequency

The minimum sampling frequency can be estimated in the following way. In order to reconstruct each pulse, at least three samples should be taken well above the noise, i.e in the interval [$+2\sigma$, -2σ] from the maximum. The fastest signals coming from the detector have a σ of 10 ns; using again equation 1, with σ_{amp} = 16.7 ns, we find that the σ of the signal at the output of the amplifier will be 19.4ns. The requirements of three samples in the interval [$+2\sigma$, -2σ results] therefore in a sampling frequency of 40 MHz. This value is in excellent agreement with the one deduced with the computer simulations. In the past two years several experimental test have been carried out with particles beam in the CERN accelerating facilities. In these tests, a silicon drift detector has been read-out with a amplifier with 50 ns peaking time and a digitising system whose sampling frequency was changed from 20 to 40 MHz. The results of this tests are consistent with the simulations and show that a peaking time of 50 ns and a sampling frequency of 40 MHz meet the resolution requirements [20].

2.3.4.2 ADC resolution

The full scale range of the ADC must correspond to the signal delivered by the amplifier for an input charge of 200.000 e^-. The resolution required to the converter can be estimated in the following way. In an ADC the error introduced by the quantisation process is characterised by the ratio between the power of a full range signal and the power of the quantisation noise. This number depends on the input signal and is:

- $6N + 1.76$ dB if the input is a sinusoid
- $6N + 4.77$ dB if the input is a square wave
- $6NdB$ if the input is a triangular wave

In the expressions above, N is the number of bits of the converter. For the sake of simplicity, we will approximate the output of the amplifier with a triangular waveform and evaluate the signal-to-quantisation noise ratio (SNR_q in the following) as $6N$ dB. The maximum signal of interest should fit with the full scale range of the converter and the minimum signal should be digitised with a resolution high enough so that the quantisation noise does not deteriorate the measurement of the energy lost by the particle.

Due to the statistical nature of the inonization process, the measurement on dE/dx has an error which is typically of the order of 10%; it is important to stress that this error is intrinsic in the physics of the interaction between the particle and the medium and is not caused by non-idealities in the processing electronics. The error introduced by the quantisation noise on low-level signal should therefore be small compared to the one induced by the stochastic fluctuations. If we consider the two source of errors as uncorrelated, an error of the 3% caused by the digitization will contribute only for a 0.4%. Therefore, on *small signals* an equivalent resolution of 5 bits is sufficient, which corresponds to a SNR_d of 30 dB. I we use an amplifier with a peaking time of 50 ns, the smallest signal (26 mV) is about 32 dB below the full scale range and the resolution required to the ADC is $30 + 32$ dB $\approx$ 10 bits. We have however to take into account the noise of the amplifier. An equivalent input noise of 250 e^- gives at the output of the amplifier a fluctuation of 1.2 mV rms. The rms error introduced by an ideal quantiser on the amplitude measurement is $V_{LSB}/\sqrt{12}$, where V_{LSB} is the size of the least significant bit. If we suppose that the two errors can be of the same size, we have that V_{LSB} = 4 mV, which, on

1V full scale range corresponds to an 8 bit converter. With two extra bit, the error introduced by the converter is negligible and the noise of the front-end electronics can be measured. This is important for the diagnostic of the system, since the noise is an indicator of the performances of the detector and the front end amplifier which are of course the most sensitive and critical blocks.

2.3.5 Timing requirements

In high energy physics only a few of the events produced in a collision are relevant and only the signals related to these events are usually recorded. The selection is provided by a system of dedicated detectors and electronic modules that generates the *trigger* signal, which is distributed to the whole apparatus. When a trigger signal is received, the information temporary stored in the front-end electronics of the subdetectors is transferred for permanent memorisation. During this read-out phase, the detectors are not usually able to process a new event. This *dead time* must be minimised, in order to keep the experimental system sensitive for most of the time.

The trigger is characterised by two main parameters: the latency and the rate. The latency is the time elapsed between the actual physical event and the activation of the trigger and is due unavoidable delays in the triggering apparatus. The rate is the mean frequency of occurrence of a trigger, which, owing to the statistical nature of the events, is random. The probability to have a given number of events in a given interval of time is described by the Poisson distribution

$$P(n) = \frac{\mu^n e^{-\mu}}{n!} \tag{2.2}$$

where n is the number of events and μ is the mean value in the considered interval. The whole ALICE detector must fully process, on average, 50 events per second and with equation 2 we can estimate the dead time that can be tolerated. Since the events which occur during the read-out are rejected, we require that the probability to have an event during the read-out time is at most 5%. In this case, using eq. 2 we have

$$P(1) = 1 - P(0) = 1 - e^{-\frac{\Delta t}{1/f}} = 0.05 \tag{2.3}$$

where f is the trigger rate. Solving for Δt, we found that a read-out time of 1 ms is allowed. Within this time the data must be digitized and transferred on tape.

2.3.6 Data volume

The amount of data produced by the SDD for each event is enormous. In fact, assuming a sampling frequency of 40 MHz and a drift speed of 6 μm/ns, 240 time samples are generated for each anode; this corresponds to 900 kbytes/ladder in layer 3 and 1.2 Mbytes/ladder in layer 4. Therefore the total amount of data produced per one event is 39 Mbytes, which as to be compressed to 1.5 Mbytes (the space allocated to the SDD on the storage media).

The main parameters which drive the design of the front-end electronics are listed in table 1.

Table 2.1: Requirements for the front-end electronics

Detector signal range (charge)	1 - 32 fC
Detector signal range (current)	20 nA - 1.3 μA
σ	10 ns - 30 ns
Target noise	$250e^-$
Number of bits	10
Sampling frequency	40 MHz
Power consumption/anode	5mW
Maximum read-out time	1ms

2.4 The front-end architecture

The first element in the front-end is the amplifier; given the small capacitance of the detector and the small analog bandwidth, this unit is not particularly critical and a number of good example can be found in the literature [13, 14]. In this work we have hence focused our attention to the other blocks in the front-end.

2.4.1 Sampling strategy

As we have seen in the previous section, the amplified signals must be sampled with a frequency of 40 MHz. Ideally, this would be accomplished by connecting each output of the amplifier to a fast ADC with 40 MS/s capability. State of the art ADC feature a power consumption of 1mW/MS and the power necessary for this scheme (40 mW/channel) is hence far beyond our limit.

A more realistic solution is to implement a temporary analog storage using a switched capacitor array. In this circuit, which is widely used in particle physics and is commonly known as "analog memory" the analog information is stored in capacitors selected by a shift register. In this way, the signal can be temporary stored with an high sampling frequency and digitised later at a lower speed only if an interesting event occurs. Since in this circuit only capacitor, switches and a minimum amount of digital logic are used, the power consumption is very low. In the literature, analog memories with a sampling frequency of 700 MHz and a power consumption of 2 mW/channel have been reported [22].

Due to the drift mechanism, when an event takes place the charge deposited by the particles is collected at the anodes with a delay which depends on the crossing point. The maximum possible delay is around 6 μs and the output of each anode must be sample for this time in order not to lose signals. Afterwards the memory can be rewritten. The maximum drift time divided by the inverse of the sampling frequency gives the number of storage units required, which in our case is 240. It is however better to have some more units, so that small variations in the detector parameters can be tolerated. A number of 256 cells has been considered suitable.

The amplifiers and the SCA can be implemented on the same chip and several good examples are found in the literature. For the time being we assume a modularity of 64 channels/chip, as a good compromise between integration density and yield

2.4.2 Analog to digital converter

Upon a reception of a trigger signal, the data of the analog memory should be converted into a digital form in less than 1 ms. Actually, the conversion procedure should be as short as possible, in order to limit the degradation of the stored analog signals due to the droop effect in the sampling cells. In a switched capacitors circuit, the droop rate is determined by the leakage current of the switches and the size of the capacitors. Assuming a leakage current of 1 pA, a storage unit with a sampling capacitor of 1 pF exhibits a droop rate of 1 mV/ms. However, the leakage current is function of the temperature and may increase during the operation of the system because of radiation induced damages. The conversion time should hence be minimized within the allocated power budget. Moreover, the 1 ms must include also the time necessary for data formatting and transmission; therefore we have allocated to the conversion process at most 500 μs, with the aim of minimising it within the limits of the allowed power consumption. In principle, this goal can be achieved either by using a fast ADC and multiplexing many channels of the analog memory or by using a slower ADC which converts only few raws. The use of a commercial ADC has of course many advantages, since state of the art ADCs features striking performances [15]. However, we must also consider that the use of a commercial converter imply the transmission of analog data out of the front-end chip, which in a huge system like ALICE may entail severe problems of signal integrity. Therefore, we have investigated the possibility of integrating the ADC and the analog memory on the same chip.

In the literature some examples are found in which the analog to digital conversion is performed on the front-end embedding one single ramp ADC in each channel [17, 18]. A single ramp ADC provides accurate conversion with a minimum of components, but since it requires 2^n clock cycles to complete a n bit conversion it is by far too slow for out application. Assuming as a reference the clock of the full system[2], about 25 μs are needed for one conversion. the generation of an accurate reference ramp on chip is also an issue. A much more attractive solution is a successive approximation converter, which requires only n clock cycles for a n bit conversion, plus the time necessary to sample the input signal. If we suppose that the sampling time and the conversion time are equal[3] an ADC working with 10 MHz clock will be sufficient to comply with the requirement of 500 μs conversion time for one arrow of the analog memory. A further option is to use a faster ADC and to multiplex the channels of the analog memory. However, even with architectures like subranging or pipeline, at least two ADCs per chip would be needed to fit in the 500 μs requirement. Given the complexity of the overall system, which is supposed to work for ten years with a minimum of maintenance, a high modularity is however a big advantage. Front this point of view a system with an ADC per channel or every few channels is preferable, because a failure in the converter will determine the loss of only a portion of the chip.

In sections 3 we have shown that the requirement of a 10 bits full scale resolution for the ADC is a consequence of the large dynamic range and of the need to preserve an acceptable resolution (5 bits) on the smallest signal of interest. Actually, 10 bits provide a redundant information on full scale signals and this fact can be used to perform a first compression of the data. A viable alternative is hence to use an ADC with smaller resolution and to adapt the input

[2]For system reason, only a single 40 MHz clock will be distributed in the detector. Of course on chip clock multiplications using PLLs are possible, but we would like to keep the clock on the front end as low as possible in order to minimise interference with the analog parts.

[3]As we shall see in the next chapter this assumption is justified by the fact that, in this kind of ADC, the sampling time constant is longer than the conversion time constant.

signal to the full scale range of the converter by introducing a nonlinearity in the signal path. In this way, there is the possibility to perform a first compression, reducing, for example, the output code of the ADC from 10 to 8 bits.

In principle, in our case two solutions are possible:

- The use of a non linear amplifier in the front end. This approach simplifies the design of the analog memory and of the ADC (which need only 8 bit accuracy) at the expense of an increased calibration complexity.
- The use of a bilinear (o multi-linear) ADC, that has in turn the drawback of demanding a supplementary analog decision circuit to detect the appropriate scale for the conversion.

As far as only the compression is concerned, a reduction from 10 to 8 bits of the ADC output preserving an adequate resolution can be easier performed in the digital domain. A possible scheme has been recently proposed and works as follow: [16]

- When the ADC output code lies in the interval $[0-127]$ it is transmitted unchanged and MSB of the 8 bit code is used to identify the range. The output is therefore: 0*xxxxxxx* and no bit is lost.
- If the ADC code is in the range $[128-255]$, the number is divided by for and mapped on 5 bits. The 8 bits output is 100*xxxxx* and 2 bit of information are lost
- When the ADC code is in $[256-511]$, the number is divided by 8 and the output is 101*xxxxx*. In this case the three LSBs are lost.
- Finally, in the region $[512-1023]$ the ADC code is still divided by 8. Also in this case three LSBs are lost at the output is 11*xxxxxx*

This scheme has the advantage that all the 256 codes are efficiently used and the compression is intrinsically monotonic. A compression from 10 to 8 bits in the digital domain still requires a 10 bits front-end. However, the linearity has to be at this level only in the lower part of the dynamic range and deviation from linearity up to 1% can be tolerated in the higher part (above 1 mip).

2.4.3 Data transmission

After the A/D conversion the data are stored in local registers and than transmitted to the end ladder board.

If we use only one level of registers, after each conversion they must be emptied in order to be ready for accommodating the result of the next conversion. As a consequence, for each detector, 512 bytes[4] have to be transmitted in less than 2 μs. If the transmission takes place at 40 MHz clock, seven 8 bit busses per detector are required. This represents a huge amount of cables that can not be tolerated.

For each events, 64kbytes of data are produced by each half detector. This data can be stored in a RAM and than transmitted at a lower speed. Actually, at least 2 RAM are needed in order to be able to accept two consecutive events. Since in this way the transmission an hence the dead time from the RAM chip onwards can be tuned on the average event rates (50 Hz) a one 8 bits

[4] We suppose here that the compression from 10 to 8 bits has already been performed

bus is sufficient for half a detector. Simulation have been run to estimate the extra dead time due to possible buffer overrun. The results show that with 2 64k RAMs for each half detector it is introduced an extra dead time of 0.1 %, which is negligible [20]

2.4.4 Radiation tolerance

The radiation levels in the ITS are not very high. For the inner layer of the Silicon Drift 13 krad are expected during the whole life of detector, which is probably below what a standard technology can afford. Nevertheless, the presence of an SCA in the electronics chain deserves particular attention. In fact the integrated dose increases the leakage current of the transistors and can significantly affect the droop rate of the sampling cells. For example, a leakage current of 1 nA, which would be negligible in most applications, changes the value of the information stored in 1 pF capacitor by 0.5 Vin 500 μs, which is of course unacceptable. Some radiation tolerant technique should therefore be implemented to prevent this risk.

2.4.5 System partitioning and technological considerations

The mechanical constraints impose that only one board of 8 x 2 cm^2 must be used to read out 256 anodes. On this space, at most 8 chips can be realistically accommodated using conventional mounting (i.e. without MCMs). The baseline choice has been to split the front-end processing into two units, one dedicated to the amplification, sampling, and analog to digital conversion and one for the double-event digital buffer. The design of this circuit does not entail any significant issue; a digital chip working with a 40 MHz clock is not a critical design for state of the art CMOS technologies and the layout can be easily synthetised from the VHDL code.

On the other and, the design of the analog unit is more critical, because imply the integration of several high resolution blocks and the coexistence of analog and digital functions on the same silicon. To minimize the interference, only the minimum logic necessary for the control and the A/D conversion will be implemented on the front-end; a careful splitting of the analog and digital supplies is of course mandatory. Due to space and power constraints, only passive power supply filtering can be used on the front end board near. The use of electrolytic capacitor is not reccomended, since a failure in such a capacitor may have unpleasant consequences on the chips and the detectors, which, of course, are all mounted unpackaged.

Particular care is required by the analog to digital converter; in order to maximise the fault tolerance, the chip should have the highest possible modularity and the integration of one converter per channel is desirable. For the resolutions required, the successive approximation ADC provides an excellent trade-off between speed and power and is a promising candidate; the use of this architecture is multi-channel applications demands a careful effort in minimising the area and in mastering the compatibility problems arising from the fact many converters have to share the same reference voltage.

As we have discussed in chapter 1, we believe that the use of a quarter micron CMOS technology in conjunction with adequate design techniques does not impair and may even improve the performances of analog circuits. The radiation tolerance, achievable simply with special layout, provides an additional benefit. After some investigation, a 0.25 μm process has therefore been selected as the baseline technology for the implementations of the final chips.

2.5 Summary

In this chapter, the specifications of the front-end system of the Silcon Drift Detectors to be used in the ALICE experiment have been described and an architecture to meet the requirements has been proposed. This architecture is based on two integrated circuits, one for the amplification and analog to digital conversion of the detector signals and one for the buffering of the events. The former chip is more critical and requires a full-custom design, which, given the system specifications in term of noise, speed and power, entails several issues. In particular, since the power/channel (5 mW maximum) is not sufficient to allow the use of a fast ADC per channel, the sampling and the conversion functions must be decoupled, implementing the former with a fast switched capacitor array and the latter with a slow or medium speed ADC. Reliability considerations suggest to introduce the highest possible level of modularity and the integration of one ADC per channel would be a desirable feature. A successive approximation architecture has been selected as the best candidate to implement the converter.

After the initial contribution to the system definition, this thesis has been mainly concerned with the front-end chip, addressing the two most critical aspects: the design of arrays of analog to digital converters and the feasibility of a linear system with 10 bits resolution in a 0.25 μm CMOS technology.

The possibility of using a non linear amplifier in the front end, thus relaxing the specs of the analog memory and the ADC has also been investigated as a back-up solution.

The steps towards the design of this chip which demonstrate the feasibility of the project will be detailed in the next chapters.

3 Design of analog to digital converter arrays in submicron CMOS technologies

The switched capacitor successive approximation technique [23] (also known as charge redistribution technique) is very popular in the design of analog to digital converters, because it offers an excellent trade-off between speed, resolution and area. The low-power consumption and the presence of only one critical node inside the circuit make this approach suitable for multichannel architectures like the one described in the previous chapter for the front-end of the SDDs in the ALICE experiment.

In this work we have therefore focused our attention on the design of low-power and high-speed successive approximation ADC with the switched capacitor approach; the investigation has been carried-out designing and testing two prototypes in two different CMOS technologies. This chapter discusses some fundamental design issues, while the circuits and the measurements are presented in detail in chapter 4 and 5.

3.1 Successive approximation analog to digital converter

3.1.1 Basic principle

A successive approximation converter basically consists of an array of binary weighted capacitors and a comparator. The operations are controlled by a digital logic, which normally is simple and is not relevant for our discussion.

Fig. 1 depicts, as an example, an 8 bit converter; in this particular case the ADC operates from a single rail power supply and with a full scale range defined by the reference voltage V_{ref}.

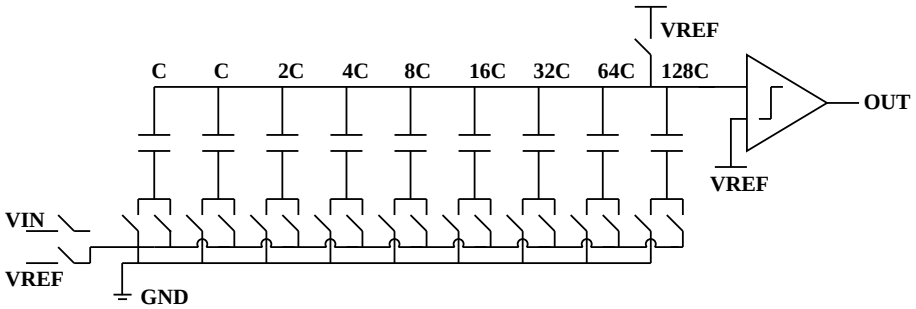

Figure 3.1: Scheme of a successive approximation ADC

The conversion is performed in three steps:

1. In the first step, the bottom plates of all the capacitors are connected to the input voltage V_{in}, whereas the top plate is connected to the reference. The ADC is in the *acquisition* mode.

2. In the second phase the bottom plates of the capacitors are switched to ground and the switch connecting the top plate to V_{ref} is opened. Due to the charge conservation, the voltage on the top plate changes from V_{ref} to $V_{ref} - V_{in}$. The ADC is in the *hold* mode.

   Its is important to note that the ADC performs intrinsically the sample & hold function, without requiring any additional circuitry.

3. In the last step, the ADC enters the *conversion* mode and finds the digital code by determining one bit per clock cycle.

   The output of the comparator feeds the digital logic, which connects the bottom plates to the appropriate voltage, according to the following scheme: in the first cycle the biggest capacitor in the array (128C) is switched to V_{ref}. The voltage resulting on the top plate is now $V_{ref} - Vin + \frac{V_{ref}}{2}$, which is compared with V_{ref} by the comparator. The voltage difference between the inputs of the comparator is $-V_{in} + \frac{V_{ref}}{2}$. If this voltage is < 0 and the top plate is connected to the inverting input of the comparator, the comparator asserts a logical 1. In this case, 128C remains connected to V_{ref}, otherwise its is flipped back to ground. In this way the most significant bit is determined. In the next cycles the other bits are found with the same procedure. Therefore, N clock periods are required for a N bit conversion. To speed-up the operation the transition to the hold mode and the test of the most significant bit can be merged in the same clock cycle

3.1.2 Power supply constraints

We assume that in the circuit of fig. 1 none of the internal nodes can move beyond the power supply rails. Two considerations motivate this constraint. First, we want to ensure always a proper functioning of the switches. For instance, if we suppose that the switch which controls the connection of the top plate to V_{ref} is implemented by a PMOS transistor, we see that it can turn on if the top plate moves beyond $V_{dd} + V_{TH}$, where V_{dd} is the positive power supply and V_{TH} is the threshold voltage of the device.

Second, we do not want to apply stress voltages on the gates of the transistors; this is of particular concern if the converter is powered at the maximum nominal power supply allowed by the technology. In particular, this limitation fixes a maximum value for the full scale range of the converter. Referring again to fig. 1, we see that the worst case in the direction of the positive rail occurs when a voltage close to zero is applied to the input. In fact, in this situation, the test of the most significant bit will rise the top plate of the capacitor array to $V_{ref} + \frac{V_{ref}}{2}$. Therefore, once V_{dd} is fixed, the input dynamic range is determined by

$$V_{ref} \leq \frac{2}{3} V_{dd} \tag{3.1}$$

and

$$V_{in} \leq V_{ref} \tag{3.2}$$

3.1.3 Power consumption

In a charge redistribution ADC the two main sources of power dissipations are the capacitor array and the comparator, whilst the control logic gives second order contributions. An high

precision comparator has fully analog elements which are permanently under bias. However, in very low power applications, this parts can be switched off during possible stand-by periods. The power consumption of the comparator strongly depends on the specific architecture, whereas the power consumption of the DAC can be very quickly evaluated [3].

As a worst case approximation, we can estimate the power dissipated when the input signal is equal to the full scale range. In fact, in this case, in the hold mode the top plate of the DAC goes to zero. During the bit search the top plate moves back towards V_{ref}. If the conversion is repeated several times, the top plates oscillates between V_{ref} and 0, so the average power consumption can be approximated as

$$P = \frac{1}{2} 2^N C {V_{ref}}^2 f_s \tag{3.3}$$

where f_s is the sampling frequency and C is the value of the smallest capacitor in fig. 1. If we assume that one clock cycle is needed for the acquisition of the signal, then $f_s = f_{ck}/(N+1)$.

The power consumption can be lowered lowering the size of the capacitors or the reference voltage. The minimum size of the capacitor is constrained by the design rules and by matching considerations and typically is of the order of 50 fF. A charge redistribution DAC with a minimum capacitor of 50 fF, 10 bit resolution and operating with a reference of 2 Volt at 1 MS/s will dissipate only 0.1 mW. From this point of view, this architecture is very attractive for multi-channel implementations.

3.1.4 Speed limitations

The limitations on the maximum speed of a charge redistribution converter come both from the DAC and the comparator. We will discuss the limitations of the comparator in the next section and we concentrate here only on the DAC.

From the point of view of speed and bandwidth evaluation, the DAC and the associated switch network can be modelled as a RC filter having two distinct time constants for the acquisition and the redistribution mode respectively [23].

When the ADC is sampling the input signal, the time constant can be estimated as

$$\tau_{ac} \simeq (R_{Vref} + R_{Vin} + \frac{R_{MSB}}{2}) C_{DAC} \tag{3.4}$$

where R_{Vref} and R_{Vin} are the equivalent resistances of the switches connecting the capacitor array to the reference voltage and the input and C_{DAC} is the total capacitance (equal to 256C for the circuit of fig. 1). R_{MSB} is the resistance of the switch driving the capacitor that is used to determine the most significant bit (MSB). The above expression holds exactly if the switches are sized according to the value of the capacitor they drive, so that $R_i C_i = R_k C_k \ \forall \ i,k$ in the array.

Within the acquisition time, the converter has to settle to the final value with an error smaller than 1/2 LSB. There is therefore the following relationship between the acquisition time, the acquisition time constant and the number of bits N

$$T_{ac} = 0.69(N+1)\tau_{ac} \tag{3.5}$$

In the acquisition mode, the small signal bandwidth has a cut-off frequency defined by

$$f_T = \frac{1}{2\pi\tau_{ac}} \tag{3.6}$$

However, this is only an approximation; actually the exact value of R_{Vin} depends on the value of the input signal and thereby can introduce signal dependent delays and harmonic distortion [24].

During the redistribution phase, the time constant is

$$\tau_R = \frac{R_{MSB}}{2} C_{DAC} \tag{3.7}$$

and has to be small compared to the time allocated to the comparator for the decision. This time is usually half a clock cycle, so, for the redistribution mode, eq. 5 can be rewritten as

$$T_R = 0.69(N+1)\tau_R = \frac{T_{ck}}{2} \tag{3.8}$$

where T_{ck} is the period of the clock. The sampling frequency of a charge redistribution ADC can then be expressed as

$$f_s = \frac{1}{T_{ac} + NT_{ck}} \tag{3.9}$$

Since τ_{ac} is greater than τ_R, the acquisition time may require more than one clock cycle.

3.1.5 Effects of DAC non idealities in charge redistribution converters

In a charge redistribution converter the main sources of errors are due to mismatches in the binary weighted DAC and to offsets in the comparator. The function of the DAC is to provide fractions of the reference voltage that are compared with the input signal. A deviation of the capacitors from their ideal values determines mismatches in the partitions of the reference voltage an hence limits the number of codes that can be generated by the ADC. A simple way to estimate this limit is to require that the DAC error, at mid-scale, is less than 1/2 LSB, which corresponds to the requirement that the integral nonlinearity of the converter is at most 1/2 LSB [24].

