Università degli Studi di Torino

Scuola di Dottorato in Scienza e Alta Tecnologia

Indirizzo di Fisica e Astrofisica

XXVII Ciclo

Development of Integrated Pixel Front-End Electronics in 65 nm CMOS Technology for Extreme Rate and Radiation at HL-LHC

Luca Pacher

Candidate

Luca Pacher

Supervisor

Dr. A. Rivetti
Istituto Nazionale di Fisica Nucleare (INFN)
Sezione di Torino

External Reviewer

Prof. G. Traversi
Dipartimento di Ingegneria e Scienze Applicate
Università degli Studi di Bergamo

Ph.D. Tutor

Prof. M. Costa
Dipartimento di Fisica
Università degli Studi di Torino

Ph.D. Coordinator

Prof. M. Gallio
Dipartimento di Fisica
Università degli Studi di Torino

Date of dissertation

April 20, 2015

Signature from head of Ph.D. committee

_____

All'unica persona a cui avrei voluto poter dedicare queste pagine.
Grazie davvero...

Now I know we said things, did things that we didn't mean
And we fall back into the same patterns, same routine.
But your temper's just as bad as mine is,
You're the same as me.
When it comes to love you're just as blinded.
Baby, please come back it wasn't you, baby it was me.
Maybe our relationship isn't as crazy as it seems
Maybe that's what happens when a tornado meets a volcano...

Contents

Abstract

The recent discovery at the CERN Large Hadron Collider (LHC) of resonances consistent with a Higgs boson with a mass of about 125 GeV/c^2 gives today a compelling and concrete case to define and justify future exploration strategies. As a matter of fact, the LHC will remain the most powerful accelerator ever built in the world for at least two decades. In order to fully exploit the discovery potential of this machine, the foreseen LHC research program extends over more than 20 years, investigating and validating the mathematical consistency of the Standard Model (SM) at the TeV energy scale and searching for evidences of new physics signatures.
According with the expected increase of the machine performance in the forthcoming years, several upgrade projects involving existing LHC general-purpose detectors ATLAS and CMS have already started. Detailed RD programs are necessary in order to explore and develop new cutting-edge detector technologies and dedicated readout electronics. This is mandatory since the typical time scale required for the design, construction and commissioning of such large and complex systems is of the order of 5 to 10 years.

In particular, the CMS experiment tracking community is already engaged in long-term upgrade plans (∼2022-2023) when the foreseen High-Luminosity (HL) LHC operating conditions will impose significant upgrades for the inner tracking system, demanding the installation of a new silicon pixel detector. With increased performance the machine will deliver *pp* collisions with an instantaneous luminosity of the order of 10^{35} cm^{-2}s^{-1}, one order of magnitude higher with respect to the current design value. The upgrade aims to reach an integrated luminosity of 3000 fb^{-1} in 10 years, providing a chance to access extremely rare physics processes.
New LHC operating conditions introduce several challenges in the design of the new pixel detector. With such a luminosity and a centre-of-mass energy of 14 TeV, the nominal collision rate of 40 MHz will lead to unprecedented pileup (up to 200), introducing extreme rates and radiation levels. More layers equipped with sensors featuring high granularity, speed and adequate radiation hardness will be required. Hence hybrid silicon pixel detectors will continue to play a fundamental role.
The innermost pixelated layer will have to cope with unprecedented Total Integrated Dose (TID), up to 10 MGy in 10 years. Smaller pixels of the order of 50 μm × 50 μm will be required to maintain high spatial resolution and tracks separation under high pileup conditions. The particle flux will increase to about 500 MHz/cm^2, leading to unprecedented hit rates (1-2 GHz/cm^2) and an estimated rate per pixel of the order of 50-100 kHz in the innermost layer. Thus a development plan devoted to the design of a new pixel Application Specific Integrated Circuit (ASIC) for CMS has started. More on-chip intelligence will be introduced to deal with the increased hit rates. The expected increased L1 trigger latency (up to 20 μs) will require more local data storage capabilities, efficient zero-suppression schemes and higher output bandwidths.
The foreseen usage of thinner sensors of the order of 100-150 μm to increase the radiation tolerance determines reduced signals, needing low-noise (ENC < 150 e^- RMS at nominal 100 fF total input capacitance) and low-threshold (about 1000 e^- minimum detectable charge) performance for the pixel analogue Front-End. Moreover, an in-time response below 25 ns is required in order to cope with the nominal LHC bunch interaction rate, while keeping bias currents to acceptable values and targeting to a maximum total power dissipation of 10 μW/pixel.