When the MSB is generated, the bottom plate of 128C (see fig. 1) is connected to V_{ref} and the bottom plates of the remaining capacitors are connected to ground. The sum of this capacitors is 128C, so in this situation we have actually only two capacitors, which are nominally equal. For simplicity we call these capacitors C_{MSB} and C_{GND} and we suppose that can deviate from their nominal value for a quantity $\pm 0.5\Delta C$, so that:

$$C_{MSB} = C - 0.5\Delta C \tag{3.10}$$

and

$$C_{GND} = C + 0.5\Delta C \tag{3.11}$$

We require now that the maximum deviation of the voltage generated at the MSB transition is less than 1/2 LSB from its theoretical value, that is

$$\left| \frac{V_{ref}}{2} - \frac{V_{ref}(C-\Delta C)}{(C+\Delta C)+(C-\Delta C)} \right| < \frac{V_{ref}}{2^{N+1}} \tag{3.12}$$

which yields

$$\frac{\Delta C}{2C} < \frac{1}{2^N} \tag{3.13}$$

Eq. 4 enables us to calculate the number of bits that can be achieved in a given technology if the matching of the capacitors is known. The maximum number of codes that can be generated with a charge redistribution approach is hence limited by mismatches between the capacitors in the binary weighted array. In fact, while the technique is insensitive to the absolute value of the capacitors, a deviation of the *ratios* of the capacitors from their ideal value is source of nonlinearity. To improve the matching the DAC is usually laid-out using only one elementary cell which is repeated many times to form the different capacitors [23]. This has the benefit that errors due to the etching of the masks used to define the capacitors affect all the cells in the same way. Another contribution to the mismatch comes from long range gradients in the capacitors oxide and can be minimised with a common centroid layout.

Even if these precautions are taken, small random errors persist and place an ultimate limit on the resolution of the ADC. The random errors can be studied with a statistical approach. In fact, if we suppose that the errors are uncorrelated and normally distributed, we can associate to each capacitor a mean value C and a standard deviation σ_C. A capacitor C' formed by connecting M capacitors C in parallel will have a standard deviation $\sqrt{M}\sigma_C$

The condition that the INL for the MSB is less than 1/2 LSB can be reformulated in the following way:

$$\frac{V_{MSB}}{V_{ref}} = \frac{1}{2}\left(1 \pm \frac{1}{2^{N+1}}\right) = \frac{C_{MSB}}{C_{DAC}} = \frac{C_{MSB}}{C_{GND} + C_{MSB}} \tag{3.14}$$

where again C_{GND} represents all the other capacitors in the array, whose bottom plates are connected to ground while the MSB is tested. For a n bit DAC built with 2^n equal capacitor of value C, C_{MSB} is given by $2^{n-1}C$. We can therefore write

$$C_{MSB} = 2^{n-1}C \pm \sqrt{2^{n-1}}\sigma_C \tag{3.15}$$

$$C_{GND} = 2^{n-1}C \pm \sqrt{2^{n-1}}\sigma_C \tag{3.16}$$

The worst case occurs if C_{MSB} has a positive deviation from its ideal value and C_{GND} a negative one, or vice versa. In this case eq. 14 becomes

$$\frac{2^{n-1}C + \sqrt{2^{n-1}}\sigma_C}{2^{n-1}C} = \frac{1}{2}\left(1 \pm \frac{1}{2^n}\right) \tag{3.17}$$

Solving for σ_C/C, yields:

$$\frac{\sigma_C}{C} = \frac{1}{\sqrt{2}\sqrt{2^n}} \tag{3.18}$$

which is the percentual standard deviation allowed on the unit capacitor in order to get a given a resolution of n bit with an integral nonlinearity smaller than 1/2 LSB. It is interesting to observe that the INL tends to improve by putting more capacitor in parallel (i.e. using DAC with bigger size) because random errors tend to average-out. Another important point is that the resolution of the DAC does not depend on the reference voltage. This provides the opportunity of adapting the reference voltage to the signal to be digitized, building multi-range converters with very high dynamic range [27].

3.1.6 Area consideration

One of the main drawbacks in using the charge redistribution architecture for high resolution is the area, which doubles for every extra bit required.

In CMOS technology, high quality linear capacitors are formed by sandwiching a thin oxide between two layers of polysilicon or metal. The density of these capacitors is between 0.8 and 1 fF/μm^2; therefore a 50 fF capacitor required about 20 $\times$ 20 μm^2 average area, taking into account also spacing between capacitors, routing and the use of dummy cells at the edge of the DAC to minimise side effects. The area consumption, reasonable up to 8 bits, becomes cumbersome for higher resolutions, especially if parallel applications are aimed.

Another problem related to the increase of the DAC size is the capacitive loading. Assuming an elementary capacitor of 50 fF, the total capacitance of a 10 bit DAC is about 50 pF. This represents an heavy load for the circuit which has to drive the ADC. We will see in this chapter 5 how a 10 bits ADC has been implemented with only a 15%area penalty compared to a 8 bit solution and without increasing the load on the driving circuitry.

3.2 Design of fast and high resolution comparators in CMOS technologies.

In synchronous applications high resolution comparators are efficiently implemented using positive feedback. An example of this principle is shown in fig. 2, where two cross-coupled transistors are used to load a differential pair.

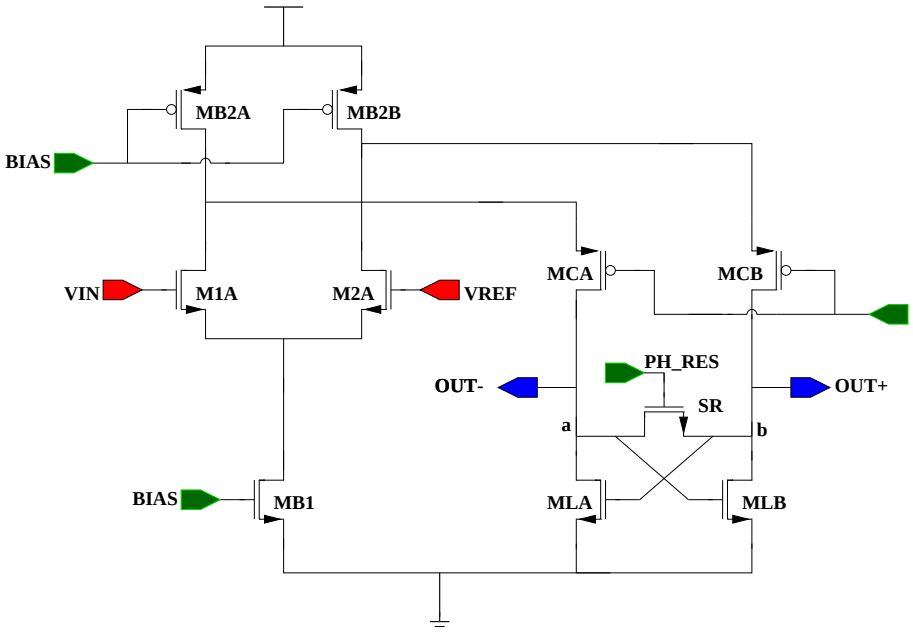

Figure 3.2: Example of a synchronous comparator

In this circuit, the comparison is carried out in two phases: in the *tracking* phase the switch S_R is closed and the latch is reset. In the *decision* phase the latch is enabled opening S_R ; the

positive feedback mechanism amplifies then the voltage difference between the nodes a and b of the latch to levels adequate for driving a digital circuit.

In order to optimise the performances of a positive feedback comparator two main issues must be properly addressed: the regeneration time constant of the latch and the offset of the whole circuit. This section is dedicated to an analysis of these aspects.

3.2.1 Speed of positive feedback comparator

Following [25] , we start investigating the dynamic properties of the positive feedback stage by modelling it with two back-to-back inverters. The model is depicted in fig. 3, which shows also the small signal equivalent circuit. Here G_m is the equivalent transconductance of each inverter, whereas R_L and C_L represent the loads a the output of each stage.

The nodal equations of the circuit can be written as

$$G_m v_b + C_L \frac{dv_a}{dt} + \frac{v_a}{R_L} = 0 \tag{3.19}$$

$$G_m v_a + C_L \frac{dv_b}{dt} + \frac{v_b}{R_L} = 0 \tag{3.20}$$

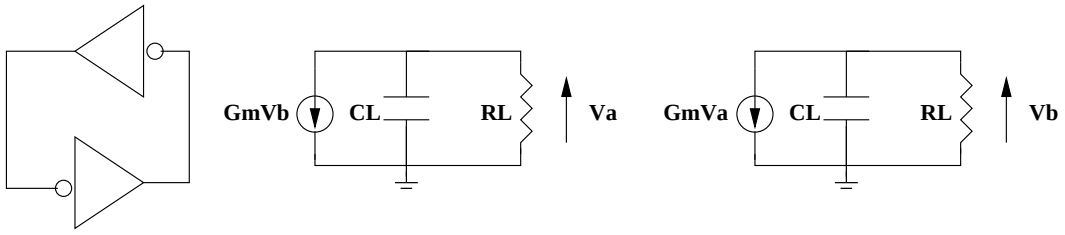

Figure 3.3: Simplified small signal model of a latch used in positive feedback comparator

The solution of the above differential equations gives the voltage difference $v_b - v_a$, which is

$$v_b - v_a = (v_{b0} - v_{a0}) e^{(G_m R_L - 1)\frac{t}{\tau}} \simeq (v_{b0} - v_{a0}) e^{\frac{G_m}{C_L} t} \tag{3.21}$$

where v_{b0} and v_{a0} are the initial voltages at nodes b and a, respectively. We assume here that the low frequency gain of the inverters, defined as $G_m R_L$, is $\gg 1$.

The regeneration time constant can be easily calculated for the comparator in fig. 1. In this case, in fact , G_m is just the transconductance g_m of MLA or MLB at the beginning of the regeneration phase. The dominant parasitic capacitance is usually the gate-source capacitance of MLA (MLB). Therefore, if we suppose that the devices are in saturation, we have

$$\tau = \frac{2}{3} \frac{L^2}{\mu (V_{GS} - V_{TH})} \tag{3.22}$$

where L is the channel length of the devices, V_{GS} is the gate source voltage, V_{TH} is the threshold voltage and μ is the mobility of the carriers (electrons in our example). This relation shows that

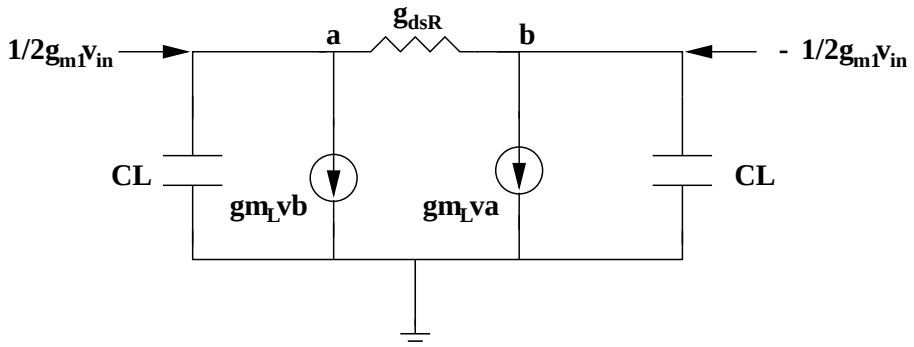

Figure 3.4: Equivalent circuit of the latch to take in account the on resistance of the reset switch (M_{SR} in fig. 2)

the time constant does not depend on the width of the transistors, but only on the length and the over-drive voltage $(V_{GS} - V_{TH})$.

For a proper design of the comparator, the contribution of the reset switch S_R must also be considered. When it is closed, in fact, the switch creates a resistive path between the two drains of the transistors forming the latch. The appropriate small signal model is shown in fig 4 [26]. This model is used to derive the nodal equations

$$\frac{1}{2}g_{m1}v_{in} - C_L\frac{dV_b}{dt} - g_{mL}v_a - g_{dsR}(v_b - v_a) = 0 \tag{3.23}$$

$$-\frac{1}{2}g_{m1}v_{in} - C_L\frac{dV_a}{dt} - g_{mL}v_b + g_{dsR}(v_b - v_a) = 0 \tag{3.24}$$

whose straightforward solution gives the the differential output voltage

$$v_b - v_a = \frac{g_{m1}}{2g_{dsR} - g_{mL}}v_{in} + \left[v_{b0} - v_{a0} - \frac{g_{m1}}{2g_{dsR} - g_{mL}}v_{in}\right]e^{\frac{t}{\tau}} \tag{3.25}$$

The on resistance of the switch enters also in the regeneration time constant τ, which now is defined by

$$\tau = \frac{C_L}{g_{mL} - 2g_{dsR}} \tag{3.26}$$

During the regeneration phase, $g_{dsR} \simeq 0$ and eq. 25 becomes

$$v_b - v_a = \frac{g_{m1}}{g_{mL}}(e^{\frac{t}{\tau}} - 1)v_{in} + (v_{b0} - v_{a0})e^{\frac{t}{\tau}} \tag{3.27}$$

Therefore the output voltage is the sum of two contributions, one due to the initial unbalance and one due to the loading effect of the latch.

From eq. 26 we see that, when the switch is closed, the condition $2g_{dsR} > g_{mL}$ must be satisfied in order to switched off the positive feedback. If a voltage difference is applied between the inputs of the comparator when it is in the reset mode, this, after a time $t \gg \tau$ will appear at the output amplified by a factor

$$A_V = \frac{g_{m1}}{2g_{dsR} - g_{mL}} \tag{3.28}$$

As we shall see in chapter 5, it is possible to use this *reset mode gain* to improve the offset performance of the comparator.

Equation 27 enables us to calculate the minimum time that the comparator needs to regenerate a given input signal to a given output level. If the comparator is not fast enough, the output voltage may not reach the value required to drive the following digital circuitry, thus determining a logic error. The study of this phenomenon, (called *metastability*) is carried out with a statistical approach and leads to the definition of an important performance metric for comparators: the bit error rate (BER).

We address the problem starting from eq. 27, which, for the sake of simplicity, we rewrite as

$$v_b - v_a = \Delta v = v_m e^{\frac{T}{\tau}} \tag{3.29}$$

where T is the time allocated to the comparator for the decision phase, Δv is the minimum output voltage to be reached in order to avoid a metastable condition and v_m is the corresponding input voltage. Let's now suppose that the comparator has a input range of v_R and that all the input values are equally likely; the probability that $v_{in} \leq v_m$ can be then expressed as

$$P(m) = \frac{v_m}{v_R} = \frac{\Delta v}{v_R} e^{-\frac{T}{\tau}} \simeq e^{-\frac{T}{\tau}} \tag{3.30}$$

that can be used to calculate the regeneration time constant needed to obtain a given bit error rate. For example, if we want a probability of error $P(m) < 10^{-10}$, solving eq. 30 yields

$$0.43\frac{T}{\tau} < 10 \rightarrow \tau < 23T \tag{3.31}$$

It is important to observe that metastability is equally likely for each bit in the code and determines huge corruption of data if it occurs in the most significant bits. Therefore, we have to require that the loss of data due to this phenomenon is negligible compared to the global efficiency desired from the system.

3.2.2 Offset minimisation techniques

In an analog to digital converter, the quality of the comparator determines the size of the LSB. Since a comparators use differential architectures, any mismatch between the two branches of the circuit creates offset and hence are source of errors. In this discussion, the offset is always referred to the input in order to directly compared it with the signal to be discriminated.

In principle, if no mismatch was present, a single latch could be used as comparator, thus reducing circuit complexity and static power dissipation. The offset of the positive feedback pair MLA - MLB of fig. 1 can be easily estimated by calculating the variation of the drain current $I_d = \frac{kW}{2L}(V_{GS} - V_{TH})^2$ with respect to all the parameters and dividing by the transconductance g_m. The calculation yields [24]

$$V_{OSL} = \Delta V_{TH} + \frac{1}{2}\left(\frac{\Delta W}{W} + \frac{\Delta L}{L}\right)(V_{GS} - V_{TH}) + \frac{\Delta Q}{C_D} \tag{3.32}$$

where V_{TH} is the mean value of the threshold voltage, ΔV_{TH} its standard deviation, $\Delta W/W$ and $\Delta L/L$ are the relative mismatches in the dimensions of the transistors, $V_{GS} - V_{TH}$ in the overdrive voltage when the latch is strobed. Actually, in eq. 32 there is an additional term, $\frac{\Delta Q}{C_D}$ which takes

into account the mismatch in the charge injection from the switch S_R on nodes a and b. A calculation of the offset for the latch with the typical parameter of a 1.2 μm technology leads to an estimation of 50 mV for the offset of the latch in fig. 2.

The gain of the differential pair reduces the total offset by a factor $\frac{g_{m1}}{g_{mL}}$. However, this gain cannot be to high without degrading the speed and ratios between 5 and 10 are commonly used. Moreover, the input stage stage suffers from its own offset as well, so to reduce the overall offset below 1 mV some compensation strategy must be applied.

In the literature, several techniques have been proposed to minimise the offset. However, the majority of these techniques can be traced to two fundamental topologies, one based on *offset storage* and one based on the use of a *servo-loop* in conjunction with an auxiliary input stage. The techniques which have been used in the ADCs presented in this thesis are discussed hereafter and their figure of merit compared.

Offset-storage techniques

In the offset storage techniques, the offset is sensed and added to the signal in such a way that its impact on the decision of the comparator is minimised. There are two basic methods to accomplish this task, that can be also combined. In the first approach, called input offset compensation (IOS) the latch is driven by an amplifier which is ac-coupled to the signal source (see fig. 5). The offset compensation is obtained by closing a unity gain feedback loop around the amplifier and storing its offset on the coupling capacitors. The residual input offset is calculated to be

$$V_{OSR} = \frac{V_{OSA}}{1+A_0} + \frac{\Delta Q}{C} + \frac{V_{OSL}}{A_0} \tag{3.33}$$

where V_{OSR} is the total input referred offset, V_{OSA} is the input referred offset of the amplifier and V_{OSL} is the offset of the latch.

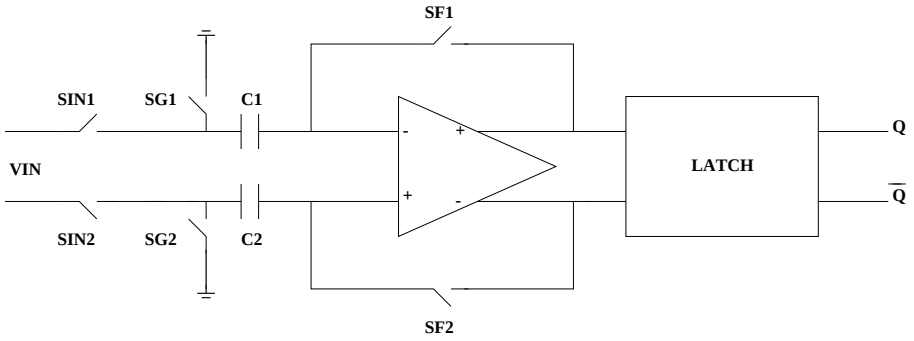

Figure 3.5: Input offset compensation

The advantage of the input offset storage is that the comparator is ac-coupled to the signal source and potentially a rail to rail input range can be achieved.

Its main drawback is that there is a residual term which is not affected by the gain of the amplifier. This term, $\frac{\Delta Q}{C}$, comes from mismatch in the charge injection between the two feedback switches S_{F1} and S_{F2}. To fully appreciate its contribution, it is enough to note that a ΔQ of 1 fC ($\simeq 6250\ e^-$) is sufficient to create a residual voltage of 1 mV on a capacitor of 1 pF. The only way to minimise the charge injection is to increase the value of the input capacitor. Moreover, to suppress the offset of the latch below 1 mV an amplifier with an open loop gain of 100 is required.

Both this remedies introduce significant delays and prevent the use of input offset storage alone in high precision and fast applications.

The complementary solution to IOS is to stored the offset in capacitor in series with the output of the amplifier, as depicted in fig. 6

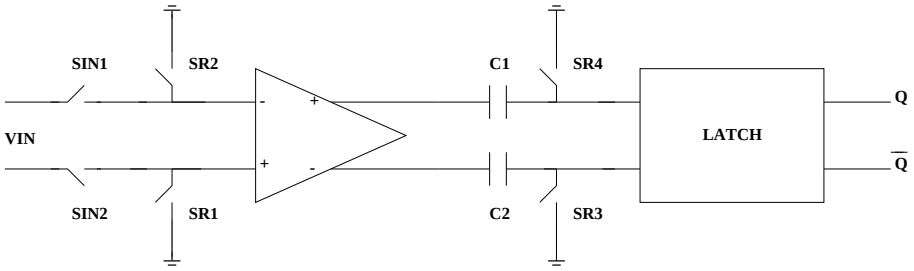

Figure 3.6: Output offset compensation

During the offset compensation phase, the inputs of the amplifier are short-circuited, and the output capacitors are grounded. The voltage difference at the output of the amplifier is then $A_0 V_{OSA}$; this value is stored in the output capacitors and can not affect the comparison anymore. The residual offset at the input of the chain is given by

$$V_{OSR} = \frac{\Delta Q}{A_0 C} + \frac{V_{OSL}}{A_0} \tag{3.34}$$

where the term $\frac{\Delta Q}{A_0 C}$ takes into account the charge injection mismatch of the switches S_{R3}, S_{R4}. As we see from equation the contribution of the amplifier offset is fully suppressed.

This second method may seem much more effective than the previous one; however we have to take into account that the amplifier always works in a open loop configuration and therefore its gain must be relatively small because otherwise the amplifier would saturate during the offset compensation phase. Hence, these implementations commonly use a gain between 10 and 20, which reduces the offset of the latch to 5 - 10 mV.

Neither of the techniques describe above is sufficient to achieve offset reduction at the level of the mV. The problem can be circumvented using more amplifiers in the chain and performing offset compensation on each stage. Fig. 7 shows an example of how this can be implemented.

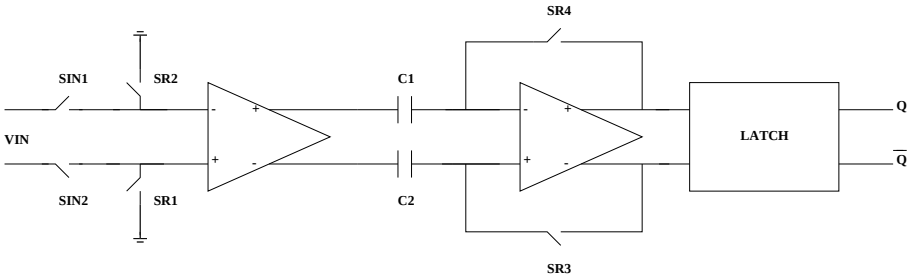

Figure 3.7: Offset cancellation using both input and output compensation

In this circuit, two stages are used, performing output offset compensation on the first stage and input offset compensation on the second one. The residual input offset of this configuration can be calculated as

$$V_{OST} = \frac{\Delta Q}{A_{01}C} + \frac{V_{OSL}}{A_{01}A_{02}} + \frac{V_{OSA2}}{A_{01}(1+A_{02})} \tag{3.35}$$

The offset of the latch is basically divided by the product of the gain of the two stages, whereas the offset of the second stage and the contribution of the charge injection are divided by the gain of the first stage, whose offset is indeed completely suppressed.

Compensation using auxiliary input port

The use of an auxiliary input port allows extremely high accuracies, usually sacrifying the speed of the compensation process. However, since offsets are DC or slowly variable signals, the compensation can be carried out only once per several comparison.

Fig. 8 depicts the principle of this compensation scheme, which requires the introduction of an auxiliary input port, here represented by M_{C1A} and M_{C1B}.

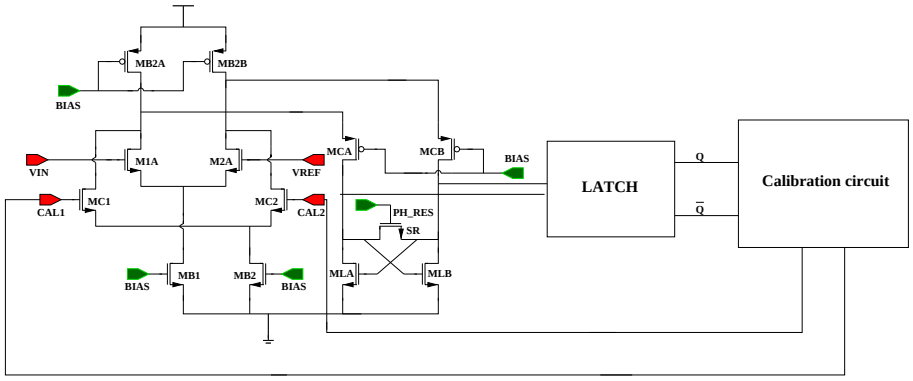

Figure 3.8: Offset compensation using a feedback loop and an auxiliary input stage

During the calibration phase, the input of the main stage are short-circuited. The auxiliary stage is driven by a calibration unit, which senses the digital output of the comparator and generates a voltage ΔV, which is applied to the M_{C1A}, M_{C1B} . When seen from the input, this voltage becomes

$$V_{err} = \Delta V \frac{g_{ma}}{g_{m1}} \tag{3.36}$$

where g_{ma} and g_{m1} are the transconductances of the auxiliary differential pair and of the main differential pair, respectively. This voltage must have a polarity opposite to the one which has determined the decision of the comparator. At each compensation cycle, the output of the comparator is sensed and a quantity ΔV is added. [1] The offset compensation is completed when the output of the comparator start flipping in opposite directions at each cycle. The residual offset referred to the input is V_{err}. With this method, comparator achieving a residual offset as low as 50 μV have been reported [27].

[1] Of course, the ΔV produced in each step must be added or subtracted to the sum of the values previously found. The analog part of the calibration unit is, in practice, an integrator, that is never reset during the compensation procedure

So far, we have supposed that all the contributions to the offset come from static mechanisms. However if we look at fig. 2, we see that the switching of the clock on transistors S_R can couple through parasitic capacitors to the arms of the latch, inducing significant voltage jumps. These jumps are in principle common mode signals, but they are converted into differential signals by the mismatch in the circuit and cause *dynamic* offset, which can be even bigger than the static offset. The servo-loop technique compensates also for the dynamic offset and is therefore very attractive when high sensitivity is a must.

The time needed to achieve a desired precision can easily be evaluated as

$$\frac{(V_{OS} - V_{OR})}{V_{err}} T \tag{3.37}$$

where V_{OS} is the offset before compensation, V_{OSR} is the residual offset after compensation and T is the clock period. Of course $V_{OSR} \geq V_{err}$. Typically, 50 clock cycles are required to reduce the offset below 250 μV.

3.3 Implementation of charge redistribution converter in submicron technologies

The implementation of a charge redistribution ADC in a scaled CMOS technology involves several considerations. The size of the DAC depends on the density of the linear capacitor used (poly-to poly or metal-to-metal capacitors), which has be found to be roughly the same (0.9 to 0.75 fF/μm^2) in many different processes from 0.7μm down to 0.25 μm.

Of course, moving towards deep-submicron technologies the densisty of the gate capacitane increases significantly (up to 6 fF/μm^2 in a 0.25 μm process) and one could think about using MOS capacitor to implement the DAC. The major drawback in this approach is that the MOS transistor used as a capacitor suffers from voltage nonlinearity, which in principle prevents its use in high resolution circuit working on the charge redistribution technique. Some methods to overcome this drawback have recently proposed and their effectiveness proved [28]. However, MOS capacitor are closer to the substrate and therefore they are more sensitive to substrate noise.

Hence, we can conclude that, if available, linear capacitor should be preferred, but in this case the size of the DAC is almost unchanged.

The situation for the comparator is more complex. If we look at eq. 22, we can conclude that the dependence of τ on L^2 makes scaling of the devices seemingly very effective in increasing the speed. In fact, using eq. 22 we can estimate that, for the same over-drive voltage, a latch implemented in a 0.25 μm process will have a time constant 23 times smaller than the same circuit implemented in a 1.2 μm technology. However, we have seen in the first chapter that in submicron devices the transconductance is limited by velocity saturation effects and that the classical expression for g_m has to be modified as following

$$g_m = WC_{ox}v_{sat} \tag{3.38}$$

where v_{sat} is the saturation velocity. The electric field at which the saturation velocity occurs can vary from 8×10^3 to 3×10^4 V/cm for NMOS transistor and from 2×10^4 to 10^5 V/cm for PMOS transistors [2]

If we assume the worst case condition, corresponding to the lowest saturation field, a device with L = 0.25 μm will be in the velocity saturation region for a drain source voltage of only 200 mV. Taking into account also this effect, the regeneration time constant becomes

$$\tau = \frac{2}{3}\frac{L}{v_{sat}} \tag{3.39}$$

and therefore the scaling effect would become proportional to L rather than to L^2.

As we have seen in the previous section the offset of the latch gives the major contribution to the overall comparator offset and is defined by

$$V_{OSL} = \Delta V_{TH} + \frac{1}{2}\left(\frac{\Delta W}{W} + \frac{\Delta L}{L}\right)(V_{GS} - V_{TH}) + \frac{\Delta Q}{C_D} \tag{3.40}$$

The threshold voltage contribution diminishes moving towards a submicron process, since it scales with the thin oxide, according to (see chapter 1)

$$\Delta V_{TH} = \frac{B t_{ox}}{\sqrt{WL}} \tag{3.41}$$

where t_{ox} is the gate oxide thickness and B can assumed to be 1 mVμm/nm in many different processes [29]. If the size of the devices and the current are kept constant, eq. 40 shows that the offset decreases. In fact, besides the reduction in the threshold mismatch, also the overdrive voltage, which gives the main contribution to the latch offset, is reduced. However, keeping the device size constant while scaling the technology increases the parasitic capacitance, since the gate capacitance increases. This in turn may worsen the speed performances, increasing the regeneration time constant. In fact, if the area of the devices is increased in the same technology or if it is kept constant implementing the circuit in a process with a smaller feature size, the effect is to decrease the overdrive voltage, with the result of decreasing the offset and reducing (see eq. 22) the speed.

For this reason the offset-delay product, defined as the product between the regeneration time constant of eq. 22 and the offset of eq. 40 is sometimes used as a useful quantity to optimize the performances of the latch [24]. After some algebra, this quantity can be written as

$$\tau \times V_{OSL} = \Delta V_{TH}\frac{2}{3}\sqrt{\frac{WC_{ox}}{2\mu I_d}}L^{\frac{3}{2}} + \frac{1}{2}\left(\frac{\Delta W}{W} + \frac{\Delta L}{L}\right)\frac{L^2}{\mu} + \frac{\Delta Q}{g_m} \tag{3.42}$$

From the above equation we can conclude that, for a *given* technology and a *given* power budget, the offset delay product is reduced using minimum length transistors; on the other hand, increasing the with improves the second term in eq. 42, which is usually dominant, but may slightly worsen the first one. Therefore, a compromise must be found and is common practice to design the transistors in the latch using the minimum length allowed by the technology and a W/L ratio of ten [24].

It is interesting to study the evolution of the offset-delay product with the scaling of the technology. For this purpose, it better to rewrite eq. 42 substituting Δ_{VTH} with eq. 41

$$\tau \times V_{OSL} = \frac{2}{3}\sqrt{\frac{B^2\varepsilon_{ox}t_{ox}}{2\mu I_d}}L + \frac{1}{2}\left(\frac{\Delta W}{W} + \frac{\Delta L}{L}\right)\frac{L^2}{\mu} + \frac{\Delta Q}{g_m} \tag{3.43}$$

This relation shows that moving to a submicron technology, the offset-delay product improves, since:

- The minimum length of the transistors becomes smaller
- For the same bias current, the overdrive voltage is reduced
- The gate oxide thickness is reduced
- The g_m is bigger.

Of course, eq. 43 holds in case there is no velocity saturation effect. If this effect has to be taken into account, the above calculations can be repeated using for g_m the expression $g_m = WC_{ox}v_{sat}$ and eq. 43 becomes

$$\tau \times V_{OSL} = \frac{2}{3}\sqrt{\frac{L}{W}}\frac{B}{t_{ox}}v_{sat} + \frac{1}{2}\left(\frac{\Delta W}{W} + \frac{\Delta L}{L}\right)(V_{GS} - V_{TH})\frac{L}{v_{sat}} + \frac{\Delta Q}{g_m} \tag{3.44}$$

We see here that the scaling effect becomes again proportional to L and that also the first term is reduced by increasing the width of the transistor. However, given the complexity of the short channel MOSFET [2], eq. 42 and 43 should be regarded only as a first approximation and more accurate evaluations must be done during the design phase with accurate models and computer simulations.

3.4 Summary

In this chapter the basic principles of the charge redistribution analog to digital conversion have been reviewed and the problem of the scaling of this architecture in submicron and deep-submicron technologies has been addressed.

It has been shown that the implementation of a switched capacitor successive approximation ADC can have some benefit from the scaling of the technology due to the improvements in the offset performance of the comparator, while the size of the DAC will stay more o less the same if metal-to-metal or poly-to-poly capacitors are used. The speed-offset trade-off of the comparator is expected to improve, thereby allowing greater accuracy and higher speed. However, the velocity saturation effect must be properly mastered in order to achieve optimum performance. Of course, the size of all the control logic will be greatly squeezed, so a moderate reduction in the overall area is expected.

4 Design and test of a low-power 16 channels charge redistribution ADC

Due to its low power consumption, the charge redistribution architecture is an attractive approach for the implementation of multi-channel ADCs and a design integrating up to 64 converters on the same die has been reported [9]. However, this circuit was limited to medium speed and resolution (8 bit over a 1.5 V full scale range and 11 MHz clock) and the realization of faster and more accurate converters entails two basic issues. On one side, in fact, in a direct implementation, the area of the binary weighted DAC doubles for each extra bit of resolution required; on the other, the parasitic inductance in series with the reference voltage can cause painful ringing, hence determining long settling times.

The problem of the resolution can be solved by a different segmentation of the DAC (see also chapter 5) and 18 bits converter based on the successive approximation techniques are found in the literature [30]. Another possibility, if a uniform resolution is not needed over the full range, is to scale the reference voltage according to the input signal [27].

The problem of the series inductance depends strictly on the characteristics of the system in which the converter is embedded, but, especially if a fast technology is used, it can be the dominant factor in limiting the speed of a successive approximation ADC.

In the work presented in this chapter we have addressed some of these issues, designing and testing a monolithic array of sixteen charge redistribution converters. In particular, in the measurements performed up to now, we have concentrated on *on chip* limitations, like cross-talk, noise and channel non-uniformity which is an important feature if the converters have to work in a multichannel data acquisition system or in a time interleaved configuration.

In the first part of the chapter the architecture of the chip is presented and some simulations on the inductive noise on the reference voltages are discussed. The second part reports the results of the laboratory measurements which show good system performances up to a clock frequency of 20 MHz, with a power budget of 3 mW/ADC.

This chip has been implemented in a 0.7 μm CMOS process.

4.1 ADC architecture

The chip contains 16 identical charge redistribution ADCs [23]. The single module consists of a 8 bits switched capacitor DAC, an offset-compensated comparator and a successive approximation register. The choice of a 8 bits resolution comes from the need of keeping the DAC within a reasonable size. However, the comparator has been designed to maintain an accuracy of 0.5 mV, in order to exploit the possibility of scaling the reference voltage down to very low values.

While the architecture of the binary weighted DAC is rather conventional (256 unit capacitors of 50 fF laid-out with a common-centroid geometry) the comparator deserves some more attention.

In order to maximise the accuracy, we used an offset compensation technique employing a servo-loop and a low-sensitivity auxiliary input port [27, 28]; the circuit is hence divided into two units, one for comparison and one for offset correction, shown in fig. 1 and 2, respectively.

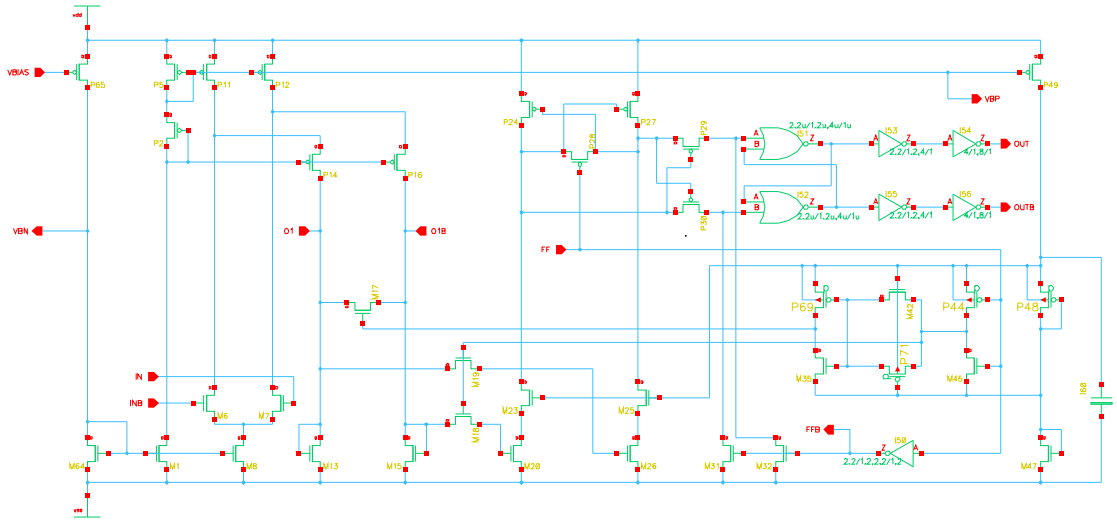

Figure 4.1: Schematic of the comparator circuit

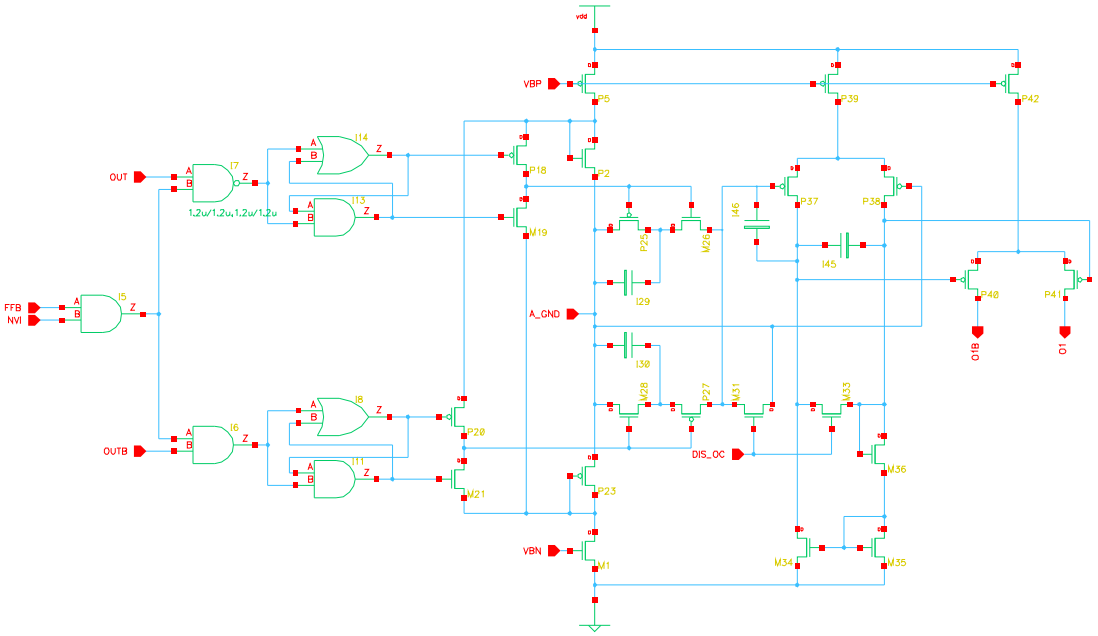

Figure 4.2: Schematic of the offset compensation unit

The comparison unit consists of a low-gain and fast folded cascode preamplifier and a positive feedback stage for signal regeneration. The two stages are coupled via the NMOS switches M18, M19, that are opened when the latch is strobed in order to prevent kickback noise from the latch into the input stage. Cascode transistors M23 and M25 in the second stage serve to

this purpose as well. The NOR gates I51 and I52 provide fully amplification of the signal to CMOS levels, while the inverters I53 - I56 buffer the two complementary digital outputs. The comparator is strobed by the low-to-high transition of the clock signal FF.

The calibration circuit has a continuous time part, composed by P37 -P41, M33 -M36 and capacitors I45, I46 and a synchronous part, composed by all the other transistors and logic gates visible in fig. 2. The non-valid-input signal (NVI) enables the offset compensation procedure. When this signal is asserted (NVI = 1), the clock FFB (where FFB = NOT FF) is transmitted to the circuit. The NVI also stops all the operations in the ADC and short-circuits the inputs of the comparator to the analog ground. In these conditions, the output of the comparator is tested and the decision is determined by the polarity of the offsets. Let us suppose that, after the first test, OUT = 1 and hence OUTB = 0. It is easy to see from fig. 2 that in this conditions P27 is off, M28 is on and the capacitor I30 is disconnected from the input of the differential stage P37 - P38 and short-circuited to the analog ground. If FFB = 0, the same conditions holds also for capacitor I29.

When FFB has a low-to-high transition, M26 is switched on, connecting I29 to the input of the differential stage, whereas P25 is switched off, injecting a small amount of charge into I29. This charged is processed by the integrator (formed by P37 - P38, M34 - M36 and I46) and converted by P40, P 1 into a small current, which in inserted into the comparator trough the nodes O1, O1B.

At each clock cycle, a small quantity ΔI is produced by the calibration circuit. This current is seen from the gates of transistors M6 - M7, as a voltage $\Delta V = \Delta I/g_m$, where g_m is the transconductance of the differential input pair. As discussed in chapter 3, ΔV must have a polarity opposite to the one of the offset which has determined the decision of the comparator.

The above considerations can be repeated for the case OUT = 0 and OUTB = 1. In fact, it easy to see that in this situation I29 would be permanently connected to ground and the charge would be injected in I30, thereby determining a compensation signal with opposite polarity with respect to the case OUT = 1. The compensation is completed when the output of the comparator changes at each clock cycle and the residual offset is equal to the incremental voltage ΔV, which is defined by the following equation

$$\Delta V = \frac{\Delta Q}{C}\frac{g_{mP40}}{g_{m1}} \tag{4.1}$$

where ΔQ is the charge injected by the switched P25 (or M28) and g_{mP40} and g_{m1} are the transconductances of the differential pairs P40 - P41 and M6 - M7, respectively. Therefore, to get an accurate compensation, the ratio $\frac{g_{mP40}}{g_{m1}}$ should be as small as possible. In our implementation, this ratio is 1/500 and the value of all the capacitors in fig. 2 is 1 pF.

4.2 Inductive noise problems in charge redistribution converters

The switching of the capacitors in a charge redistribution DAC draws from the reference voltage significant currents that reach their peak value in few nanoseconds [24]. If an inductance is present is series with the reference, a significant ringing can result; while the basics of the phenomenon are easily understood with a trivial RLC series circuit, their exact impact on the accuracy of the conversion is quite difficult to determine.

In fact, the simulation of the transient behaviour of an array of converters with an analog simulator is impractical since it would require very long CPU times. Even more important, is not so easy to model all the parasitics that may contribute to the final result with a precision adequate to predict the effective resolution of the ADC.

We have therefore used a simplified approach, simulating only the critical blocks, i.e the DAC and the switch network and including all the on-chip parasitic capacitance estimated by the CAD tool. An inductance in series with the reference voltage has been introduced as the free parameter to emulate various experimental situations. The worst case occurs when the ADCs are working in parallel mode (i.e. the bits of equal weight are evaluated simultaneously) and the MSB is tested. For brevity only the simulations inherent to this situation are shown in the following for the case of sixteen DACs, which corresponds to the circuit actually implemented.

Since in this design we used a unit cell of 50 fF, the total capacitance of a single DAC was 12.8 pF; the switches driving the bottom plates were binary scaled according to the value of the capacitors and the equivalent resistance of the MSB switch was 20 Ω. Therefore, the value of the redistribution time constant (see chapter 3) is 0.13 ns, which allows a 8 bits settling in 0.8 ns, as shown in fig. 3.

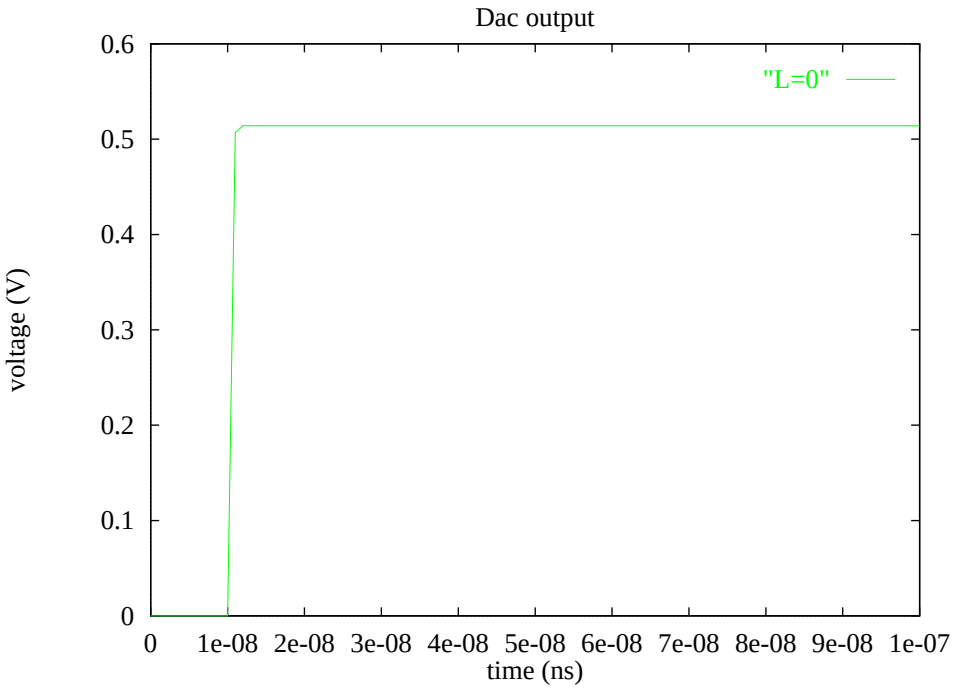

Figure 4.3: Simulation with no series inductance

In the simulation reported in fig. 4, an inductance of 2.5 nH (corresponding to the parasitic inductance of a bonding wire [9]) has been inserted in series with the reference voltage. The effect on the settling time is apparent and in this situation a settling to the 8 bits level is reached only after 15 ns. As a consequence, if we suppose that a full clock period is used half for the settling of the DAC and half for the completion of the comparison, which is usually the case, we can easily calculate that the clock frequency can not go beyond 33 MHz.

The situation further worsens when also the PCB trace is taken into account; the sum of the contribution of the PCB trace and of the bonding wire leads to the disatrous behaviour of fig. 5, where the settling is achieved only after 95 ns, thereby limiting the overall clock frequency

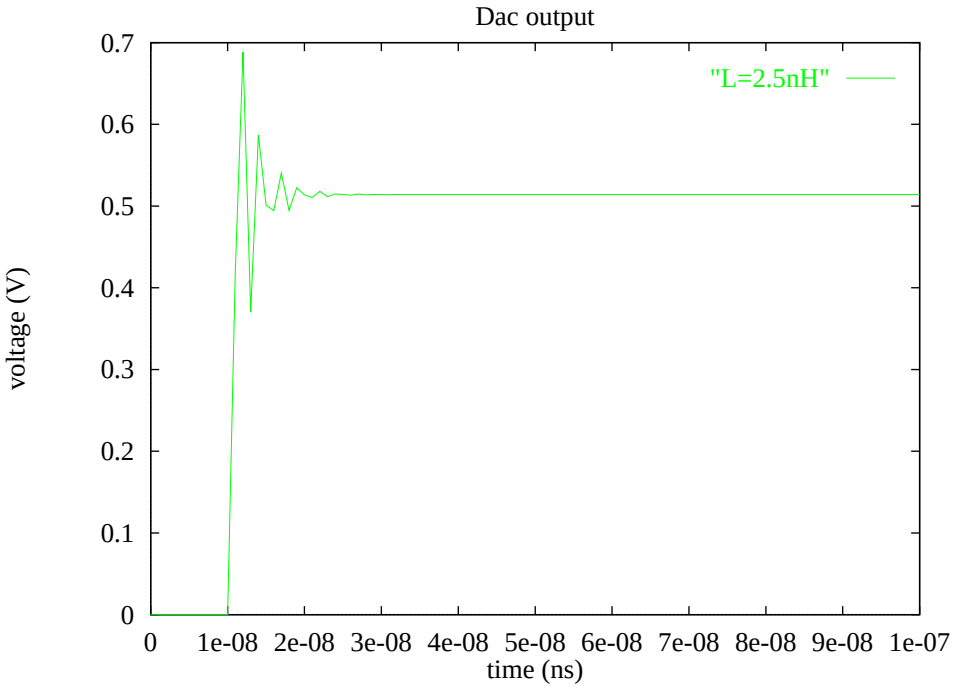

Figure 4.4: Simulation with 2.5nH series inductance

to 5 MHz. Actually, some simple remedies can be introduced to alleviate this problem. One possibility is to use more reference lines on the PCB and more bonding pads on the chip. Fig. 4 depicts the case in which four independent lines are used to bring the reference voltage to the chip. The settling time drops from 95 ns to 20 ns, allowing a clock of 25 MHz in the redistribution phase. Another alternative is to use a voltage regulator very close to the chip, in order to minimise the length of the trace seen from the converters. However, these solutions are not always practical. For instance, we have see in chapter 2 than in the front-end electronics of the Silicon Drift Detectors of the ALICE experiment, a number of charge redistribution converters are foreseen. In this case, given the severe constraints on power and material budget, the use either of an on board voltage regulator or of multiple PCB paths would be a serious issue. On chip voltage regulator can be also considered, but in this case to meet high speed requirements usually an high power dissipation is needed [24].

Of course, a trivial way to dump the oscillations is to insert a resistance in series with the reference voltage. Fig. 5 shows the case in which the switched have been sized to give a resistance 10 times higher than in the previous simulations. The total parasitic inductance was still 12.5 nH and a 8 bits settling is achieved in 14 ns.

In our design, since we were more interested in investigating the intrinsic limitations due to the technology and the particular architecture, we have sized the switches to give the faster time constant (i.e. 0.13 ns); four bonding pads have been used for the reference voltages and during the test phase, care has been taken in minimising all the parasitic effects.