Different approaches for both the analogue signal processing (preamplification and shaping) and the charge digitization (binary-only readout, time-over-threshold techniques, usage of local or shared A/D converters) are under consideration. However, in order to fully exploit speed and integration densities offered by ultra-deep submicron (UDSM) technologies, most of the signal processing will be performed into the digital domain, with a chance to move as much as possible temporary data storage (buffering) and trigger matching from the chip periphery to the pixel level.
Radiation tolerance and higher integration level constraints led to the choice of a commercial 65 nm CMOS technology as the presently favoured IC fabrication technology for the design of the new pixel chip. At present 65 nm represents the most advanced technology node adopted to implement full-custom solutions for radiation detection and measurements in particle physics and medical imaging applications. Technology qualification and radiation hardness studies using this 65 nm CMOS process are now part of the international RD53 collaboration research program officially supported by CERN, that involves both ATLAS and CMS pixel ASIC communities as well as other non-LHC experiments and groups interested in designing in 65 nm. Furthermore, Italian CMS and ATLAS groups have submitted in July 2013 a detailed proposal to INFN/CSN5 to finance a new three-years research program on CMOS 65 nm, leading to the approval of the CHIPIX65 project.

This thesis presents my personal contributions on the design and test of both analogue and digital integrated circuits in such a commercial 65 nm CMOS technology. All design activities were carried out with extensive usage of professional and industry-standard Computer Aided Design (CAD) softwares for circuit simulation and mask layout provided by the VLSI Design Laboratory of the Torino section of INFN. The entire research work has been supported by the CHIPIX65 collaboration and the Torino CMS Tracker group.
Preliminary pixel Front-End test structures, small pixel matrix prototypes and other analogue, digital and mixed-signal building blocks have been submitted on October 2014 by the CHIPIX65 collaboration to the foundry access service. They were received back from the manufacturer for laboratory test measurements and bench characterizations at the beginning of 2015. These studies have provided the necessary first steps towards the design of a future complete hybrid pixel ASIC demonstrator suitable for the long-term CMS pixel detector upgrade.

Part I
Background and motivations

Chapter 1

ASIC requirements for the CMS pixel detector upgrade at HL-LHC

An extrapolation to HL-LHC unprecedented operating conditions of state of the art integrated circuit solutions for the readout of hybrid pixel detectors introduces major challenges on several fronts. This chapter provides necessary background information and motivations to frame the overall Ph.D. research activity, devoted to the development of first prototypes of integrated pixel Front-End electronics in a commercial 65 nm CMOS technology suitable for the long-term CMS experiment pixel detector upgrade at HL-LHC.

Keywords: Particle tracking, silicon pixel detectors, LHC, luminosity, pileup, radiation damage, CMS experiment, VLSI, ASIC, readout chip, trigger, CMOS technology, upgrade, RD53, CHIPIX65

1.1 Introduction

Precise and efficient tracking and vertexing procedures are of utmost importance in high energy physics experiments. The reconstruction of the tracks provides measurements of charged particles momenta exploiting the curvature of trajectories in a magnetic field. Hits measured closest to the interaction region are fundamental in determining the position of the primary vertex (PV) and of secondary vertices (SV) originated by the decays of short-living particles. Pattern recognition, reconstruction of vertices and precise measurements of their impact parameters in a high track multiplicity environment necessary require sensors featuring high granularity, spatial resolution, speed and radiation hardness.
The last generation of particle physics experiments has seen substantial progresses in the usage of segmented silicon detectors for particle tracking. [Weilhammer 2000, Spieler 2005, Hartmann 2009]. Silicon sensors offer high spatial resolutions (5-10 μm) and maximum granularity, therefore can cope with huge track densities. Sensors in which the charge collection is driven by drift exhibit fast response and radiation hardness. In addition, good energy resolutions can be obtained with a low amount of material. The linear energy transfer for a Minimum Ionizing Particle (MIP) traversing a bulk of silicon is $\approx$ 390 eV/μm. This leads to about 32'000 electron/hole pairs in 300 μm of thickness thanks to the low ionising energy (3.6 eV) of silicon. Hence internal charge multiplication is usually not required. At the same time, a particle hit can be measured with no appreciable effect for the particle itself, because the energy loss is small ($\approx$ 0.1 MeV in 300 μm). Furthermore, multiple Coulomb scattering effects are minimized. However, maintaining good signal-to-noise ratios in hostile radiation environments requires low operating temperatures in order to keep leakage currents at low levels. Finally, silicon exhibits excellent mechanical properties and affordable production cost to instrument large areas.