4.3 Digital controls and chip layout

The chip has been provided with simple digital facilities in order to simplify the interface with the data acquisition system. Each of the sixteen ADCs is connected to a 8 bit register, in which

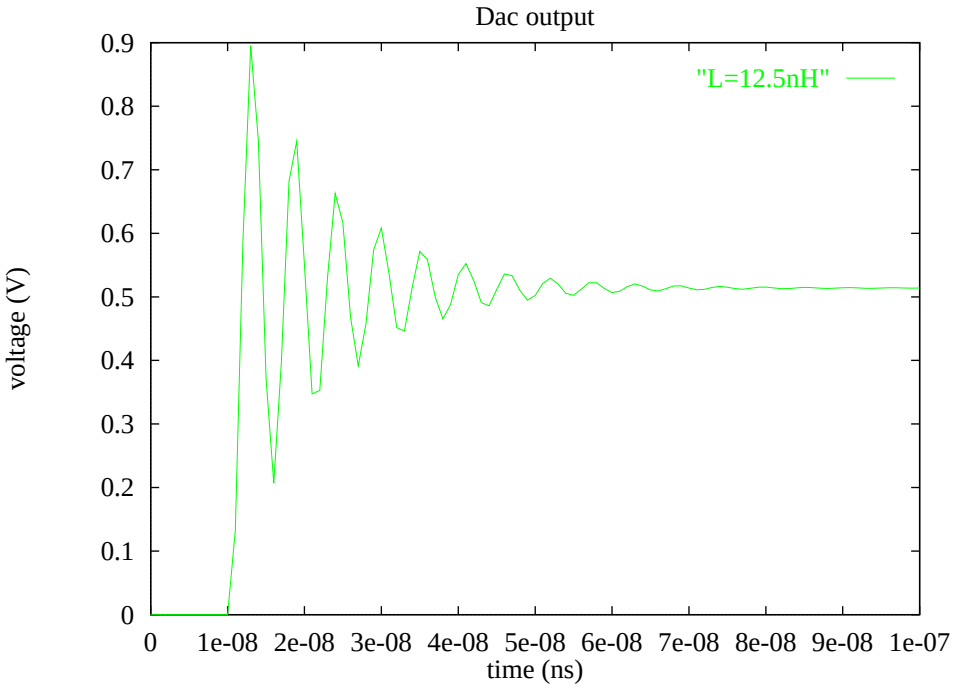

Figure 4.5: Simulation with 12.5nH series inductance

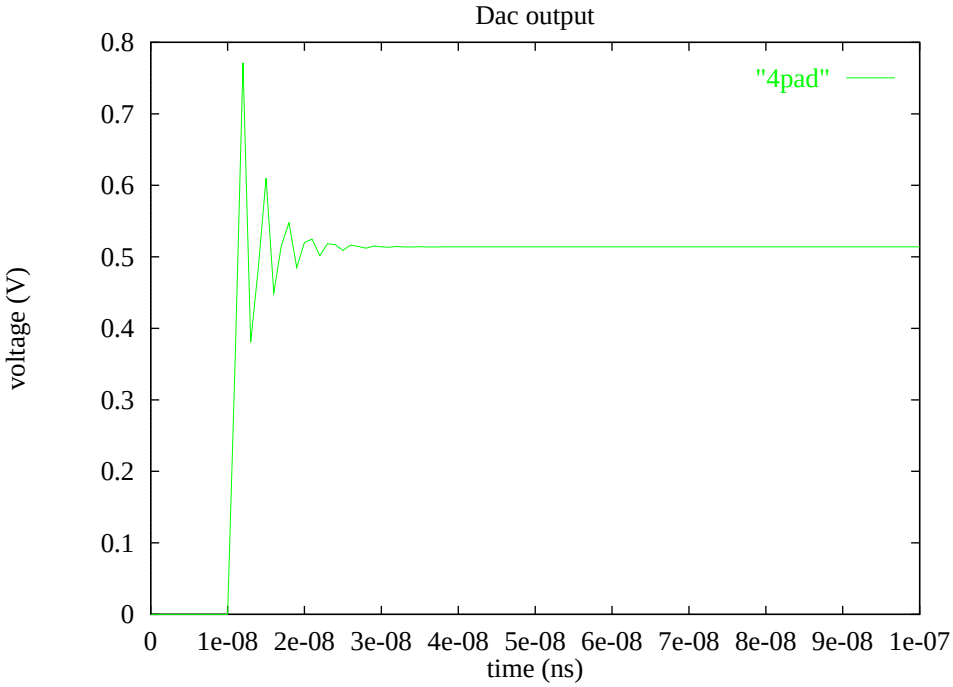

Figure 4.6: Simulation with four reference lines

the result of the conversion is stored. The data are then transferred to the output pads via a multiplexer. During the the design phase, the ADC and control logic were simulated using the mixed-mode simulator hspiceS-Verilog. In this way, since all the digital blocks are represented by behavioural models, a complete conversion cycle can be simulated within a reasonable time

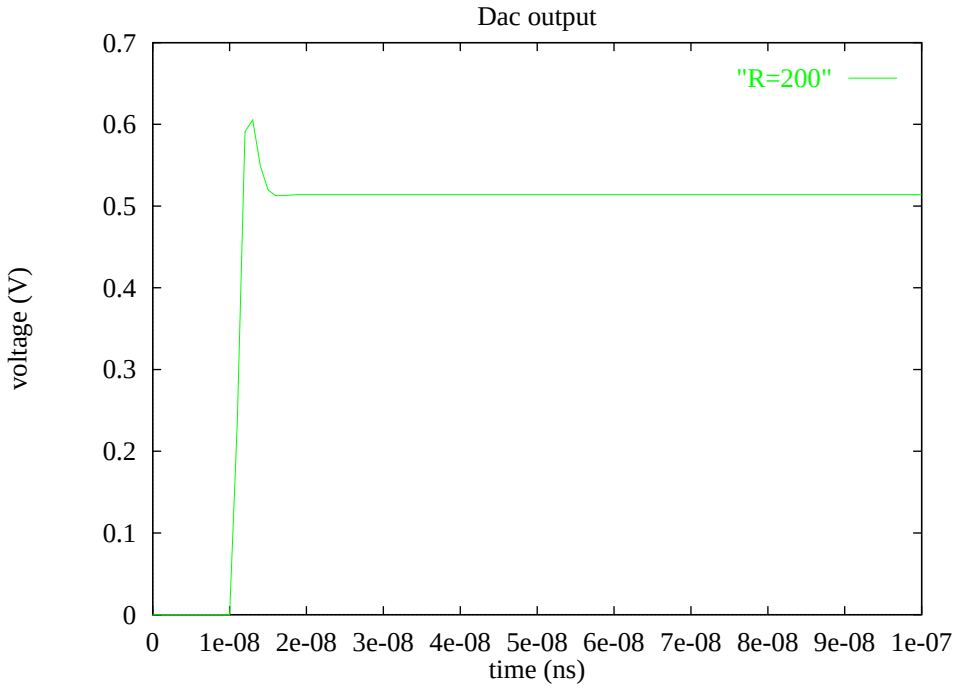

Figure 4.7: Simulation with a switch resistance of 200 Ω

(about 15 minutes of CPU time on a Ultra Spark processor with 1 Gbyte of RAM) and most of the ADC codes could be explored.

For the digital parts we have used CMOS standard cells from the silicon foundry. As we have seen in chapter 1, CMOS logic is not the best choice for a mixed mode chip; however, the design of full custom digital cells adds a considerable time over-head. We have not considered this options since our frequency of interest in this design are at most 20 MHz e the amount of switching noise should be limited. In fact, all the digital logic is not active while the ADCs are in the sampling mode and only one shift register per chip and one SAR per channel are used during the conversion. The internal fully differential structure of the comparator should help to reject common mode noise. Nevertheless, we have carefully splitted the digital and the analog power supplies and grounds and the digital output buffers, which can drive considerable peak current, have been put as far as possible from the analog parts. Also the layout of the digital blocks has been hand-crafted.

Each ADC is provided with three independent input ports, one for the input signal, one for the reset of the offset compensation circuit and one for disabling the converter. In this way each converter can be test while the others are switched off and the effects due to simultaneous operation can be clearly identified. The single ADC occupies an area of 500 x 500 μm^2 and the sixteen ADCs have been laid-out in 8 x 2 matrix, occupying an area of 1.5 x 6 mm^2. The layout of the chip is depicted in fig. 8. The ADC matrix is visible on the left. The analog input pads are on the left of the die, the digital pads on the right and the power pads on the top and the bottom. All the powers pads are double and two pads per side are provided for the reference voltage. The total die size is 30 mm^2.

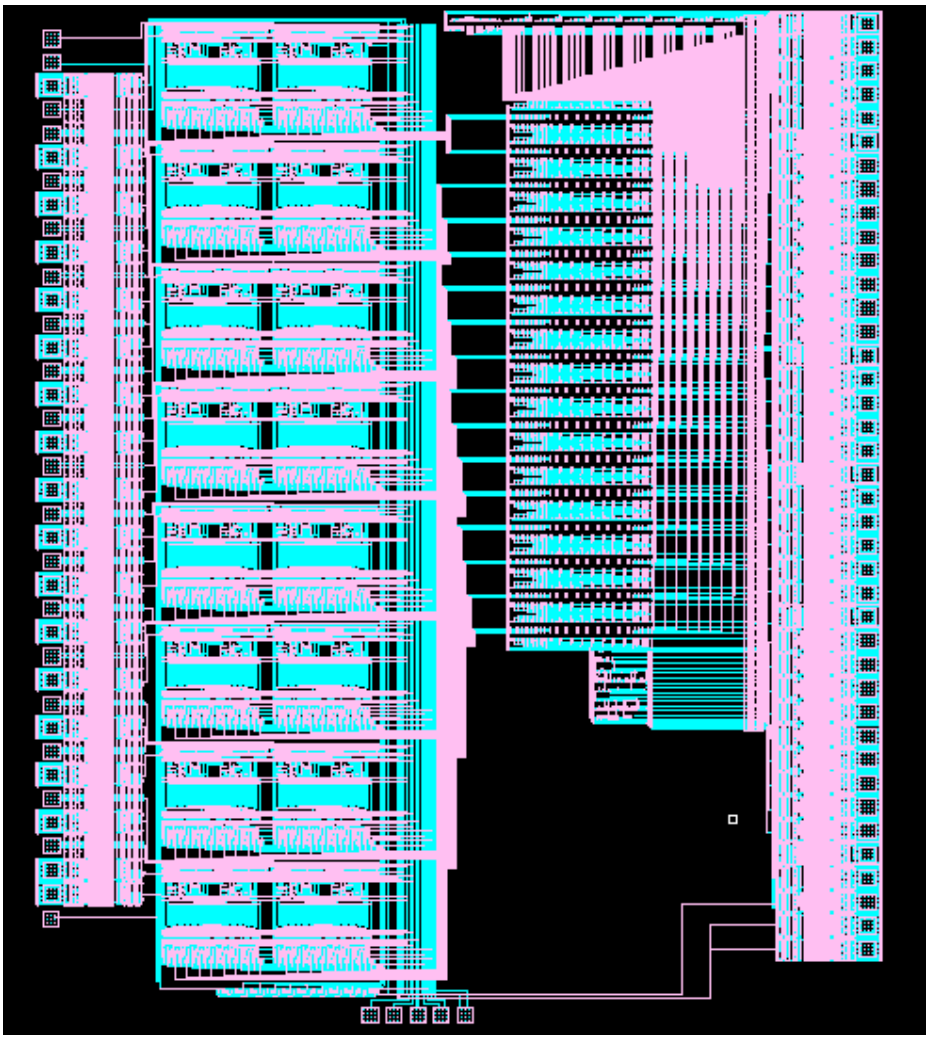

Figure 4.8: Layout of the complete chip

4.4 Considerations on the test of analog-to-digital converters

The test of high-resolution analog to digital converters is a difficult task, since very small signal have to be discriminated and the accuracy of the testing equipment has to be better than the accuracy of the ADC under test.

In many commercial data-sheet, the measured specifications of the ADC are given only when the input is a DC voltage; this is misleading, because all the dynamic limitations of the analog to digital converter, due to slew rate limitations, settling time, aperture uncertainty remain unexplored.

The first step in the test of an analog to digital converter is therefore the choice of a suitable *dynamic* signal. A signal that is very often proposed is the voltage ramp, since an ideal ramp with various slope would be suitable for testing both static and dynamic performances. Unfortunately, high frequency ramps of very good quality are difficult to generate and very often the accuracy is no better than 1%, which is inadequate even for testing a seven bits converter. Though less intuitive, a sinusoidal signal is much more suitable for the following reason:

- A sinusoidal signal with high accuracy is easy reproducible; very good sine wave genera-

tors with second harmonic distortion down to -95 dB are found on the market. Therefore the results of the test can easily be reproduced by more independent users

- A sinusoid is easy to describe and manipulate mathematically, so the analysis of the data is simplified
- A sinusoid is bidirectional and points out problems depending on the sign of the slope of the input signal.

For a complete characterisation of the converter it is important to use full scale signal, in order to explore all the codes of the ADC. In fact, the performances measured with a smaller signal can to be reliably extended to the full scale range, since slew rate limitations have a different impact on small an big signal.

Using a sinusoid as the input signal, main three different tests can be carried out. These tests are shortly described in the following.

4.4.1 Histogram testing

In the histogram test a full scale sine wave is used as the input signal for the ADC. It is important to note that the ratio between the sampling frequency of the converter and the frequency of the input signal should not be a rational number. In fact, if the ratio between the periods of the two waveform is a rational number, i.e. T_{in}/T_{ck}=m/n, $nT_{in} = mT_{ck}$ is an integer multiple of the period of both waveforms, which "beat" every nT_{in} second. Hence a "coherent sampling" results. In particular, if nT_{in} is small only few codes are examined, giving misleading results. Once the sampling frequency and the input signal frequency have been properly chosen, a set of several hundred thousands samples of the input signal must be taken and stored on a computer. A the end of the acquisition, an histogram is plotted with the possible codes along the x axis and the effective occurrence of each code along the y axis. The results obtained from the ADC under test must be compared to the ones given by the ideal ADC. In fact, for a sine wave input, a perfect ADC would produce a distribution described by the following probability density function

$$p(V) = \frac{1}{\pi\sqrt{A^2 - V^2}} \tag{4.2}$$

where A is the peak amplitude of the sinusoid and $p(V)$ is the probability of occurrence of the voltage V.

A real ADC exhibits differential nonlinearity, so not all the quantisation steps have the same size, as it would be for an ideal converter. The histogram test provides an immediate grasp on the performance of the ADC, because codes associated with large steps will occur more frequently than codes associated with smaller steps. Missing codes are also easily detected.

4.4.2 Fast Fourier Transform Test

In the FFT test, the output of the ADCs is studied in the frequency domain. Since the output of the ADC is discrete in time, the conversion between the time domain and the frequency domain is achieved by mean of a Discrete Fourier Transform algorithm. This test is particularly useful for evaluating the integral nonlinearity of the ADC. In fact, if the ratio between the fundamental and the highest harmonic in the spectrum is highest than 6N dB, it is legitimate

to say that the error contribution of the integral nonlinearity is negligible, since the amplitude of the harmonic is smaller than the least significant bit. However, even very low spurious harmonic can determine encoding errors, by forcing a voltage near a threshold level into the adjacent quantisation step.

4.4.3 Sine wave fitting

In the sine wave fitting test, a full scale wave of given frequency is digitized by the ADC. A computer program calculate best fit to the data using least squared errors minimisation techniques. The idealized sine wave is then digitized by a software which emulates an perfect ADC with the same resolution of the device under test. The actual rms error between the best fit and the actual data is calculated. The ideal rms error between between the idealized data produced by the perfect ADC and the sine wave is calculated as well. The two errors are used to define a performance metric which is called effective number of bits (ENOB)

$$ENOB = N - log_2 \frac{rms_A}{rms_I} \tag{4.3}$$

where rms_A is the actual rms error including all the nonidealities of the real ADC (INL, DNL, missing codes, noise, etc) and rms_I is the rms error of the perfect ADC, only due to the unavoidable quantisation error.

In order the sine wave fitting test gives reliable results, three conditions must be satisfied:

- The number of data points has to be large
- The frequency of the test input signal must not be harmonically-related with the sampling frequency of the ADC. In fact, if this condition holds, some codes occur more than others and if these codes are good codes, an overestimation of the ADC performances result. Moreover, a correlation between the two frequencies can alias certain harmonics back onto the fundamental, seemingly rising the signal-to-noise ratio
- The input signal should be full scale, in order to explore all the codes and to test also slew rate and other dynamic limitations

4.5 Test results

4.5.1 Test set-up

In the tests, the chip has been bonded unpackaged on a dedicated printed circuit board of 9 x 7 cm^2; in the layout of the pcb care has been paid in minimising the cross-talk between the digital and the analog lines.

The power supplies have been filtered with passive components, while a ZRA124F01 zener regulator has been used to stabilise the reference voltage. The clock and digital control signals have been generated with a data pattern generator (DG 2020 by Tektronix) and the outputs have recorded with a logic state analyzer (Tektronix Prism series 3002).

An arbitrary waveform generator (Tektronix AW 2020) has been used to generate the input signals, whose frequency contents have been verified with a spectrum analyzer (Tektronix 495P). The system was controlled by a PC running a dedicated LABVIEW program.

4.5.2 Test procedure

The chip has been powered with the nominal power supplies and all the test have been performed at a *fix* analog power budget, estimated to 2 mW. This is essentially the power drain by the comparator, which is the only part requiring a DC current bias.

Unfortunately, due to limitations in the data pattern generator, a full offset compensation cycle had to be carried out before each conversion, thereby slowing down the effective sampling rate to 200 kHz. However, since the clock was set to 20 MHz, the effective conversion time was 400 ns and only two clock cycles were allowed for sampling the input signal. A sinusoid with a frequency of 5.5 kHz and amplitude equal to the full scale has been used as the input signal in the tests presented hereafter. All the plots report the FFT and the INL and DNL profiles. For clarity the plots are accompanied by a short discussion.

4.5.3 Optimum offset compensation time

In the comparator of this ADC we have used an offset compensation technique employing a servo-loop and an auxiliary input stage. As we have seen in chapter 3 with these structures the compensation time depends on the correction voltage generated at each step, which is our case is defined by eq. 1. As we see from eq. 1 this term depends on a parasitic charge injection, ΔQ, which is difficult to determine accurately in the simulations. Therefore, the first measurements aimed at finding the optimal compensation time. The measurements have been done on a typical channel of the converter. The test conditions were the following:

- Clock frequency 20 MHz (i.e. conversion time: 400 ns
- Reference voltage: 1 V (LSB=3.9 mV)
- Sampling frequency: 200 kHz
- Input signal: 1 V peak-peak sinusoid with a frequency of 5.5 kHz

In each measurement, the time devoted to the offset compensation has been changed, starting from a minimum of 10 clock cycles up to a maximum of 65 clock cycles. The results of this measurements are shown from fig. 9 to fig. 11. On the basis of these results, an offset compensation time of 65 clock cycles has been chosen.

4.5.4 Measurements with different clock frequencies

In this measurements (shown from fig. 12 to fig. 14) we changed the clock frequency to the ADC; in fact, the clock frequency fixes both the conversion time (equal to eight clock periods) and the acquisition time (set in these measurements to 2 clock cycles). The condition of the measurements were the same as for the previous test, with an offset compensation time of 65 clock cycles. The ADC shows excellent performances up to a clock of 20 MHz, while a degradation starts appearing from 30 MHz above.

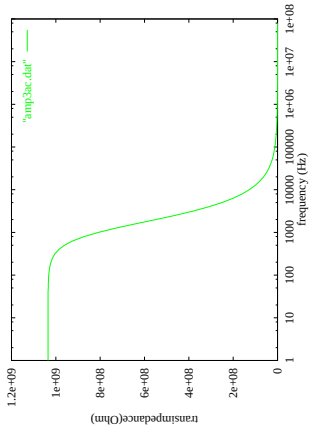

Figure 4.9: Test with an offset compensation time of 10 clock cycles

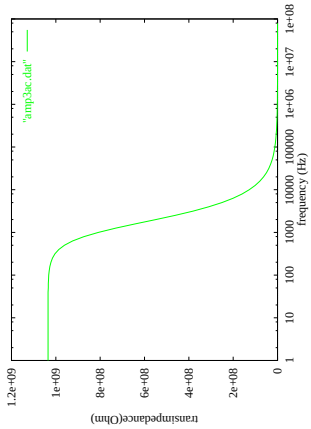

Figure 4.10: Test with an offset compensation time of 50 clock cycles

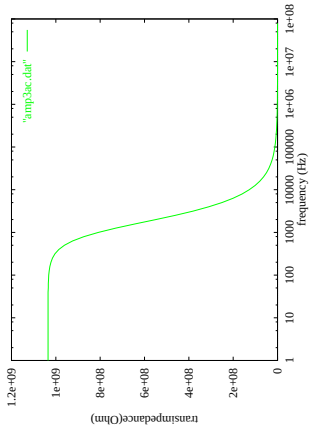

Figure 4.11: Test with an offset compensation time of 65 clock cycles

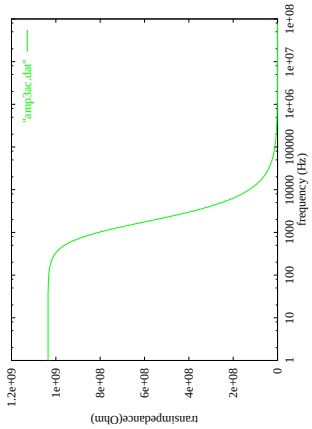

Figure 4.12: Test with a clock frequency of 10 MHz

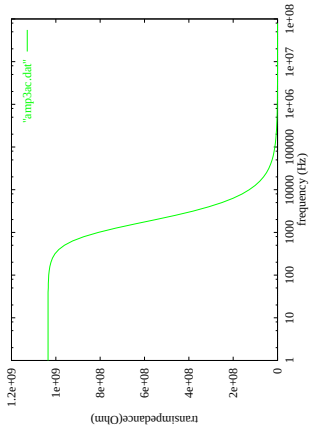

Figure 4.13: Test with a clock frequency of 20 MHz

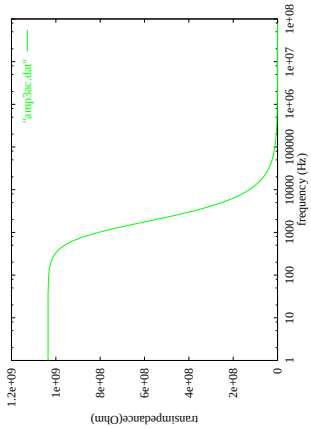

Figure 4.14: Test with a clock frequency of 30 MHz

4.5.5 Tests scaling the voltage reference

The dynamic range of a converter can be increased by adapting the voltage reference to the signal to be digitized. This feature is especially useful in applications in which a large dynamic range is combined with not to high precision, as we have seen is the case of an high energy physics experiment. The minimum voltage difference that the comparator can correctly discriminate in the useful time poses a limit on the scaling on the voltage.

We have performed this in the following conditions

- Clock frequency 20 MHz
- Offset compensation time: 65 clock cycles
- Sampling frequency: 200 kHz
- Input signal: sinusoid with a frequency of 5.5 kHz and a peak-peak amplitude adjusted to the full scale range

Table 4.1: Reference voltages and corresponding values of the LSB

Reference voltage	LSB
1 V	3.9 mV
0.5 V	1.96 mV
0.125 V	0.49 mV

The results of these measurements are presented in the plots from fig. 15 to fig. 17. A we can see from these figure, the converters operates as a true 8 bits converter down to a reference voltage of 0.5 V. For the case in which the reference voltage is 0.125 V (fig. 17), we note that the Differential Nonlinearity is still below 1 LSB, while the Integral Nonlinearity ranges from -1.5 to +2 LSB. This means that the converter has no missing codes (DNL$<$1 LSB), but some distortion must be expected. In fact, the FFT plot shows a second harmonic at - 39 dB, that reduces the linearity of the circuit to the level of an ideal quantizer with 6.5 bits of resolution.

4.5.6 Uniformity measurements

This test was intended to see the uniformity between different ADCs on the same chip. The tests have been carried-out with a reference voltage of 1 V, a input sinusoid of 1 V peak peak and 5.5 kHz frequency. The clock of the ADC was 20 MHz. All the ADCs on one die have been measured, and, besides the usual FFT, DNL and INL tests, also the sinusoidal fit has bee done, in order to measure the effective number of bits. The ENOB ranges from 7.5 to 7.8 bits, showing excellent uniformity between the channels and full 8 bits resolution for all the converters.

4.5.7 Cross-talk and noise measurements

Cross talk

The problem of the interference between the different ADCs is one of the major issue. The cross-talk has been measured on three adjacent channels, sending one sinusoid of 1 V peak-to-peak in the lateral channels and measuring the output of the central one. We performed three

different sets of measurements, changing the frequency of the input signal from 5 to 70 kHz. A cross-talk of 3 LSB was observed, with no remarkable dependence on the frequency of the input signal.

Another test has been performed, sending the sinusoidal signal only to one converter, while all the other channels on the chip were converting a DC level. Using the Non Valid Input (NVI) signal, a different number of converters was activated in each measurement. No difference has been seen between the case in which only one ADC is activated and the case in which all the channels are working simultaneously. This shows that the internals cross-talk between the channels, as well as the loading of the reference voltage is negligible and does not affect the conversion.

Therefore, the observed cross talk occur-ed either between the traces on the PCB or between the input lines of the converter. The estimations of the parasitic capacitance on the layout suggests that the capacitance between the input lines is to small to justify the observed cross talk and a significant contribution should come from the PCB traces. In fact a measurement on a PCB mounting all the component but the ADC showed a cross talk between the input lines of 1%, which is the big fraction of the total observed cross talk (1.2%) of the full scale range).

Noise

In an ideal ADC, the transition between two adjacent codes should be sharp. However, in a real circuit the transition is affected by the noise and does not occur always at the same point. The measurement has bee done putting a DC voltage at the input of the converter and plotting the relative occurrences of each code as a function of the input signal. The noise has been considered as the voltage range in which the occurrence of a given code jumps from 10% to 90%. An example of the plots obtained in these measurements is shown in fig. 18. The maximum observed noise was 3 mV, but the input signal was already affected by a noise of 2 mV; hence a maximum noise of 2 mV (equal to 1/2 LSB) can be attributed to the ADC.

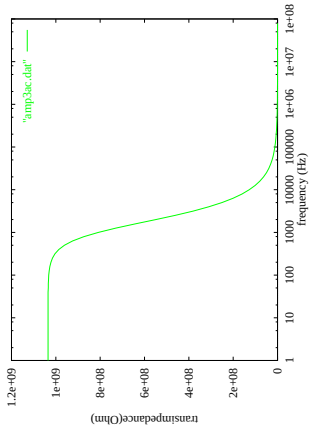

Figure 4.15: Test with a reference voltage of 1 V

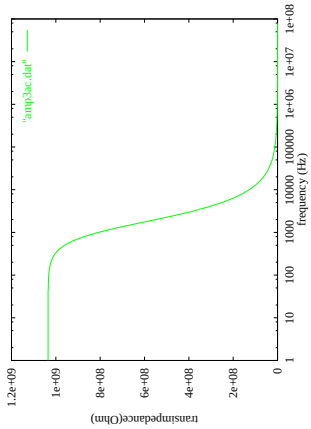

Figure 4.16: Test with a reference voltage of 0.5 V

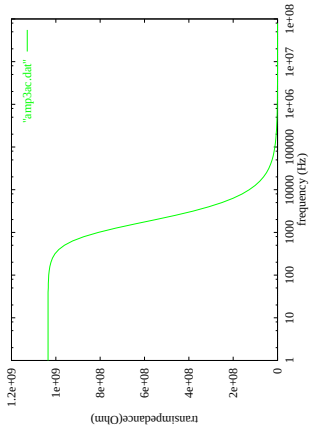

Figure 4.17: Test with a reference voltage of 0.125 V

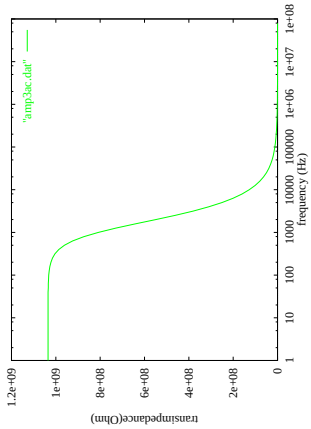

Figure 4.18: Example of noise measurements on three different codes

4.6 Summary

This chapter has discussed the design and test of a sixteen channels charge redistribution ADC, developed in the framework of the SDD front-end project for the ALICE experiment. The aim of the work was to investigate the feasibility of an integrated array of successive approximation ADCs with a resolution , speed a power suitable for the experiment. The chip is based on a 8 bit DAC coupled with a comparator whose offset is compensated with a precision technique. The combination of these two blocks allows a compact design and provides the opportunity of achieving a large dynamic range by scaling the reference voltage used for the quantisation of the signals. Since the power that can be dissipated in the experiment is very limited, the characterisation has been done with a fixed power budget of 3 mV per ADC. In these conditions, very satisfactory performance was measured with a clock frequency of 20 MHz and allowing a sampling time of 100 ns. Therefore, the total time required by each conversion was 500 ns, which is adequate for the specifications of the experiment. In fact, 16 ADCs will be sufficient to convert 64 channels of the analog memory within a dead-time of 512 μs. No interference has been seen due to the simultaneous switching of the ADC and to the loading on the reference voltage. A a cross talk has been observed between the inputs, but it has been traced to coupling between the lines on the printed circuit board. The converters show excellent uniformity; the performance is not degraded by scaling the reference voltage down to 0.5 V. With a reference voltage of 0.125 V a 8 bit resolution is obtained with a 6.5 bits linearity, making the part suitable for multi-range operation. The converter is powered from a 5 V power supply and all the specifications have been attained within the given power budget of 3 mV/channel.

5 Design and test of a charge redistribution ADC in a 0.25μm CMOS technology

The design of the converter presented in this chapter has been carried out under the CERN RD49 project, aiming at a comprehensive investigation of the use of standard CMOS technologies in radiative environments. In fact, the thin gate oxide inherent in modern deep submicron CMOS processes[1] combined with dedicated layout techniques strikingly enhances the resistance of the circuits to total dose radiation [32]. The main purpose of this design was to demonstrate the effectiveness of our rad-tolerant approach on switched capacitor circuits, which are fundamental building block in data acquisition system for high energy physics[2].

Unfortunately, deep submicron technologies operate with a power supply of at most 2.5 V and due to the tight constraints on space and power dissipation typical of high energy physics experiments, conventional switched capacitor circuits can not be easily replaced with alternative architectures more suitable for low voltage operation.

Therefore, the investigation on experimental cases of the limitations induced on switched capacitor building blocks by the reduction of the power supplies is also important. The choice of a charge redistribution ADC as a test vehicle stems from the fact that this circuit has only two critical blocks and failure sources can be easily identified. Additionally, this kind of converter is needed in the implementation of some detector front-end presently under design, like, for instance, the front-end chip of the Silicon Drift Detectors described in chapter 2. In this project, we have target a resolution of 10 bits (which usually can be attained without the need of cumbersome calibration procedures) and a conversion speed of 250 ns (40 MHz) clock with a power budget of 1 mV.

The first part of the chapter deals with the design of the ADC, focusing in particular on possible limitations coming from the switches and on the design of the comparator. The measurements before and after irradiation are detailed in the second part.

5.1 Switch limitations

The implementation of charge redistribution analog to digital converters in a deep submicron technology entails one fundamental issue. In fact this circuit, as any other conventional switched capacitor circuit, suffers from the reduced power supply which limits the over-drive voltage of the switches. We can start to study this aspect by calculating the on resistance of a NMOS switch in the simple case in which the source is at ground potential, i.e. the bulk-source voltage is zero. In this situation the on resistance is given by

$$R_{ON} = \frac{L}{KW(V_G - V_{T0})} \tag{5.1}$$

[1] With the term "deep submicron" we refer here to technologies with a minimum gate length of 0.25 μm ore less

[2] The feasibility of continuous time circuits has also been demonstrated in the framework of the same RD project [33]

where W and L are the width and the length of the transistor, V_{T0} is the threshold voltage and K is the technological parameter ($K=\mu C_{ox}$, with μ mobility of the electrons and C_{ox} the gate capacitance per unit area).

For a 0.7 μm technology K can be assumed to be 95 $\mu A/V^2$, V_{T0}=0.75 V and for a minimum size transistor and a power supply of 5 V, eq. 1 gives an equivalent resistance of 2.5 kΩ. If we repeat the calculation for a minimum size transistor implemented in a 0.25 μm technology with K = 250 $\mu A/V^2$, V_{T0}= 0.5 V and a power supply of 2.5 V, we find that R_{on} is 2 kΩ, so the potential advantage of a better K is partially wasted by the reduction in the power supply. Of course, the area of the switch is much smaller in a quarter micron technology (about eight times), so, for the same silicon area, the conductance is bigger in the process with the smaller feature size.

However, the switches have often to operate with their terminals not at a fixed potential; this is, for instance, the case of the sampling cell of fig. 1. In this case eq. 1 must be modified as

$$R_{ON} = \frac{L}{KW} \frac{1}{[V_G - V_{in} - V_{T0} - \gamma(\sqrt{2|\phi_F| + V_{in}}) - \sqrt{2|\phi_F|})]} \tag{5.2}$$

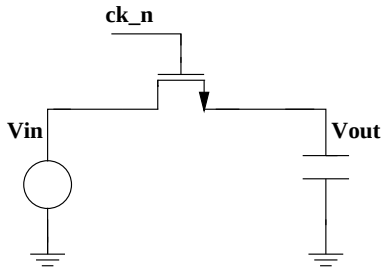

Figure 5.1: Elementary sampling cell with NMOS transistor. The bulk terminal, connected to ground, is omitted

where γ is the bulk-effect coefficient introduced in chapter 1 and ϕ_F is the Fermi potential. Since the source and the bulk are not at the same voltage, the bulk effect can rise the threshold voltage V_{T0} from 0.5 V up to 1 V and if V_{in} is greater than 1.5 V the transistor is switched-off . Hence, in a quarter micron technology, the systematic use of complementary switches is mandatory to achieve wide dynamic range. However, even this topology fails if the power supply is less then the sum of the threshold voltages of the devices. In this case, a conductive path between the two terminal of the switch for any value of the input signal is not assured anymore. The limit at which this effect start appearing is fixed by some authors at 2.4 V [34], which would make a conventional switched capacitor circuit hardly feasible in a 0.25 μm technology. Though this limit seems exceedingly conservative, care must be paid in the design of the switches, especially if their are placed in the signal path. In fact, as we see from eq. 2, the resistance of the switch depends also on the input signal and may change significantly while the input span the whole dynamic range.

In a charge redistribution converter, (see, e.g., fig.) the switch connecting the top plates of the capacitors of the DAC to the reference voltage is closed only during the acquisition mode and operates always at a fixed potential; hence its resistance does not depend on the input signal and introduces bandwidth limitation, but not signal-dependent harmonic distortion. The switches driving the bottom plates operate between fixed potentials during the redistribution mode and are connected to the input signal during the acquisition mode. Therefore, during the sampling,

their resistance depends on the value of the input signal and can be source of nonlinearity. A simple model to evaluate the contribution of the switches to the signal distortion is depicted in fig. 2; here the capacitor represents the total capacitance of the array.

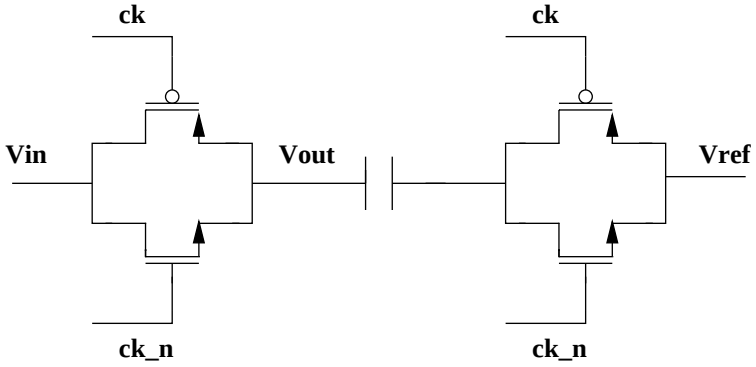

Figure 5.2: Model of the charge redistribution DAC in the sampling mode

As an example, in the following are reported the results of some simulations done to evaluate the effect of the scaling of the power supply. In these simulations, the switch connected to the input signal was sized to 200/0.36 and the switch connected to the reference voltage to 26/0.36. A reference voltage of 1.024 V was used and the input signal was a sinusoid with a frequency of 2.5 MHz and an peak-peak value equal to 90% of Vref. Each plot represents the difference between the signal sampled on the capacitor V_{out} in fig. 2 and and the input signal; in the last plot a simulation done by substituting the switch with an ideal resistor of 20 Ω is reported for comparison. The estimated second harmonic distortion is reported in the captions of the plots. These simulations indicates that a converter with a resolution of 10 bits and a sampling frequency of 5 Ms/s should be feasible in this technology and that it can operate from a 2.5 V supply with a reasonable safety margin. Higher performances can be obtained with switches of bigger size; in principle, bigger switches add bigger parasitic capacitance to the bottom plates of the DAC, but the impact of these parasitics in negligible, since the bottom plates are always connected to low impedance sources [23].

5.2 Design of a capacitive-only 10 bits DAC

5.2.1 DAC architecture

The first step in designing a charge redistribution DAC is the choice of the capacitors. The process used in the implementation of this ADC features very linear metal to metal capacitors with a satisfactory density. Unfortunately, at the time of the project, no data about the matching of these structures were available, so we had to assume “a priori” that the matching was good enough for the design of a 10 bits ADC.

The elementary cell in the DAC is a capacitor with a nominal value of 75 fF and occupies an area of 18 x18 μm^2, including inter-cell spacing and routing. The bigger capacitors are obtained by connecting in parallel a suitable number of these units.

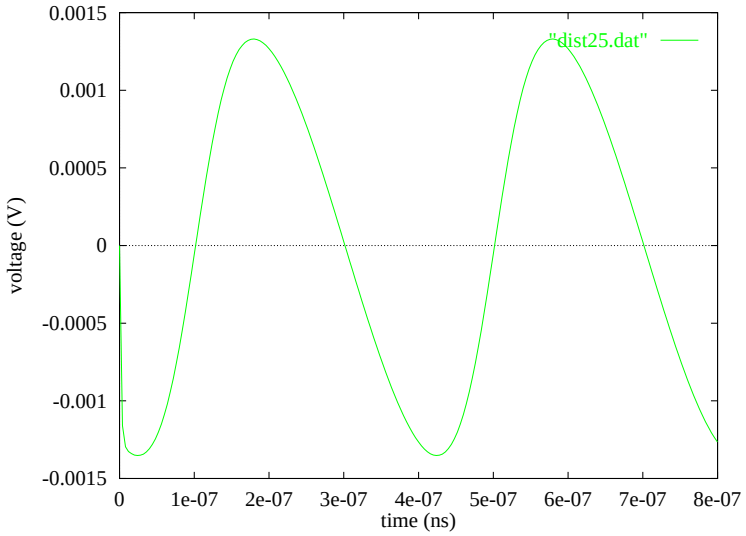

Figure 5.3: Difference between the input and the output signal for the circuit of fig. 2. Switch size: 200/0.36; Vdd=2.5 V. Second harmonic: -67 dB

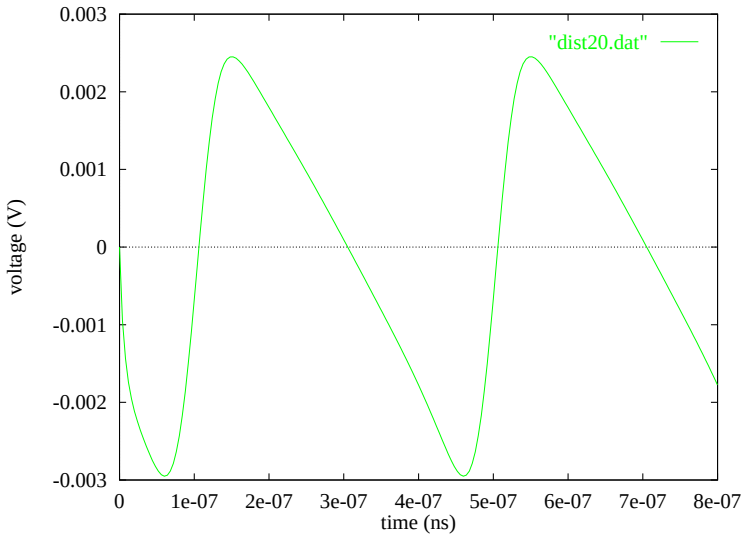

Figure 5.4: Difference between the input and the output signal for the circuit of fig. 2. Switch size: 200/0.36; Vdd=2 V. Second harmonic: -54 dB

Table 1 shows the total area and the total capacitance needed in this technology to implement a binary weighted DAC in function of the number of bits

Table 5.1: Area occupation and DAC capacitance in function of the number of bits

Number of bits	Area	Total capacitance
5	100 x 100 μm^2	2.4 pF
8	288 x 288 μm^2	19.2 pF
9	407 x 407 μm^2	38.4 pF
10	576 x 576 μm^2	76.8 pF

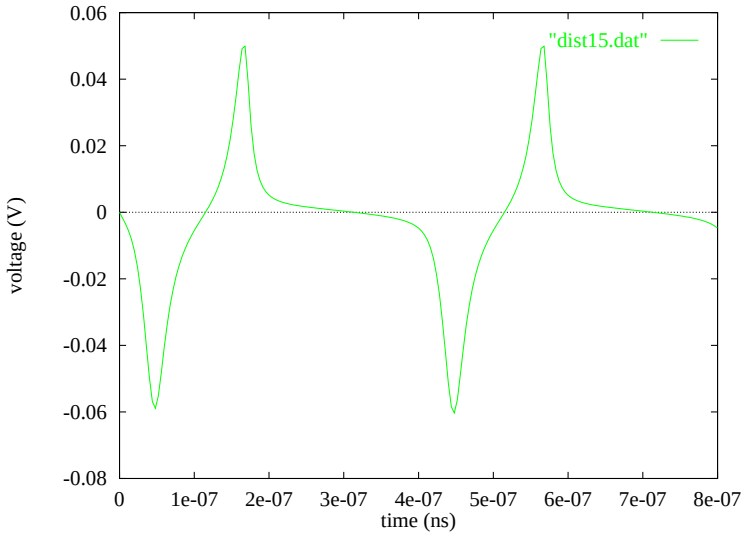

Figure 5.5: Difference between the input and the output signal for the circuit of fig. 2. Switch size: 200/0.36; Vdd=1.5 V; second harmonic: -28 dB

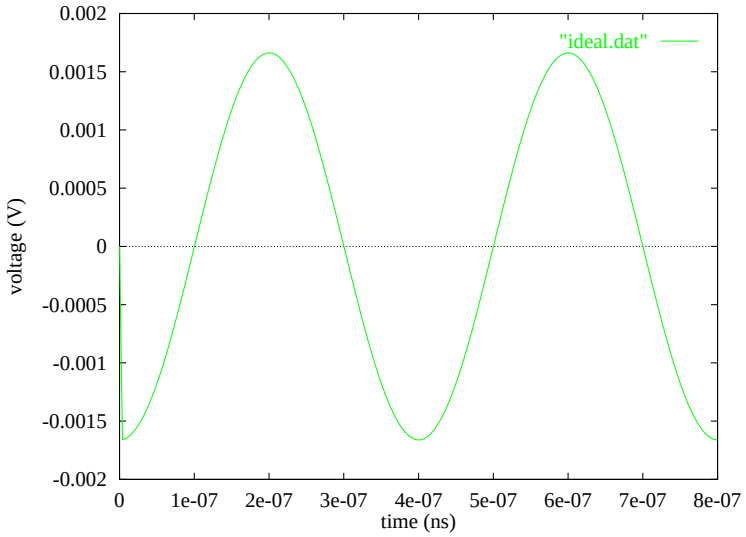

Figure 5.6: Difference between the input and the output signal for the circuit of fig. 2 when the switch is replaced with an ideal resistor 20 Ω

It is apparent from this table that a direct implementation of a 10 bits DAC is not very practical both for the area (especially if more converters have to be integrated on the same chip) and for the total capacitance, which would represent a heavy load for the circuit driving the ADC. This is, in fact, one of the more serious drawback of charge redistribution converters; the solution to overcome this problem is to scale down the voltage on the bottom plates of the capacitors, providing in this way further division of the reference voltage. Fig. 7 depicts a possible implementation of these principle [35] and is used to illustrate the basic idea. In this case the first 8 bits are decided with a conventional binary weighted DAC; the smallest fraction of the reference voltage that can be generated by an 8 bit DAC is

$$\frac{C}{256C}V_{ref} = \frac{V_{ref}}{256} \tag{5.3}$$

After the first 8 bits have been encoded, the termination capacitor C is connected to a second

DAC, in this particular case implemented with a resistor string. When the bit N+1 is tested, the output of the auxiliary DAC is $V_{ref}/2$ and, by applying eq. 1, we see that a step voltage equal to $V_{ref}/512$ is generated at the output of the main DAC. The use of a resistive sub DAC is very common in the literature, but we have not considered it because the resistive string dissipates static power.

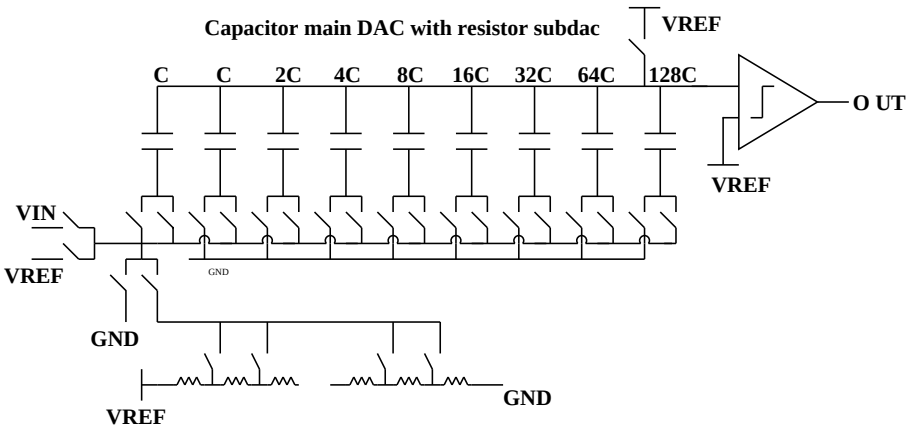

Figure 5.7: Charge redistribution ADC using a binary weighted main DAC with a resistive subDAC.

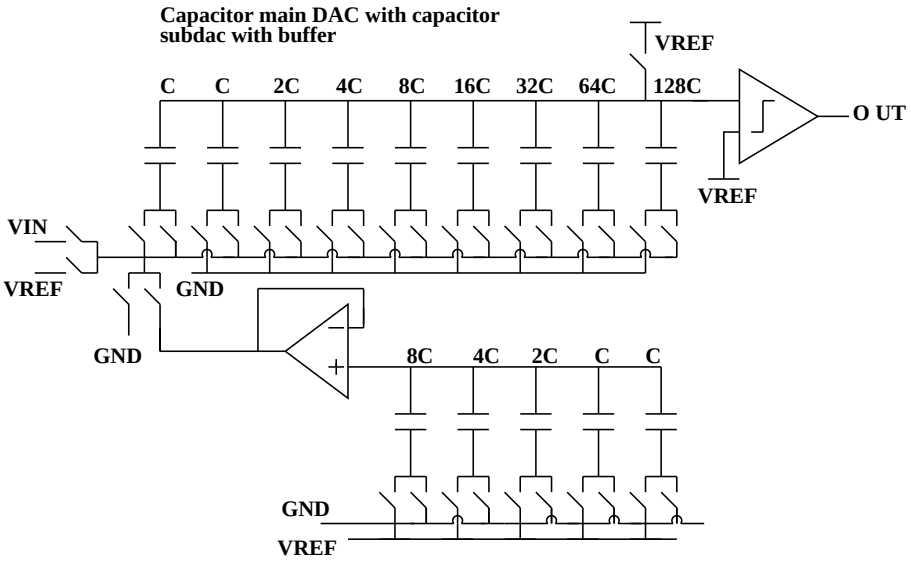

Figure 5.8: Charge redistribution ADC using two binary weighted DACs.

A possible alternative is to use a second binary weighted DAC, as shown in fig. 8. This scheme has been often adopted in the implementation of a high resolution converters embedded in circuits for telecommunications [36]. The first N bits are determined by the main DAC in the classical way and the during this phase the termination capacitor is grounded. In order to determines the M LSBs, the termination capacitor is connected to the output of the voltage buffer while the second DAC provides the supplementary voltage partitioning. The aim of the driver is to guarantee a decoupling between the two DAC, thereby avoiding any loading effect.

The presence of this stage has the drawback that it dissipates static power and may require some time to settle, thus limiting the overall speed. Moreover, if we want to operate the converter in the configuration of fig. 8, i.e. using only a single power supply and a unipolar reference, the design of the buffer can require some care, because at least from time to time it has to handle signals very close to one of the supply rails (GND in this case). Actually, in the original architecture described in [36] the ADC was operating using a bipolar reference. Though it is evident that in a configuration like the one depicted in fig. 7 a decoupling between the two DACs is mandatory for very high resolutions, we calculate now if a direct coupling allows accuracies at the level of 10 bits.

For the time being we suppose that a 8 bit DAC is used as the main stage, as illustrated in fig. 9 and a direct coupling between the two DACs occurs via a unit capacitor C.

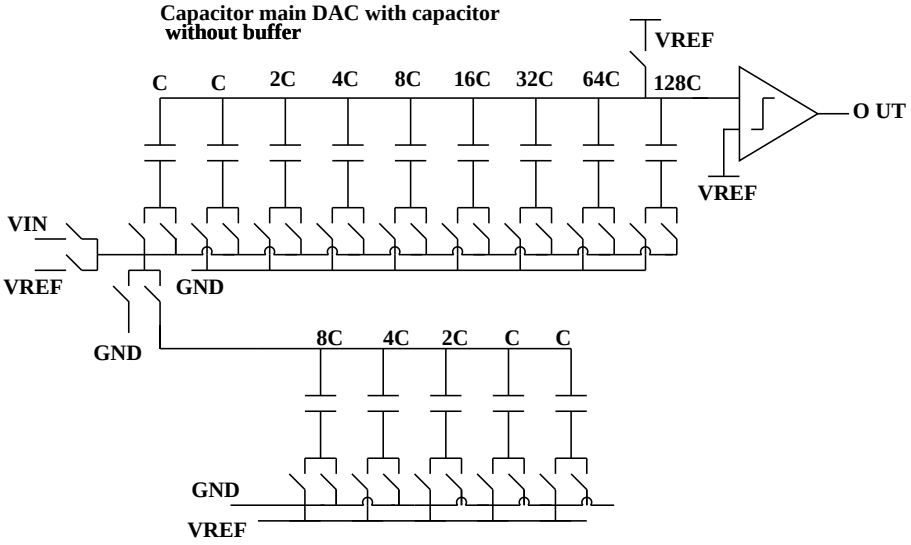

Figure 5.9: Charge redistribution ADC using two binary weighted DACs with direct coupling

The presence of the second DAC loads the main stage, so that the termination capacitance, instead of C is the series combination of C and the total capacitance of the second DAC, $CDAC_2$. Table 2 gives the fractional errors induced by the direct coupling on the main DAC, in function of the number of bits of the second DAC, under the hypothesis that the main DAC has 8 bits resolution.

Table 5.2: Errors induced by the direct coupling in function of the size of the second DAC

Number of bits in the subDAC	$CDAC_1/CDAC_{1TH}$
2	0.999
3	0.9996
4	0.9998
5	0.9999

It is obvious that the impact of the direct coupling is minimised by increasing the size of the sub DAC and a subDAC with 5 bits capability will provide an error at the level of 13 bit, which is adequate for our application.

In our circuit, the signal is sampled only by the first DAC, that operates in stand-alone mode for the decision of the first N bits; the second DAC is connected only for the last M-N bits via the termination capacitor and coupling errors are introduced only at this level. As it can be seen from fig. 9, we made a very conservative choice, determining eight bits in the main DAC; the second DAC has actually 5 bits, among whom the two MSBs are use to determine the last two bits of the conversion. Still, this solution is more efficient in term of area and power than the solution using the voltage buffer. The timing of the circuit is the following:

- During the *sampling* phase, the signal is sampled on the main DAC, which works in stand-alone mode and the sub DAC is grounded
- In the first *redistribution* cycle, the first eight bits are determined by the main DAC
- In the second *redistribution* cycle, the termination capacitor of the first DAC is connected to the output of the sub DAC and the two last bits are found

Due to direct coupling and to the asymmetry in the size of the two DACs, the voltage steps generated by the second DAC will be slightly different, thereby giving a small contribution to the overall integral nonlinearity. However, it can be easily verified that this contribution is irrelevant for a 10 bit resolution and the main limitations to the circuits will come from capacitors mismatch. The capacitive loading on the circuitry driving the ADC is limited to the main DAC and is equivalent to that of an eight bits ADC. The area is increased only by 15% with respect to an 8 bit solution, which represents a great saving compared to a straightforward 10 bits implementation. Since no buffer is used, the converter can easier work from a single power supply and a unipolar reference.

5.2.2 Speed optimisation

The speed calculations have been done for the main DAC and then scaled to adapt to the sub DAC. The speed of a binary weighted DAC is limited by two different time constants

$$\tau_s = (R_{SWref} + R_{SWin} + \frac{R_{ON}}{2})C_{DAC} \tag{5.4}$$

and

$$\tau_r = \frac{R_{on}}{2}C_{DAC} \tag{5.5}$$

for the sampling and redistribution mode. In these equations, C_{DAC} is the total capacitance of the array an R_{ON} is the equivalent resistance of the switch driving the MSB capacitor, R_{SWref} and R_{SWin} are the equivalent resistance of the switches linking the array to the input and the reference voltage, respectively. If we want to operate our converter with a 40 MHz clock, we have to assure that the redistribution time constant is small enough to let the converter settle at each step within the desired accuracy. Actually, only half clock cycle will be available for settling, since the other half is used for comparison. The requirement of settling to 10 bits resolution in 10 ns implies an equivalent resistance for the switch driving the MSB capacitor of 115 Ohm [3]. In practice, due to the timing in the comparator, the time allocated for settling must

[3] In this calculation, we have taken into account the fact that the absolute value of the capacitor can change of a 15% due to process variations and we have assumed the highest value, that would lead to a total capacitance of 22.1 pF for the 8 bits DAC

be smaller. We have chosen a nominal value of 20 Ohm, in order to accommodate also process variations which might degrade the quality of the transistors. Once the size of the MSB switch is found, the sizes of the other switches are determined by scaling them according to the values of the capacitors they control.

The optimisation of the redistribution time constants does not entail particular issues, because the bottom plates of the capacitors are always connected to low impedance nodes and the sizes of the switches (and hence their parasitic capacitors) are not of particular concern. The same statement holds also for the switch used to connect the bottom plate of the array to V_{in}.

For the switch that links the top plate to V_{ref} the situation is slightly different, because, its parasitic junction capacitance can be of the same order of the LSB capacitance. Since at the end of the conversion the voltage on the top plate has converged back to its initial value, the non linearity of this capacitance has little influence on the accuracy of the conversion. However, it does provide an attenuation at the output of the DAC which reduces the actual signal available for the comparator; therefore this switch should not be oversized. The fine optimisation, done with the help of computer simulations, led to a choice of a W/L ratio of 26/0.36; the resulting resistance should be small enough to allow accurate sampling within one clock cycle.

5.2.3 DAC layout

The DAC has been laid-out in the conventional way: all the capacitors are obtained by replicating the same fundamental cell and using a common centroid geometry to attenuate the impact of systematic gradients. Since in this technology the capacitor are built by sandwiching two higher level of metals, the first level of metal has been used to implement a ground shield underneath the arrays, in order to reach a better insulation from the substrate. The layout of the full DAC (including the main DAC and the sub DAC) is shown in fig. 10.

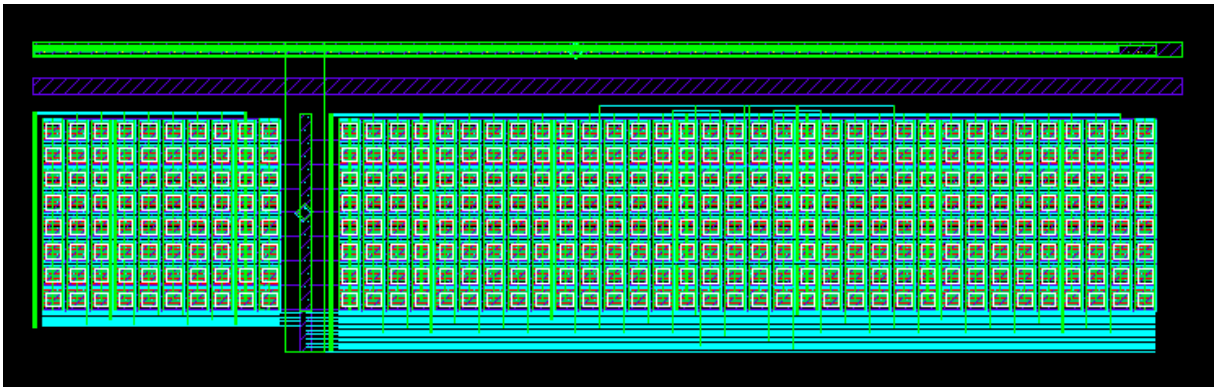

Figure 5.10: Layout of the 10 bit DAC

5.3 Comparator design

5.3.1 Comparator architecture

The second basic element in the ADC is the voltage comparator, whose scheme is shown in fig. 11 . It consists of two ac-coupled differential stages, a positive feedback latch and some

digital logic (not shown in the figure) to fully regenerate the outputs to CMOS levels and drive the shift register. A minimum length of 0,5 μm has been chosen for all the transistors, except the current mirrors and the switches. The choice has been done in order to assure that all the critical transistors work outside the velocity saturation region.

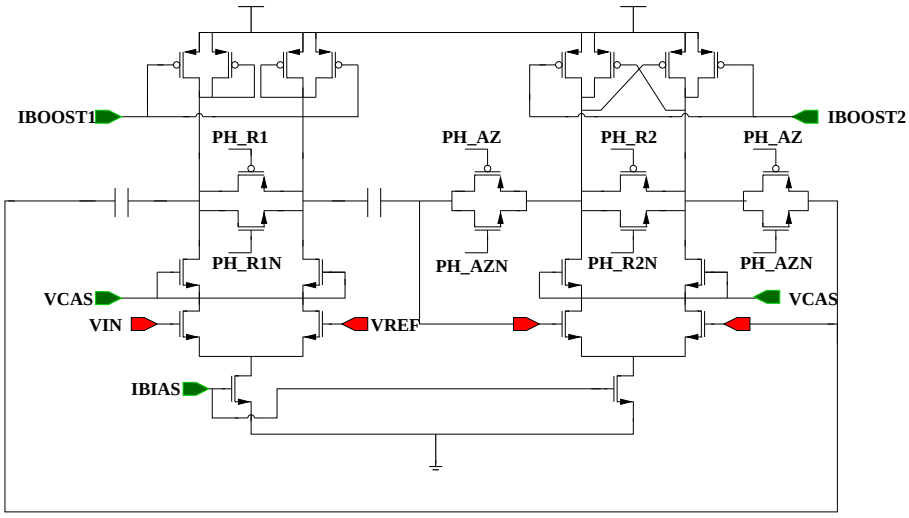

Figure 5.11: Schematic of the voltage comparator

Despite the 2.5 V power supply, a cascode configuration has been used in the differential pairs. The cascode transistors have been introduced to improve the speed performances and to reduce the drain-source voltage on the input transistors below the velocity saturation voltage. The loads of the first stage are implemented with diode-connected transistors coupled with current mirrors; in this way the current flowing in the loads can be adjusted with an external control and the small signal gain can be changed. The second stage is loaded with a cross-connected PMOS pair and according to what we have seen in chapter 3, the width of these transistors has been chosen to 5 μm. Also the loads in the second stage are coupled to current mirrors and can be controlled by an external bias.

5.3.2 Offset compensation

The offset compensation procedure, necessary to maximise the accuracy, is carried-out on both stages of the comparator before each conversion. It is performed while the ADC is in the sampling mode in order not to add to much time over-head to the conversion. During the sampling mode, in fact, both inputs are connected to the reference voltage and the offset of the first stage is stored on the coupling capacitor and hence its effect is completely eliminated. The offset on the second stage is compensate by using the *reset mode gain*, as proposed in [37]

As shown in chapter 3, the behaviour of the cross-coupled loads can be regenerative or not depending on the sign of the time constant, which is defined by

$$\tau = \frac{C_L}{g_{mL} - 2g_{dsR}} \tag{5.6}$$

where g_{mL} is the transconductance of the load device M_{L2A} or M_{L2B} and g_{dsR} is the conductance of the switch. If $2g_{dsR} > g_{mL}$, τ is negative and the second stage behaves like an amplifier. After

the transients have died-out, the gain of the second stage in the amplifying mode is

$$A_{v2} = \frac{g_{m1}}{2g_{dsR} - g_{mL}} \tag{5.7}$$

Therefore, closing a feedback loop around this stage with the autozero switches controlled by the signals PH_{AZ}, PH_{AZN} , its offset is stored on the coupling capacitors and reduced by a factor $1 + A_{v2}$. The residual offset referred to the input of the chain is:

$$V_{OSR} = \frac{V_{OSR2} + V_{inj}}{A_{v1}} \tag{5.8}$$

where the term V_{inj} has been introduced to take in account the effect of the mismatch in charge injection from the autozero switches of the second stage into the coupling capacitors. The term V_{OSR2} is the residual offset of the second stage after the compensation and is defined by

$$V_{OSR2} = \frac{V_{OSN2} + V_{OSL}\frac{g_{mL}}{g_{m1}}}{1 + A_{v2}} \tag{5.9}$$

In determining the gain of the two stages of the comparator, we have calculated the offset of the structure from the parameter of the technology, concluding that a gain of 10 both in the first and the second stage should be sufficient to reduce the input referred offset below 1 mV[4]. However, thanks to the current mirrors, the gains of both stages can be independently changed during the test and their effect on the overall system performances can be investigated.

5.3.3 Speed optimisation

The time allowed for comparison is in principle half a clock cycle. However, between two consecutive conversions the comparators should be reset, in order to prevent hysteresis. The mechanism of the reset can be better understood referring to fig. 12, in which the two main control signals are shown.

For clarity, we display here only the signal driving the NMOS transistors in the complementary switches. When the signals are high, the switches are closed; the first stage is fully reset, whereas the second stage has a residual gain given by eq. 7 When P_{HR1} goes low, the input stage is enabled and amplifies the voltage difference between its inputs by a factor $A_{v1} = \frac{gm_1}{gm_{L1}}$; this voltage is further amplified by the second stage via the gain A_{v2}. When also P_{HR2} goes low, its load becomes regenerative and the positive feedback strongly amplifies the voltage between the cross-connected terminals; an output latch (not shown in the figure) is strobed at the end of the clock phase to store the decision of the comparator. We must observe that the time allocated to the reset phase is not half a clock cycle, but only the time during which P_{HR1} is high. In our implementation, P_{HR1} has the same period of P_{HR2} and a duty cycle of the 25%. Therefore, if a clock of 40 MHz is assumed, the comparator has about 6 ns for the reset phase and 6 ns for the preamplification phase. If we require a 1% settling during this time, the value of the reset time constant if 1 ns. The value of the parasitic capacitance at the output node is primary determined by the parasitic capacitances introduced by the current mirrors and the cascode transistors and

[4]Since we are designing with a *fixed* power budget, to use unnecessary gain in the amplifiers worsen the speed of the comparator

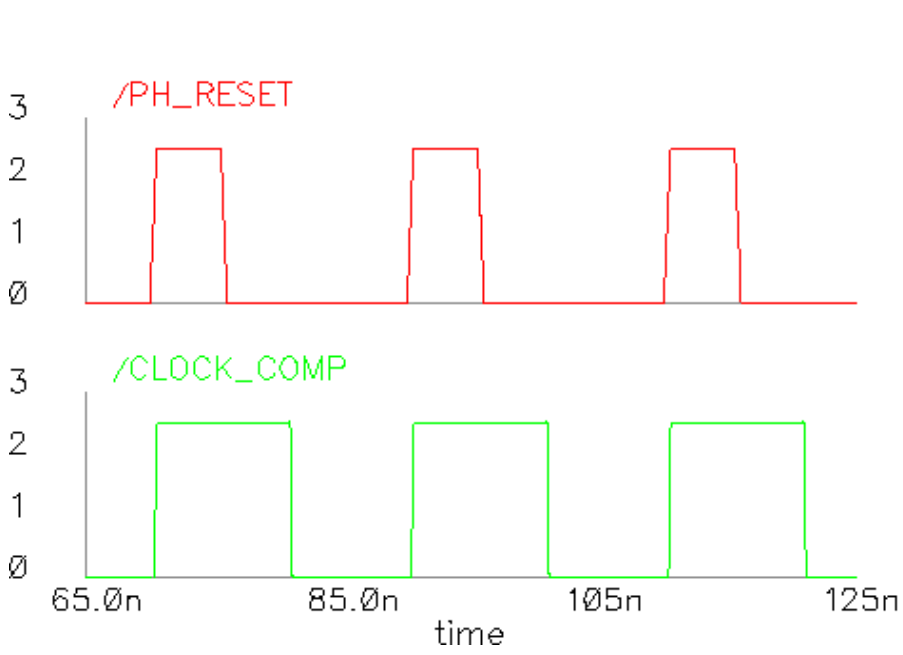

Figure 5.12: Schematic of the voltage comparator

has been estimated to 160 fF. Therefore, the term $g_{mL} - 2g_{dsR}$ should not be smaller than 160 μS. Since this value determines also the reset mode gain, a transconductance of at least 1.6 mS is required in the input transistors of the second stage in order to achieve a reset mode gain of 10. The value of g_{dsR} and of g_{mL} have been chosen equal, in order to guarantee with a good safety margin that during the reset phase, the positive feedback is switched off. Since the circuit is powered from a 2.5 V power supply, it is mandatory to use complementary reset switches to assure an adequate robustness against power supply variations.

5.3.4 Comparator layout

Fig. 13 shows the layout of the voltage comparator. The structure has to be very symmetric, since any asymmetry can transform a common mode signal into an unwanted differential signal, thus degrading the accuracy of the circuit. As it can be seen from the figure, a big fraction of the area is occupied by the coupling capacitors; this is mainly due to the fact that to optimize matching, a ring of dummy capacitors has been laid-out all around the signal capacitors. The same technique has been applied to all the other blocks in the circuit. To reduce as much as possible the coupling between digital and analog parts the control signals are routed at the periphery of the comparator and therefore all the reset switches are placed near the edges. For the same reason, the small logic block which generates the CMOS signals for the successive approximation register is placed far from the analog parts. In order to enhance the radiation resistance, enclosed layout transistors have been used everywhere. The total area occupied by the circuit is 390 x 150 μm^2.

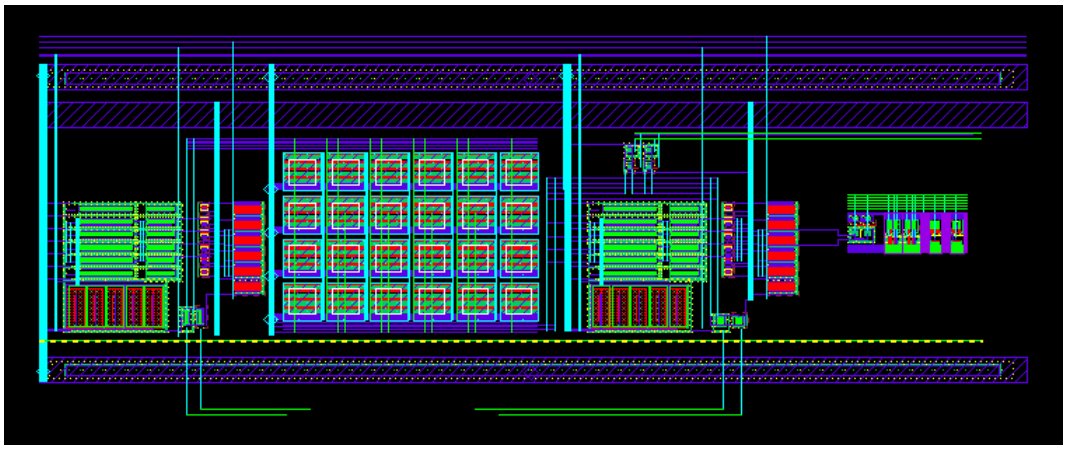

Figure 5.13: Layout of the voltage comparator

5.4 Global converter architecture

Two complete analog to digital converters and a spare comparator have been integrated on a die of 2 x 2 mm^2 Due to space constraints, the outputs of the two ADCs have been multiplexed. The size of a single converter is 0.3 x 1 mm^2 Since the circuit has been implemented in a epitaxial technology with an heavy doped substrate, all the digital and the power supply lines have been carefully splitted. A dedicated line has been used to bias the substrate of the digital standard cells. A double pad has been provided for the bonding of reference voltage, in order to minimize parasitic effects due to the inductances of the bonding wires.

5.5 Test results

The part has been delivered from the foundry only at the end of October 1999, so only preliminary tests have been accomplished up to now. We have performed first functional measurements on few samples before and after irradiation; some representative measurements are discussed hereafter.

The ADCs have been tested with the same procedures described in the previous chapter for the converter in the 0.7 μm technology, i.e. applying a full scale sinusoid and calculating the FFT, the INL and the DNL.

The tests with the nominal clock frequency detects some missing codes near the MSB transition, which are clearly visible in the DNL profile and in the transfer characteristic. This problem is common for all the converters measured and disappears if the clock frequency is scaled to 20 MHz. One converter exhibits a similar problem also at 20 MHz clock, but on a different code, corresponding to the second most significant bit. All the others codes are correctly detected and the sizes of the steps are quite uniform; this explain the very low level of distortion, which is always at the edge of a 10 bits converter. In fact the second harmonic ranges from -62 dB to -58 dB below the fundamental.

The problem with the missing code, depending on the clock frequency is clearly a dynamic problem that does not depend on the matching of the capacitors chosen to implement the DAC. We have observed that the number of missing codes changes by modifying the bias of the loads of the second stage of the comparator. A preliminary hypothesis is that the value of the time

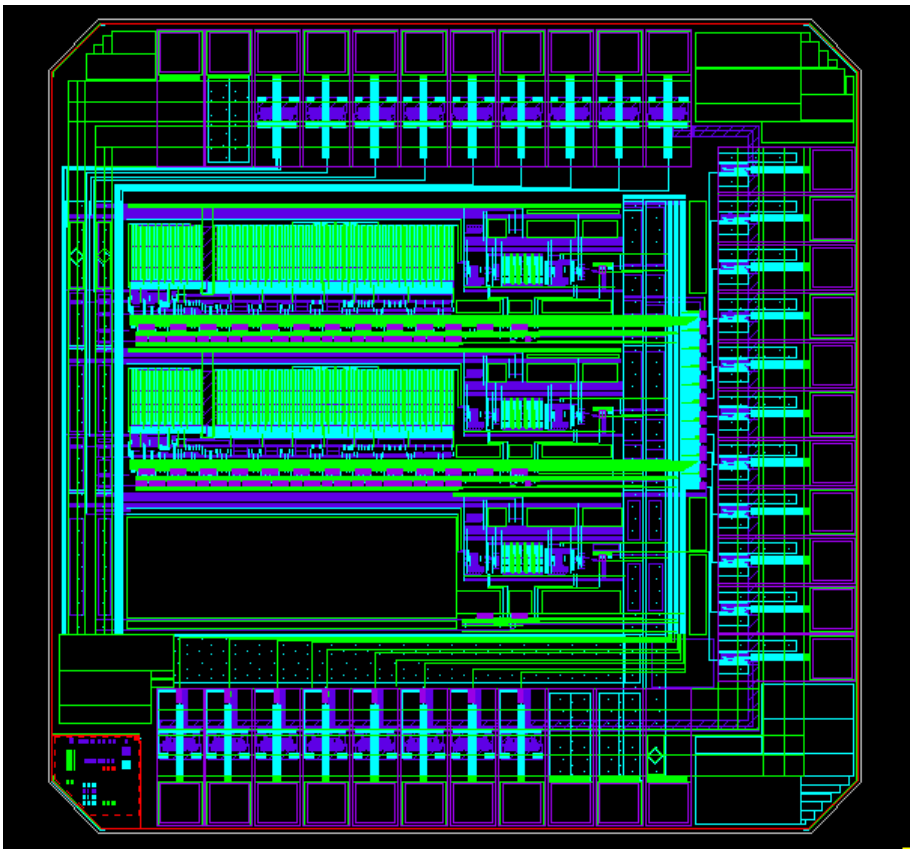

Figure 5.14: Layout of the full ADC chip

constant of the second stage was underestimated during the design phase, thereby slowing down the comparator.

One of the major goals of these design was to demonstrate the resistance to total radiation dose of a switched capacitor circuit laid-out with the enclosed-layout techniques. Therefore, as a first step. three ADCs have been tested to a total dose of 200 krad, 1 Mrad and 10 Mrad. No degradation in circuit performances was seen, thus demonstrating the effectiveness of the enclosed structures in hardening a conventional technology.

Figure 5.15: Response of the converter: measurement before irradiation with a clock frequency of 20 MHz

Figure 5.16: Response of the converter: measurement before irradiation with a clock frequency of 40 MHz

Figure 5.17: Response of the converter: measurement after 100 krad

Figure 5.18: Response of the converter: measurement after 1 Mrad

Figure 5.19: Response of the converter: measurement after 10 Mrad

5.6 Summary

A 10 bits charge redistribution converter has been designed in a 0.25µm CMOS technology. The converter is mainly intended for parallel data acquisition systems for high energy physics experiments. Therefore the design has been optimized for very low power consumption and the chip has been laid-out using systematically enclosed layout transistors in order to enhance the resistance to total dose radiation. The chip operates from a 2.5 V power supply and a reference voltage of 1.5 V. The first tests show a ten bits capability with a 20 MHz clock, while a 40 MHz a problem has been detected with the codes adjacent to the MSB. In this early stage of the tests, the problem has not been jet clearly identified, but being a dynamic failure, it could be explained by speed limitations in the comparator. Three chips have been irradiated to three different total dose (200 krad, 1 Mrad and 10 Mrad) without measuring any degradation.

6 Design of integrated high gain transimpedance amplifiers in CMOS technologies

The preliminary research presented in this chapter deals with the design of integrated circuits for the measurements of very low photocurrents. The work has been carried out in the VLSI Laboratory of the Politecnico di Torino in the framework of a R & D effort aiming at the development of an integrated circuit for the read-out of electrochemiluminescence sensors.

In the past few years electrochemiluminescence (ECL in the following) has gained popularity as an alternative detection technology for applications both in diagnostics and in fundamental biological research.

ECL consists in light emission due to an oxidation reduction reaction of a ruthenium ion triggered by an applied voltage. The reaction is based on the use of two components, a ruthenium chelate, which is recycled and tripropylamine (TPA) which is consumed. The ruthenium chelate serves as a marker and is bound with a chemical procedure to the substance to be detected. The labelled component is captured on the surface of paramagnetic beads, that are pulled by a magnetic field to the surface of an electrode. The TPA is introduced in excess into the flow cell and a low voltage ($\simeq$ 2 V) is applied to the electrode. The low voltage determines an oxidation reduction reaction, in which the TPA loses a proton and becomes a reducing reagent, transferring one electron to the ruthenium. The electron is captured in an excited state and then decays, emitting a photon.

While the TPA is consumed by the process, the ruthenium is recycled; therefore the *same* label on the *same* molecule can be used for subsequent reactions, thus enhancing the sensitivity of the analysis. With this technique, quantities down to the picogram level can be measured. Of course, when a very low amount of substance has to be detected, few markers are used and the light emission is weak. Hence, very sensitive photodetection systems are needed. Commercial systems employ in fact photomultipliers tube, with the drawback that the overall apparatus is big and expensive.

Since the ruthenium emits in the visible band ($\lambda \simeq 600$ nm) a silicon photodiode could be used as a detector. A read-out chain composed by a silicon photodiode and an integrated front-end will lower dramatically the costs making the technique available also to small size labs, which can not afford the price of a conventional instrument. At the end, also the photodiode could be integrated on the same silicon substrate of the front-end electronics; in fact a photodiode custom-designed in a standard process will have worse performances than a commercial and specialised product, but this can be compensated by the attenuation of the parasitics, which are much larger in hybrid systems.

Basically, two kinds of front-end are suitable for a photosensor. One possibility, illustrated in fig. 1a, is to use a transimpedance amplifier. In this configuration a resistor is used as the feedback element of an operational amplifier. The amplifier must have a low input bias current, therefore amplifiers with JFET or CMOS input transistors are preferable. The transimpedance configuration has the advantage of providing a real-time image of the sensor signal; the drawback is that for very high sensitivity the feedback resistor must have a value of 100 MΩ or more, which makes impractical the implementation of the whole circuit in a monolithic form.

Therefore, for a fully integrated solution, the configuration of fig. 1b, using a capacitor in the feedback loop, is preferred. The input current is integrated onto the capacitor C_f for a given time; afterwards, the output of the amplifier is sampled and digitized an the integrator is reset by closing the switch S_f.

The advantage of this architecture with respect to the previous one is clear, since capacitors are more easily implemented in a monolithic form than resistors. The output voltage is defined by $V_{OUT} = \frac{It}{C_f}$ where I is the input current, C_f is the feedback capacitor and t is the integration time. Hence, the sensitivity is improved either by increasing the integration time or by *decreasing* the value of the feedback capacitor. This is in principle very interesting, because a more sensitive circuit requires less area, which is the opposite of what happens with the transimpedance configuration. However, to assure a proper collection of the charge, the condition $A_v C_f \gg C_d$ must hold, where A_d is the open-loop gain of the op-amp and C_d is the parasitic capacitance associated with the photodiode. As a consequence, C_f can not be decreased below a certain limit, which depends on the specific application. Still, the sensitivity can be increased using longer integration time, at the expense of speed.

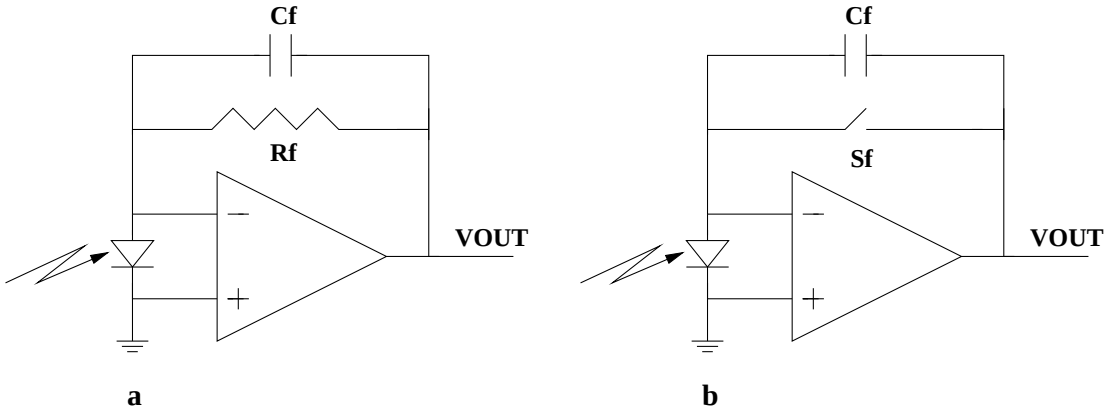

Figure 6.1: Front-ends for photodetectors: **a** Transimpedance amplifier; **b** Integrator

The analog to digital converter can be implemented on the same die of the integrator; actually two integrators can be used in a time interleaved configuration, so that while the output of one integrator is processed by the ADC, the input current is steered to the other, thereby avoiding any dead time. Systems with a resolution up to 20 bits based on this approach are found on the market; such an high resolution is usually obtained by using sigma-delta converters and precision analog techniques, like correlated double sampling and chopper stabilisation. However, integrated circuits achieving a very good resolution with simpler and lower cost architectures have also been reported [38].

Thought a current integrating system has many advantages, it must be observed that the transimpedance configuration is nevertheless attractive, since, providing a real time representation of the light source, it is best suited for studying the intrinsic characteristic of the detector.

Additionally, in ECL sensors the light emission can be triggered applying a modulating potential to the electrode and very high sensitivity could be achieved by detecting the output signal with a lock-in technique. In this case front-end reproducing the signal with good fidelity is needed as an interface between the photodiode and the lock-in apparatus.

In the early stage of this investigation, a discrete transimpedance amplifier built with a low-noise BiFET operational amplifier and a feedback resistance of 1 GΩ was used as a the front-end of the photosensor, which was a commercial photodiode. With this simple system a sensitivity of 100 pA was possible. The integration of the feedback resistor on the same die of the amplifier, would reduce the contribution of the parasitic capacitance and of the interference noise, hence increasing the resolution.

The aim of our R & D is to develop low cost solutions with the ultimate goal of integrating the full chain, from the photodiode to the digitizing element, on the same silicon substrate. We think that a current integrating system would be probably preferable for the final application; however, a fully integrated high-gain transimpedance amplifier is also desirable for the intermediate characterisation of the custom-developed photosensor and to exploit the lock-in technique to reach very high sensitivity in some specific measurements.

The rest of the chapter discusses the study on possible implementations of monolithic high-gain transimpedance amplifiers. In the last section. a first circuit implementing on the same chip the photodiode and the processing electronics recently sent to the foundry is presented.

6.1 Design of monolithic high-gain transimpedance amplifiers

The design of a very high gain monolithic transimpedance amplifier demands an efficient implementation of the feedback resistor. In fact, integrated resistors beyond few hundreds kΩ are usually not practical because they require large area and add considerable parasitic capacitance. Moreover, high value resistors are not available in all processes. Therefore, the alternative of emulating the resistive function with active elements has to be considered. This solution, in order to be competitive with its discrete counterpart, should allow to reach comparable values of equivalent resistance (i.e. in the range of 1 GΩ) occupy an acceptable die area and not worsen the noise of the system. This last point is particularly important in our case, since we are aiming at the detection of very small signals.

The current spectral noise density of a transimpedance amplifier can be calculated with the following equation [6]

$$i_n = 4kT\frac{\Delta f}{R_F} + s_n\frac{1+\omega^2[(C_d+C_{in})R_F]^2\Delta f}{{R_F}^2} \tag{6.1}$$

where R_F is the value of the feedback resistance, s_n is the voltage noise spectral density of the core amplifier, C_d is the parasitic capacitance of the photosensor and C_{in} is the input capacitance of the amplifier. Furthermore, at low frequency, the contribution of the core amplifier can be made easily negligible compared to the one of the feedback resistor. From this equation we see that the noise is reduced by increasing the feedback resistor of the amplifier and by reducing the input parasitic capacitance; therefore a fully integrated solution may represent a significant advantage.

Usually, the feedback resistor is implemented with a passive component and the core amplifier is a simple common source or cascode stage [39]. Actually, in the front-end of detectors single ended input stages are often preferred to differential ones, because the theoretical superiority of a differential configuration is easily wasted by the mismatches between the inputs introduced by the parasitic capacitances. Moreover, for the same power budget, a differential input stage has an intrinsic noise 3 dB higher than a single ended stage.

In our application the use of an input stage with very low input bias current is mandatory and hence a CMOS technology has been chosen[1]. For an optimal front-end design, the noise contribution of the amplifier is due mainly to the input transistor and for a MOS device the quantity s_n is defined as[2]

$$s_n = \frac{8kT}{3g_m} + \frac{K_f}{C_{ox}{}^2 WL} \tag{6.2}$$

where

- k is the Boltzmann constant
- T is the absolute temperature
- g_m is the transconductance of the device
- K_f is the flicker noise coefficient
- C_{ox} is the gate oxide capacitance per unit area
- W and L are the width and the length of the device

In the standard design flow, the noise optimisation starts with simplified calculations considering the contribution of only few critical devices and is then refined with more comprehensive computer simulations. However, the noise models themselves are often defective[3] and the final performances can only be assessed by laboratory measurements. This general statement is particularly true when the design specifications are as tight as in our case. We have also considered important to compare the noise of the considered circuits with the one of a benchmark circuit, in which the feedback is obtained with an ideal but noisy resistor. In fact when the feedback network is implemented with more devices, some more noise has to be expected.

We have studied three circuits, using always the same core amplifier (a direct cascode stage) and different feedback topologies. The three architectures are described in the following pages. For each circuit a simplified schematic is presented and the main design issues are discussed, together with the results of computer simulations. The plots show the magnitude of the transfer function and the transient response of each circuit. The transient response is reported in the range 20 pA - 100 pA, with steps of 20 pA. In order to compare the different techniques, in all circuits the transimpedance has been adjusted to about 1 GΩ and the bandwidth to 1.5 kHz. All the simulated systems had first-order low-pass transfer functions. An additional capacitor of 1 pF has been added in parallel to the input current source to emulate the parasitic capacitance of the sensor.

[1] A JFET solution has not been considered because it would increase the cost significantly.

[2] For the sake of simplicity, we neglect here the excess noise factor. However more accurate equations for the noise of the MOS transistor are reported in chapter 1.

[3] For instance, in some models, the thermal noise of a transistor working in the triode region is taken to be zero, which of course is not true. Another example is the excess noise factor we have discuss is chapter 1, which is almost never considered.

6.1.1 Single MOS transistor feedback

In MOS technology, resistor of very high value can be implemented in a small area using a MOS transistor biased in the weak inversion region. In the weak inversion region the drain-source current of a MOS device can be expressed as

$$I_{DS} = I_{D0}\frac{W}{L}e^{\frac{V_{GS}-V_M}{nV_t}}(1-e^{-\frac{V_{DS}}{V_t}}) \tag{6.3}$$

where

- I_{DS} is the current flowing in the device
- I_{D0} is a constant depending on the particular process used to fabricate the device
- V_t is the thermal voltage $\frac{kT}{q}$
- n (see chapter 1) is a parameter that can be assumed to be 1.5
- V_M is the upper limit of the weak inversion region.

Differentiating eq. 3 with respect to V_{DS}, and evaluating the result for $V_{DS} = 0$ we get

$$R_{DS} = \frac{\partial I_{DS}}{\partial V_{DS}} = I_{D0}\frac{W}{LV_t}e^{\frac{V_{GS}}{nV_t}} \tag{6.4}$$

This relation shows that the equivalent resistance of the device can be made small by decreasing the $\frac{W}{L}$ ratio. Fig.2 show the implementation of this topology; with this circuit an equivalent resistance of 1 GΩ could be obtained with W/L=2/70 and $V_{GS} = V_{GD} = -300\,mV$. The results of the simulation are shown in fig. 4 and 5 for the transient and the ac response, respectively.

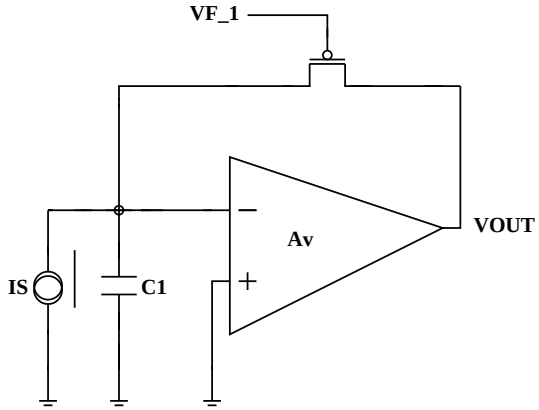

Figure 6.2: Transimpedance amplifier with single MOS transistor feedback

The circuit attains a gain of 1.05 GΩ and the simulated input referred noise is 0.276 pA r.m.s. over a bandwidth of 1.5 kHz. This is very close to the noise of the amplifier with the passive feedback resistor, which is 0.260 pA. This last number is interesting, because allows us to calculate the noise contribution of the core amplifier itself. In fact, in case the parallel noise contribution of a passive resistor is $\frac{4kT}{R_F}$ and a resistor of 1 $G\Omega$ will give in our condition (i.e. first order system with a cut-off frequency of 1.5 kHz) a contribution of 0.2 pA r.m.s. Therefore, the noise introduced by the main amplifier and the input parasitic capacitance is 0.18 pA r.m.s.

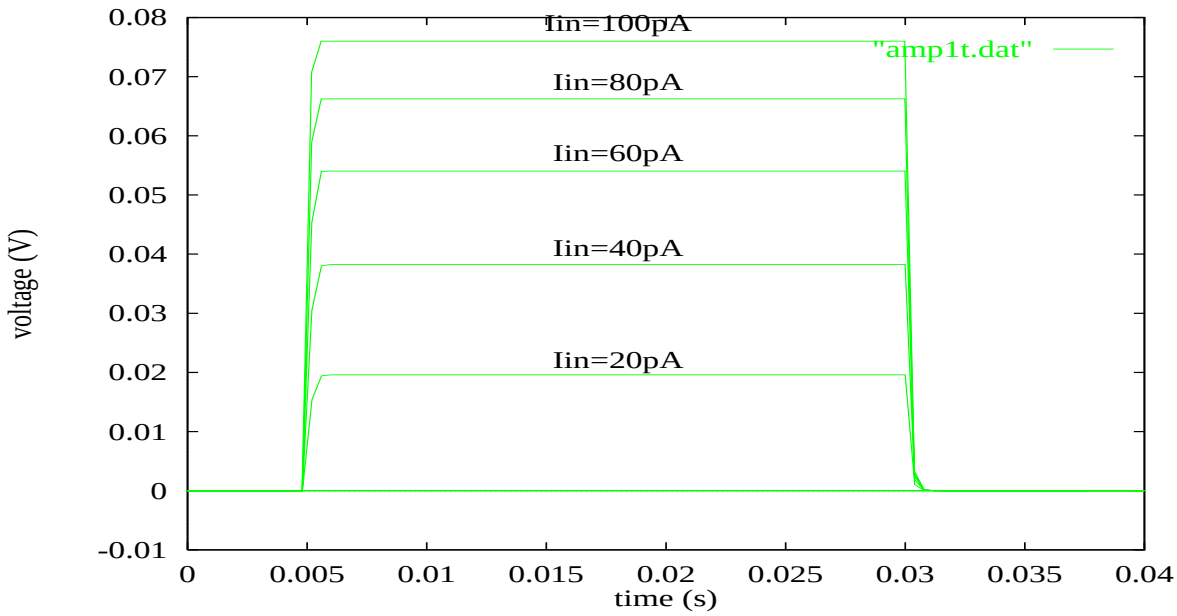

Figure 6.3: Transient response of the circuit of fig. 2

Though the circuit of fig. 2 has a very good gain-area trade-off, its practical use entails one serious issue. In fact, the small signal gain of the circuit heavily depends on the value of the voltage applied on the gate of the feedback transistor. From eq. 4 it is easy to calculate that a variation in V_{GS} of only 27 mV can change the equivalent resistance and thereby the gain by a factor 2. Therefore, a very tight control on this voltage is needed in order to assure a reliable operation. Moreover, the input signal moves up the output node, hence changing the gate-source voltage on the feedback device; the gain of the circuit is dependent on the value of the input signal and the linear dynamic range is quite poor. In fact, from fig. 3 we can argue that the voltage step response to an input of 20 pA is 20 mV, which is in good agreement with the expected gain of 1 GΩ; however, the response to an input of 100 pA in only 76 mV and the linearity error in the range [20 pA, 100 pA] is 25 %.

Actually, the technique of fig. 2 is widely used to build compact resistor for charge sensitive amplifier, where the resistor is used just to provide a DC feedback path for the core amplifier and to slowly discharge the feedback capacitor after a pulse has been detected. However, the configuration is not very well suited to mimic a transimpedance function.

In principle, a simple way of improving the circuit in fig. 2 is to add in the feedback transistor a small bias current, as depicted in fig. 5. Since the current mirror forces a current into the feedback device, the output node of the amplifier follows the gate voltage of the feedback transistor and a shift in this voltage translates in a offset, but not in gain variation. The small signal gain of this circuit is defined by $\frac{1}{g_{mf}}$ where g_{mf} is the transconductance of the feedback device. If we want high gain, we must operate the feedback with very small current and the hypothesis that the transistor works in the weak inversion region is justified. The g_{mf} can be evaluated as

$$\frac{I_f}{nV_t} \tag{6.5}$$

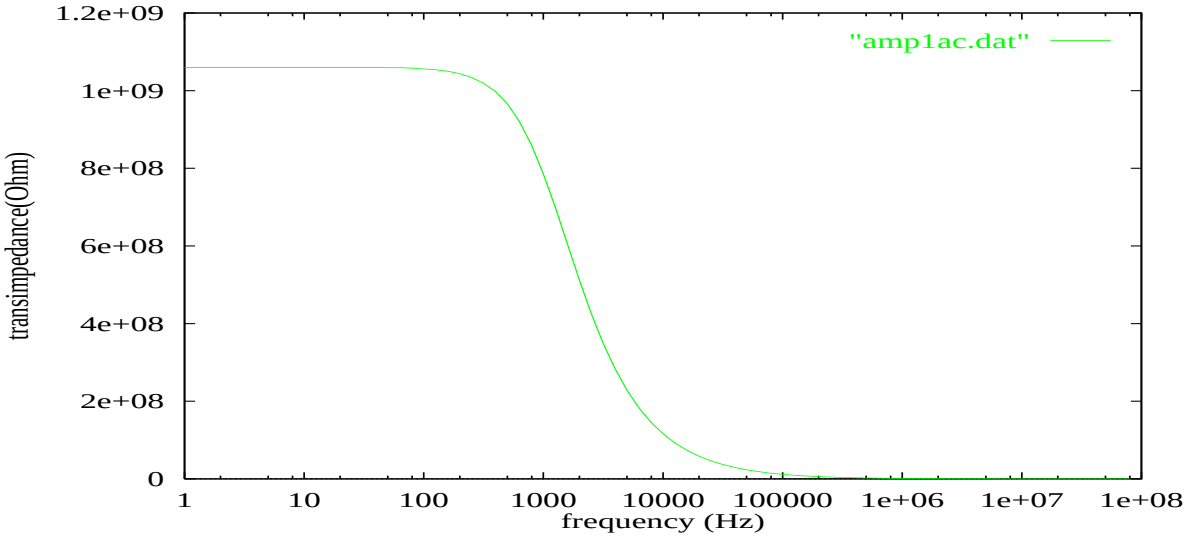

Figure 6.4: Small signal response of the circuit of fig. 2

where I_f is the dc bias current in the feedback branch and the other parameters have been defined above.

feedback

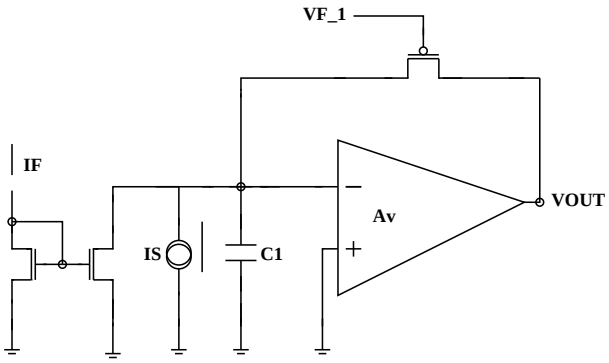

Figure 6.5: "Linearized" transimpedance amplifier with single MOS transistor feedback

The current I_f "linearises" the circuit in the sense that if the variation induced by the input signal is small compared to I_f, the behaviour of the circuit can be regarded as linear. If we want to have a nonlinearity for at most 10 % for a 100 pA input, we can calculate from eq. 5 that we need a bias current of at least 1 nA; the transimpedance gain would be in this case 40 MΩ which fairly agree with the SPICE simulation shown in fig. 6, indicating a gain of 49.5 MΩ. Therefore, the gain achievable with the circuit of fig. 5 is by far too low for our application.

6.1.2 OTA feedback

An interesting alternative for the integration of very large equivalent resistor consists in introducing a complete Operational Transconductance Amplifier (OTA) in the feedback path [40],

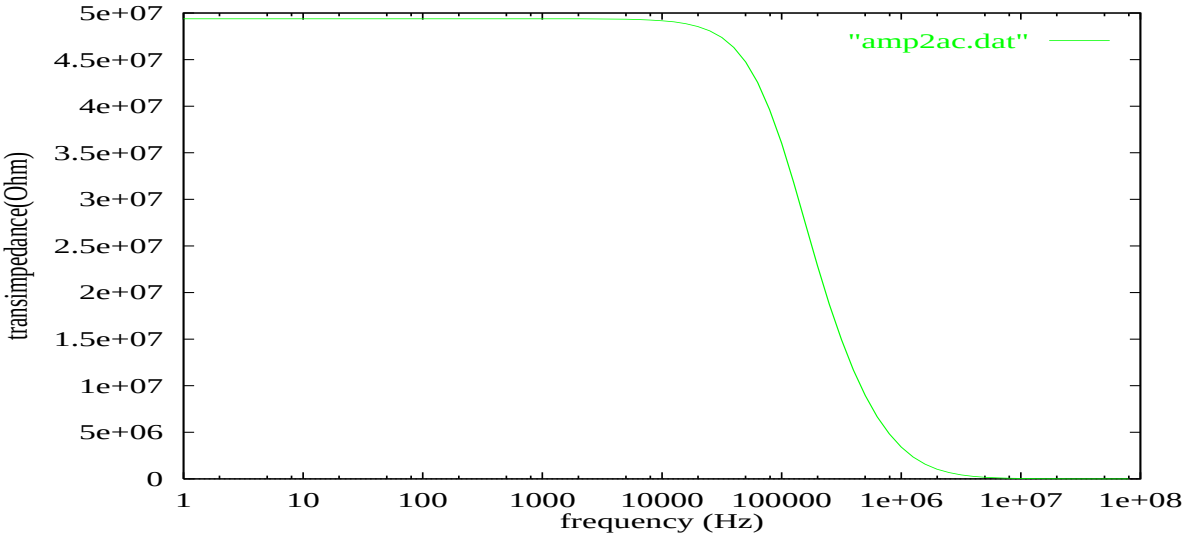

Figure 6.6: Small signal gain of the circuit of fig. with I_f=1 nA

as shown in fig. 7. In principle, very high transimpedances can be emulated by properly scaling the ratio of the current mirrors M2A - M2B, M3A - M3B.

To determine the small signal low-frequency gain of this configuration, we refer to the small signal equivalent circuit of fig. 8. In this design:

- A_v is the open-loop gain of the core amplifier
- g_{mf} is the transconductance of the input differential pair of the OTA
- g_{m2} is the transconductance of the diode connected transistor M2A, M3A
- g_{m3} is the transconductance of the output transistor M4B and R_3 its output resistance
- i_{in} is the input signal current ad R_s is the output resistance of the sensor. This resistance is supposed to be very high an therefore is neglected in the calculation.

The output voltage of the amplifier is defined by

$$v_0 = A_v v_1 \tag{6.6}$$

The voltage v_1 is the "error voltage" at the input of the core amplifier and can be expressed in the following way

$$v_1 = R_3 i_\varepsilon \qquad\qquad i_\varepsilon = i_{in} - i_f \tag{6.7}$$

where i_ε is the "error current" and i_f is the feedback current. The feedback current i_f, in turn, can be written as

$$i_f = g_{mf}\frac{g_{m3}}{g_{m2}} \tag{6.8}$$

Using the above equations v_0 can be expressed only in function of the input current

$$v_0(1 + g_{mf}\frac{g_{m3}}{g_{m2}}) = A_v R_3 i_{in} \tag{6.9}$$

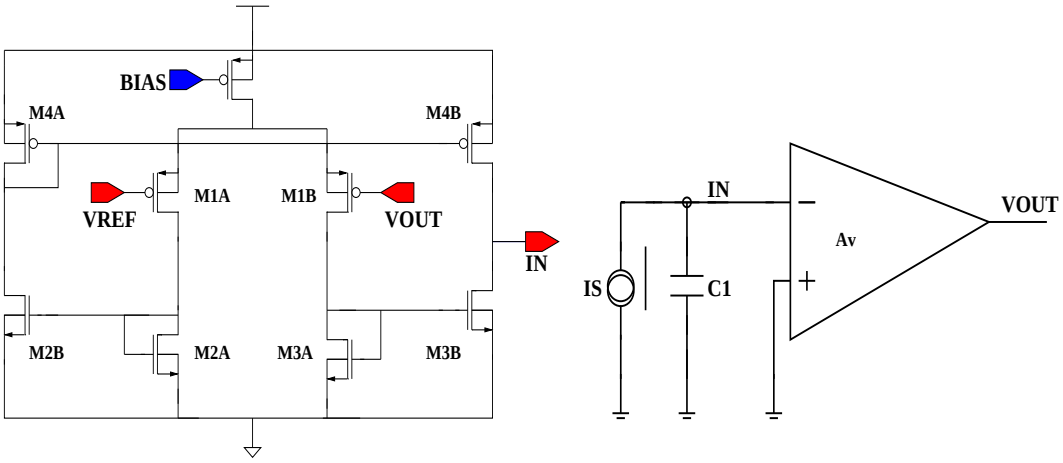

Figure 6.7: Transimpedance feedback implemented with an OTA stage

From eq. 9 it is easy to calculate the ratio between the output voltage and the input current, which is

$$\frac{v_0}{i_{in}} = \frac{A_v R_3}{1 + A_v R_3 g_{mf} \frac{g_{m3}}{g_{m2}}} \tag{6.10}$$

Under the hypothesis that $A_v R_3 g_{mf} \frac{g_{m3}}{g_{m2}} \gg 1$, eq. 10 is simplified as following

$$\frac{v_0}{i_{in}} = \frac{1}{g_{mf} \frac{g_{m3}}{g_{m2}}} = \frac{1}{k g_{mf}} = R_f \tag{6.11}$$

where $k = \frac{g_{m3}}{g_{m2}}$.

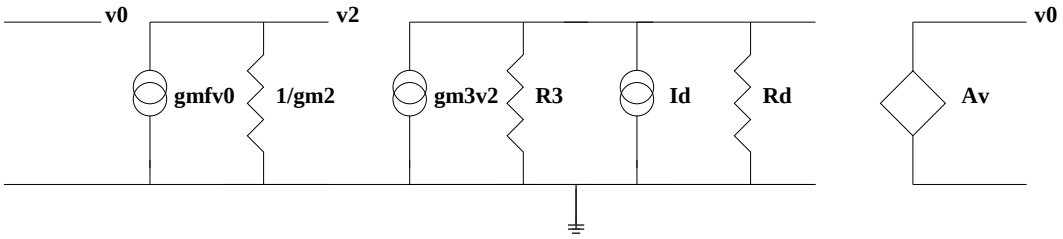

Figure 6.8: Low frequency small signal model of the circuit of fig. 8

Therefore the system composed by the core amplifier and the OTA acts as a transimpedance amplifier, whose gain can be estimated using eq. 11; very high equivalent resistances can be obtained by properly sizing the transconductance of the input stage of the OTA and the term k. It is interesting to observe that k actually depends on the ratio of the *size* between the transistors M_{3A} and M_{3B} and hence is a *linear* term also for large signal swings. The limitation on the overall linearity will then be determined, in first approximation, by the transconductance of the input stage of the OTA.

Since we aim at very high transimpedance, the term $g_{mf} \frac{g_{m3}}{g_{m2}}$ will be very small (10^{-9} for 1 GΩ transimpedance gain). However, making this term small requires also a very asymmetric mirroring factor between M3A and M3B (and of course M4A and M4B). As a consequence,

the current flowing in the output transistors will be very small as well, thereby rising R_3; the approximation that $A_v R_3 g_{mf} \frac{g_{m3}}{g_{m2}} \gg 1$ is so justified, provided that the open-loop gain of the core amplifier is high enough. For instance, for a transimpedance gain of 1 GΩ, R_3=50 MΩ and A_v=2000 are sufficient to give an error smaller than 1% in the approximation.

The minimum noise achievable by the topology of fig. 7 is limited by the parallel noise of the transistors in the output branch. Since a very small current flows in these devices, is natural to assume that they work in the weak inversion region. Their current noise spectral density is then defined as

$$i_n = 4kT\frac{1}{2g_m} = 4kT\frac{1}{\frac{I_d}{nV_t}} = 3qI_d \tag{6.12}$$

where I_d is the current flowing in M3B, M4B. This current must be significantly bigger than the leakage current of the transistors and possibly of the dark current of the sensor and hence can not be reduced below few hundreds pA. Assuming a value of 200 pA and applying eq. 12[4], we calculate that the noise contribution of M3B, M4B is 0.5 pA r.m.s; this value is remarkably close to the one furnished by the SPICE simulations (0.58 pA r.m.s.), showing that the current noise of the output stage places an ultimate limit to the minimum possible noise. The only way of decreasing this limit without reducing the current in M3B, M3A is by reducing the bandwidth. Actually, in our application bandwidths of few tens of Hertz are adequate and if the bandwidth is reduced, for instance, to 100 Hz, the noise scales to 0.16 pA r.m.s.

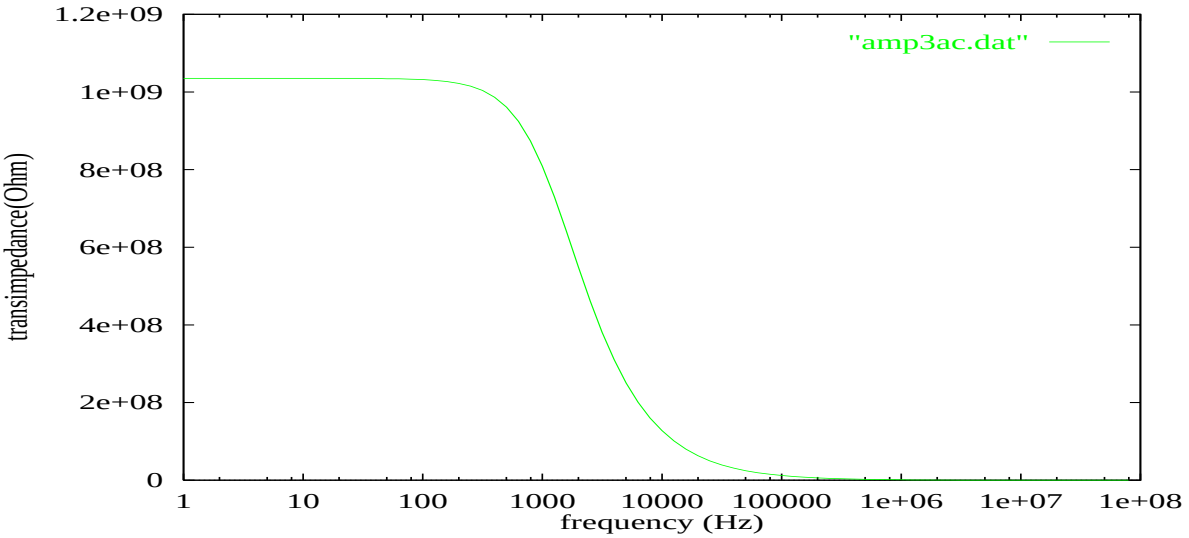

Figure 6.9: Small signal gain of the circuit of fig. 7

Fig. 9 shows an ac simulation of the circuit of fig. 7; the DC gain of 1.05 GΩ is reached by sizing the ratio of M3A to M3B and M4A to M4B equal to 20000/1. Tough this ratio may seem abnormal, values up to 40000/1 are found in the literature for similar applications [40]. The plot in fig. 10 shows an example of the transient performance of the circuit. The improve in the linearity is clear; a small distortion (about 4 %) starts appearing from 100 pA. This distortion

[4]We stress again that the system has a first order low-pass transfer function with a bandwidth of 1.5 kHz)

is in fact determined by the input differential pair of the OTA, since its transconductance enters directly in the expression of the equivalent feedback resistance.

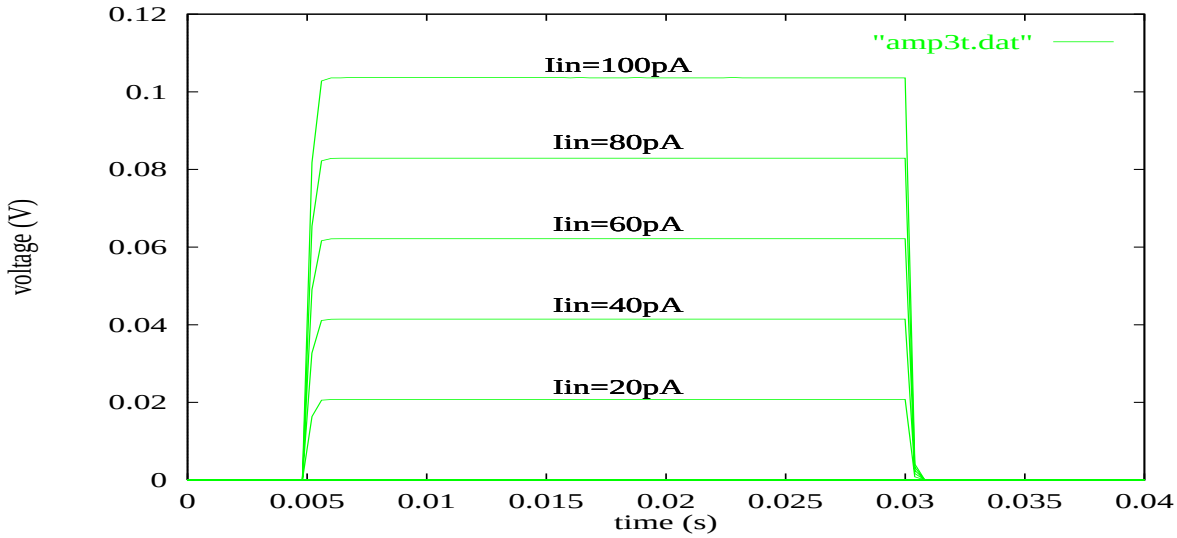

Figure 6.10: Transient behaviour of the circuit of fig. 7

Non linearities in MOS differential stages can be alleviated, as for the bipolar stages, by inserting *degeneration* resistors on the source of the transistors. Unfortunately, the MOS transistor has a poorer transconductance with respect its bipolar counterpart and the degeneration technique, to be effective, requires either the use of large resistor or of large bias currents in the transistors in order to increase the g_m.

A much more effective solution is to *boost* the transconductance of the input differential pair with an auxiliary stage. Many configurations are possible, but for our applications the use of a simple common source amplifier has been sufficient [41]. The resulting circuit is depicted in fig. 11.

Let I_1 be the bias current flowing in M1A, M1B and I_2 the bias current flowing in MB1, MB2. The current flowing in the boosting transistors M3A, M3B is $I_1 - I_2$ and is mirrored to the output branch, composed by M4B M5B. When an input signal v_{in} is applied to the OTA, it determines a current defined by

$$i_{in} = \frac{v_{in}}{R_d + \frac{2}{g_{mf}(1+A_0)}} \frac{A_0}{1+A_0} \tag{6.13}$$

where A_0 is the gain of the boosting amplifier.

Eq. 13 shows that the bigger A_0 the smaller the influence of the transconductance of the differential pair on the gain. The transimpedance gain of the overall circuit will therefore be $R_f = \frac{k}{2} R_d$, where k is the mirroring ratio between M3B and M4B.

In order to get a transimpedance of 1 GΩ with the circuit of fig. 11, a degeneration resistance of 100000 $k\Omega$ and a mirroring factor of 20000 are necessary. The linearity has a big improvement (basically no distortion has been seen in the range [20pA, 200pA]), at the expense

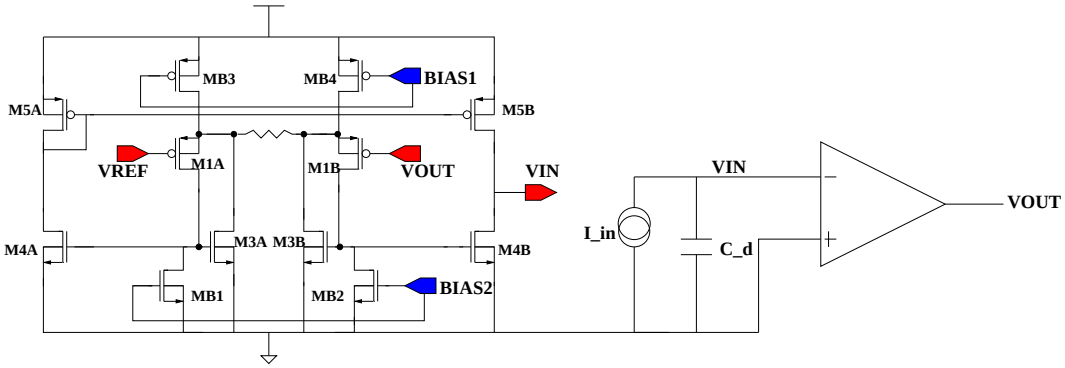

Figure 6.11: Transimpedance feedback implemented with an linearized OTA stage

of introducing a further component of remarkable size. However, this circuit has the big advantage that the gain is not dependent on any bias current or voltage, since its rely only on a passive component and the ratio between transistor sizes.

In principle, an alternative to reduce the value of the degeneration resistor and/or the mirroring factor is to introduce a voltage partitioning at the output of the core amplifier and to apply to the feedback OTA only one fraction of the total output swing. With this technique, using a degeneration resistor of 10 kΩ, a k factor of 1000 and feeding the OTA with 1/200 of the input voltage swing, an equivalent transimpedance of 1 GΩ is obtained as well.

However, the great area saving is paid by an increase in the overall noise; in fact, the attenuation introduced by k acts not only on the signal, but also on the noise, thereby reducing the noise contribution of the OTA. Hence, if very high sensitivity are aimed, very large attenuation factors must be used. The system of fig. 2 with a R_d of 100000 $k\Omega$ and a k of 20000 has a total input referred noise of 0.7 pA r.m.s, which jumps to 2.5 pA r.m.s in case the "compact" version is used. Of course, any intermediate solution is possible, depending on the need of the application. The last important remark is that, since the mirroring ratio is very asymmetric, it would be also not precise; so only the order-of-magnitude of the gain can be estimated in the simulation. Moreover, the gain mismatch between different amplifiers can be huge. However, the gain is *stable* an these problems can be easily solved introducing on-chip gain tunability.

6.1.3 Technological considerations

As we have seen in chapter 1, a deep submicron technology (of the generation 0.25 -0.35 μm) may offer better analog performances if proper design strategies are used; the layout is also more compact even for transistors with gate length far from the minimum size.

Nevertheless, in our circuit low-frequency noise is an issue and care must be paid in choosing a technology with good $\frac{1}{f}$ noise parameters. The estimated layout of the circuit of fig. 11 is 0.3 mm^2 in a 0.25 μm technology and 0.6 mm^2 in a 0.7 μm process. The latter is however economically more convenient, since the price per mm^2 is much less for a 0.7 μm than for a 0.25 μm process.

Additionally, in more conventional processes, analog libraries are available and this is an advantage if a more complex system (including, for instance, A/D converters) has to be implemented around the full-custom block. The capability of the technology to work with 5 V power supply is also important, since it makes easier the interface between the part and the measurement an testing equipment.

6.2 Current to frequency conversion

As we have seen at the beginning of this chapter, for the final application a charge digitizing system would probably be preferable. In order to lower the costs, this circuit should be as simple as possible; hence we have considered a solution based on the architecture described in [38].

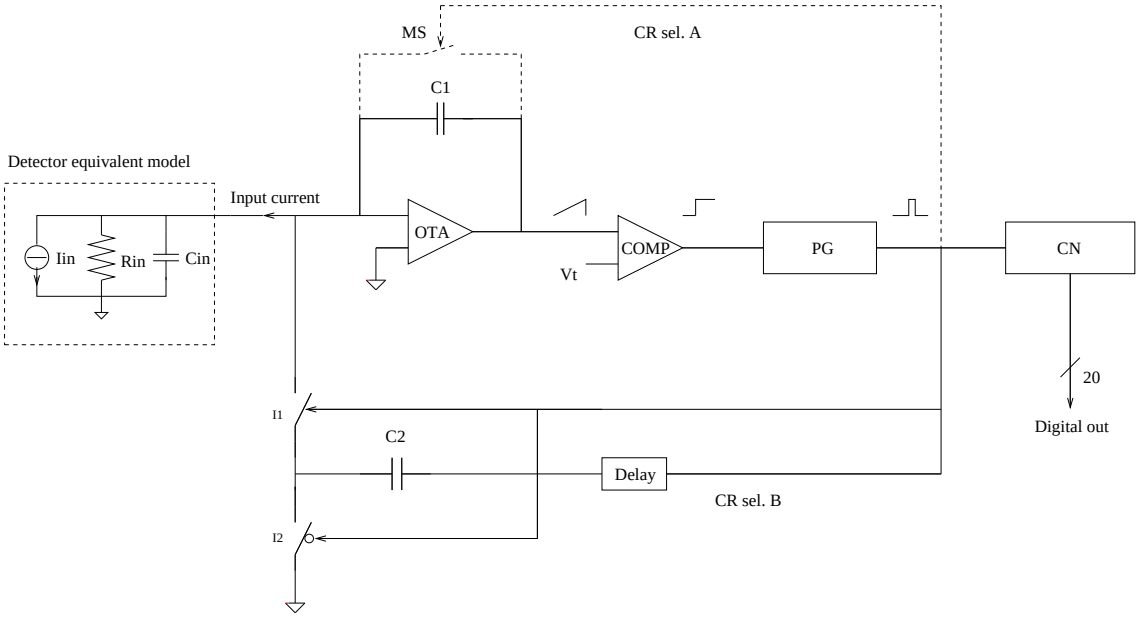

Figure 6.12: Block diagram of the charged digitizing circuit

As can be seen in fig. 12 the input stage of the circuit is an integrator, built by inserting a capacitor in the feedback path of an operational transconductance amplifier. The input current is integrated onto C_1 and whenever the voltage at the output of the op-amp exceeds a given threshold, the comparator (COMP in the figure) asserts a logical 1. This signal feeds a digital circuit, that generates a 100 ns pulse; the pulse is sent to an asynchronous counter (CN) and to a reset network, which subtracts to the integrator a fixed amount of charge. The digital number at the output of the counter indicates how many times the quantum of charge has been subtracted and hence is a measure of the total charge accumulated on C_1. If the time of the integration is known, the mean value of the input current can be deduced.

A remarkable feature is that the subtraction mechanism does not determine any dead time and the system is continuously operating. It is worth noting that in this scheme the quality of the analog components is not particularly critical and hence a simple and compact design results. Moreover, most of the circuit is composed by digital blocks and an implementation in a deep submicron technology would save considerable amount of silicon.

In the original implementation, the circuit was working width a feedback capacitor of 600 fF and a threshold of 500 mV in the comparator. Clearly, the frequency of the operation speeds up if the size of the feedback capacitor is reduced and if the threshold voltage is decreased. The system proved to be very reliable and input currents down to 20 pA could be measured [42]. We have therefore considered this design very promising for our application; an upgraded version, in which also the photodiode is integrated on the chip has been submitted to the foundry in November 1999.

6.3 Summary

In this chapter some circuits suitable for the measurement of very low currents have been described. The work has been carried-out in the framework on a R & D project whose aim is to build a fully integrated read-out chain for electrochemiluminescence sensors in standard CMOS technologies.

A first theoretical analysis, supported by computer simulations, shows that the integration of high-gain transimpedance amplifiers in a reasonable silicon area is feasible.

The transimpedance function is emulated by an operational transconductance amplifier or by an a linearized g_m stage and the desired gain is achieved by current division techniques. For an optimal design (i.e. minimizing the contribution of all the noise sources) the minimum achievable noise is determined by the parallel noise contribution of the transistors in the output branch of the OTA. In order to assure a reliable operation, this current can not go below few hundreds pA; assuming a minimum limit of 200 pA, an input noise spectral density of 10 fA/$\sqrt{Hz}$ results as the best possible noise figure. Depending on the architecture and on the required bandwidth (which, in any case, would not exceed few kHz) a total input r.m.s noise between 100 fA r.m.s and 1 pA r.m.s. is expected.

This amplifier would be an useful tool in the development phase, while for the final circuit implementation a system based on current integrating technique should be preferred.

A first prototype of this second circuit, integrating on board also the photosensor, has been recently submitted to the foundry.

7 Conclusions

This work has explored some design issues related to the use of sumbicron and deep-submicron CMOS technologies in analog design, with emphasis on the circuits used for the read-out of silicon sensors. The studies have been carried out investigating the analog properties of the transistors and designing some representative circuit. In fact, if the characterisation of basic devices already provide some insight, the extrapolation from the transistor level to the circuit level is not always straightforward and the final analog performance can be better assessed by implementing and testing circuits which have to face specific applications.

The theoretical analysis, supported by experimental measurements, has shown that superior analog performance is obtained in deep-submicron technologies, provided that non minimum gate length are used in the design. The noise figure is very good if the transistors are working in the weak or in the moderate inversion region. An excess noise has been observed in strong inversion , when high currents are flowing in the device, especially for short channel length. This phenomenon has been traced to short channel effects (like hot carriers, impact ionization and weak avalanche), which are exacerbated when the channel length is squeezed. However, moving towards deep submicron processes the gate oxide becomes thinner and thinner and the "technological parameter" K=μC_{ox} increases; as a consequence, the transistor enters the strong inversion region for higher current densities and the weak and moderate inversion can be better exploited. Furthermore, in weak inversion the drain-source saturation voltage is of the order of 200 mV, which makes easier the use of cascode configurations even at low power supply.

The design of the system for the Silicon Drift Detectors has required a consistent preliminary work, in order to the determine the optimal architecture. This investigation has lead to the conclusion that the preamplification and analog to digital conversion functions should be merged on the same chip. The analog to digital converter has been considered as the critical block and has been investigated in great detail, by designing and testing two prototypes. These efforts showed the feasibility of the proposed architecture and a chip implementing on the same substrate the amplifier, the analog memory and the analog to digital converter in presently under design.

From a more general point of view, the ADC was also an excellent test vehicle to compare analog performance of two CMOS technologies of different generation (0.7μm and 0.25μm of minimum gate length). The same architecture, based on the switched capacitor charge redistribution approach, has been used for both converters. In the more advanced process the same resolution is obtained with half the area and dissipating only 30% of the power. This result is due the great squeezing in the size of the control logic; furthermore, the improvement in the matching of the transistors for the 0.25μm technology allows a simpler and more compact design of the comparator. However, even in the analog part, where very conservative sizing of transistors was used, the reduction in area occupation is significant, thanks to the scaled design rules (e.g. distance between metals, contacts, etc).

This part of the work as been done in close contact with the Microelectronics Group of the European Laboratory for Particle Physics, in Geneva. Actually, the ADC implemented in the deep-submicron process serves also as a demonstrator for the RD49 collaboration, which is investigating the use of standard commercial technologies in environment with high level of ionizing radiation, like high energy physics experiments or space applications. By systematically using enclosed layout geometries and guardrings around NMOS transistors, we were able

to stand radiation level of 10 Mrad (SiO_2) without measuring any degradation in the parameters of the circuit.

The analysis carried out on the photodetector front-end showed that very high gain transimpedance amplifiers can be fully integrated with reasonable noise performance; this circuit would be useful in the intermediate steps of the project and to perform some particular measurements. For the final application, a charge digitizing system would probably be preferable. A preliminary system, based on an existing design and integrating also the sensor, has been recently submitted to the foundry.

It is a quite common opinion between analog designers that the scaling of the technology automatically worsens the performance of analog circuits. Actually, the experience of this work suggests the situation is more complicated, since the reduction in dynamic range caused by the squeezed power supply is alleviated by the improvement in the characteristics of the transistors. We think that high performance analog circuits can be designed in present deep-submicron processes, even employing conventional architectures.

8 References

[1] M. J.M. Pelgrom and M. Vertgret, “CMOS Technologies for Mixed Signal ICs”, *Solid-State Electronics*, vol. 41, no. 7, pp. 967-974, 1997.

[2] Y. P. Tsividis, “Operation and Modeling of the MOS transistor”, McGraw-Hill, 1999.

[3] F. Maloberti, F. Francesconi, P. Malcovati and O. J. A. P. Nys, “Design Considerations on Low-Voltage Low-Power Data Converters”, *IEEE Journal of Solid State Circuits*, vol. 42, no. 11, Nov. 1995.

[4] K. R. Laker and W. Sansen, “Design of analog integrated circuits and systems”, McGraw-Hill, 1994.

[5] A. Rivetti et al.”Analog Design in Deep Submicron CMOS Processes for LHC”, *Proceedings of the fifth workshop on electronics for the LHC experiments*, Snowmass, Colorado. Sept. 1999.

[6] Y. Chang and W. Sansen, “Low-Noise wide-band amplifiers in bipolar and CMOS technologies”, Kluwer Academic Publishers, 1991.

[7] S. Tedja et al. “Noise spectral density measurements of a radiation hardened CMOS process in the weak and moderate inversion”, *IEEE Transactions on Nuclear Science*, vol. 39, n0. 4, pp. 804-808, 1992.

[8] M. Pelgrom et al. “Matching Properties of MOS Transistors”, *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, October 1989.

[9] T. Smerbeck, “Practical Aspects in Analog and Mixed-Mode ICs”, Course Notes, EPFL, Lausanne, September 1999.

[10] E. Gatti and P. Rehak, “Silicon Drift Chambers - First results and optimum processing of signals”,
Nuclear Instruments and Methods in Physics Research, A226 (1984), pp 129-141

[11] E. Gatti, A. Longoni, M. Sampietro and A. Castoldi, “Electron injection in semiconductors drift chambers”, *Nuclear Instruments and Methods in Physics Research*, A295 (1990), pp 489-491

[12] E. Gatti, P. Rehak and M. Sampietro, “Double particle resolution in Semiconductor Drift Detectors”, *Nuclear Instruments and Methods in Physics Research*, A274 (1989), pp. 469-476

[13] W. Dabrowski et al., “OLA, A low-noise bipolar amplifier for the read-out of Silicon Drift Detectors” *Nuclear Physics B*, vol. 44, pp. 637-641, 1995.

[14] G. Gramegna, P. O'Connor, P. Rehak and S. Hart, "Low-noise CMOS preamplifier-shaper for Silicon Drift Detectors", in *Proceedings of the Second Workshop on the Electronics for the LHC experiments*, Lake Balaton, September 1996.

[15] B. Brandt, "High-speed data converters", Course Note, EPFL, Lausanne, September 1998.

[16] A. Werbrouk, private communication.

[17] S.R. Klein et al., "Front end electronics for the STAR TPC", *IEEE Transactions on nuclear science*, vol. 43, no.3, June 1999.

[18] M. French, M. Morrisey, T. Pritchard and S. Warren, "AROW - Test results of a 128 channel analogue pipeline and Wilkinson ADC with sparsification IC for ATLAS", *Proceedings of the Second Workshop on the Electronics for the LHC experiments*, Lake Balaton, September 1996.

[19] V. Bonvicini, P. Burger, A. Gregorio, A. Rashevski, A. Vacchi and N. Zampa, "Characterising large area silicon drift detectors with MOS injectors", *Il Nuovo Cimento*, Vol. 112 A, n. 1-2, Jan-Feb 1999.

[20] The ALICE collaboration, *Technical Design Report of the Inner Tracking System*, CERN/LHCC, 99-12, June 1999, pp. 83 - 173.

[21] The ALICE collaboration, *ALICE technical proposal*, CERN/LHCC, 95-71, Dec. 1995.

[22] Gunther M. Haller, "High-speed, high-resolution analog waveform sampling in VLSI technology", PhD thesis, Standford University, 1994.

[23] J. L. McCreary and P. R. Gray, "All-MOS Charge Redistribution Analog-to-Digital Conversion Techniques-Part I", *IEEE Journal of Solid State Circuits*, vol. SC 10, no. 6, pp. 371-379, Dec. 1975.

[24] B. Razavi, "Principles of Data Conversion System Design", IEEE Press, NewYork, 1995.

[25] D. A. Johns and K. Martin, "Analog Integrated Circuit Design", John Wiley & Sons, 1997.

[26] G. M. Lyn, F. Op't Eynde, and W. Sansen, "A High Speed CMOS Comparator with 8-b Resolution", *IEEE Journal of Solid State Circuits*, vol. 27, no. 2, Feb.1992.

[27] V. Valencic et al., "A Low-Power Piecewise Linear Analog to Digital Converter for use in Particle Tracking", *IEEE Transactions on Nuclear Science*, vol. 42, no. 4, Aug. 1995.

[28] H. Yoshizawa, Y. Huang, P. F. Ferguson and G. Temes, “Mosfet-Only Switched-Capacitor Circuits in Digital CMOS Technology”, *IEEE Journal of Solid State Circuits*, vol. 34, no. 6, June 1999.

[29] M. J. M. Pelgrom, H. P. Tuinhout and M. Vertregt, “Transistor matching in analog CMOS applications”, *Technical Digest of the International Electron Devices Meeting 1998,* San Francisco, Dec. 1998.

[30] G. A. Miller, “An 18 b 10 μs Self-Calibrating ADC”, *ISSCC Dig. Tech. Pap.*, pp. 168-169, Feb. 1990

[31] C. Henz and G. Temes, “Circuit Techniques for Reducing the Effects of Op-Amp Imperfections: Autozeroing, Correlated Double Sampling, and Chopper Stabilisation”, *Proceedings of the IEEE*, vol. 84, no. 11, Nov. 1996.

[32] P. Jarron, “Radiation Tolerant Electronics for the LHC experiments”, *Proceedings of the fourth Workshop on Electronics for LHC Experiments*, Rome, Sept. 1998.

[33] M. Campbell et al., “A Pixel Readout Chip for 10-30 MRad in Standard 0.25 μm CMOS, *IEEE Transactions on Nuclear Science*, vol. 46, no. 3, June 1999.

[34] W.Sansen, “Low-Voltage, Low-Power Analog IC Design”, Course Notes, Lausanne, June 1998.

[35] B. Fotouhi and D. A. Hodges, “High Resolution A/D Conversion in MOS/LSI, *IEEE Journal of Solid State Circuits*, vol. SC-14, pp. 920-26, Dec. 1979.

[36] Y. P. Tsividis et al., “A segmented μ-255 Law PCM Voice Encoder Utilizing NMOS Technology, *IEEE Journal of Solid State Circuits*, vol.SC-11, pp. 740-747, Dec. 1979.

[37] F. Brianti, A. Manstretta and G. Torelli, “High-speed autozeroed CMOS comparator for multistep A/D conversion”, *Microelectronics Journal*, no. 29,pp. 845-853, 1998.

[38] G. C. Bonazzola and G. Mazza, “ A VLSI circuit for charge measurement of a strip ionization chamber”, *Nuclear Instruments and Methods in Physics Research*, A 409, pp.336 - 338, 1998.

[39] Philips Semiconductors, “A family of wide band low noise transimpedance amplifiers” Application note AN1435, Dec. 1991

[40] M. Steyaert, P. Kinget and W. Sansen, “Full Integration of Extremely Large Time Constants in CMOS”, *Electronics Letters*, vol. 27, no. 10, May 1991.

[41] J. J. F. Rijns, “CMOS Lo-Distortion High-Frequency Variable-Gain Amplifier” *IEEE Journal of Solid-State Circuits*, vol. 31, no. 7, July 1996.

[42] G. Mazza, private communication.