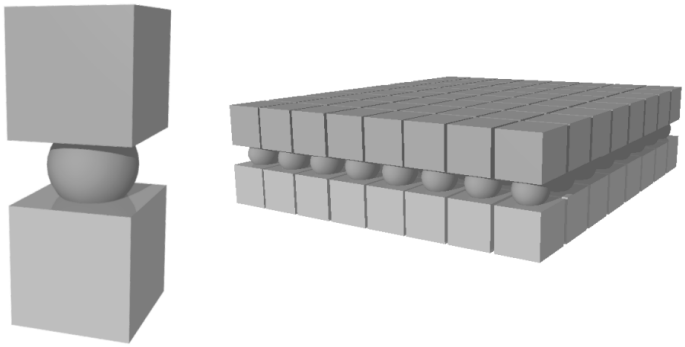


Figure 1.1: Schematic view of hybrid pixel sensors technology. Sensor and readout electronics are fabricated on different silicon wafers and then connected through the bump-bonding technique.

Among the reach variety of silicon sensor technologies developed for radiation detection, hybrid pixel detectors are a well-established and mature detector technology. They have been proved to be an efficient and reliable solution for particle tracking.
As reviewed in Figure 1.1, hybrid pixel detectors consists of an array of sensitive cells connected to a dedicated readout chip through the bump bonding technique [Rossi 2006]. Sensor and readout electronics are therefore fabricated separately on different silicon wafers and than mated together using a flip-chip processing. Tiny balls of conducting solders, typically In or Pb-Sn, are deposited on special bonding pads. Despite flip-chip technology is extensively adopted in semiconductor industry for the packaging of integrated circuits with ball grid array (BGA) surface mounting, the small pixel sizes demanded for particle physics and imaging applications require the usage of advanced cutting-edge processing.
Each pixel cell is a reverse-biased junction formed by a high resistivity substrate and a collection electrode connected to its own Front-End electronics. A ionizing particle that crosses the sensor generates electron/holes pairs that move in the depletion region under the action of the electric field. Hence the charge collection in the sensor is driven by drift, providing fast response and radiation hardness. The current signal induced at the electrode is processed by an optimized analogue Front-End chain that provides necessary amplification and filtering. Additional signal processing in the analogue and digital domains is performed according to the target application.
The usage of a custom designed mixed-signal Application Specific Integrated Circuit (ASIC) is the only means to achieve the integration density required to read out the large number of pixel cells and implement all functionalities demanded for the application. Furthermore, commercially available deep-submicron CMOS technologies have been also demonstrated to be radiation tolerant, as discussed later in the chapter.
Innermost layers of pixel detectors arranged in a barrel geometry around the interaction region and equipped with highly specialized readout circuits represent therefore a standard configuration for tracking detectors of a collider experiment. This is the case of pixel tracking systems installed at the CERN Large Hadron Collider (LHC) near Geneva. Operating conditions for a pixel tracking system at the LHC are challenging. In order to better understand and justify upgrade scenarios as well as pixel ASIC specifications and performance requirements discussed through the chapter, the next section introduce some fundamental concepts and common terminology related to the LHC environment. A description of the Compact Muon Solenoid (CMS) experiment is given in Section 1.3. According to the aim of this work, most of the attention will be then focused on pixel readout electronics.

1.2 General concepts and definitions

The LHC machine [Evans 2008] was conceived to deliver both proton-proton and nucleon-nucleon (lead ions) collisions at nominal unprecedented operating conditions. The choice of a hadron collider is well suited to the task of exploring new energy domains, resulting de facto into a discovery machine. For a symmetrical collider with two head-to-head[1] interacting beams of energy E_b each one, the available centre-of-mass energy in the lab frame is $\sqrt{s} = 2E_b$. Indeed, the fundamental constituents participating in the scattering are partons, which carry a fraction x of the 4-momentum of the particles in the beam. Hence the centre-of-mass energy of the effective hard scattering process is $\sqrt{\hat{s}} = \sqrt{x_1 x_2 s}$ and can span several orders of magnitude. According to proton parton distribution functions, on average $x \approx 0.15 - 0.20$. The nominal centre-of-mass energy of the LHC has been therefore chosen $\sqrt{s} = 14$ TeV in order to explore particle physics at the 1-2 TeV energy scale.
For a given physics channel with production cross section σ_{ev} the number of events per second (event rate) generated in a collider system is usually written as

$$\frac{dN_{\text{ev}}}{dt} = \sigma_{\text{ev}} L(t)$$

where $L(t)$ is a fundamental machine parameter referred to as *instantaneous luminosity*. It depends on collider parameters such as number of particles per bunch, number of bunches per beam, collision frequency (bunch crossing rate) and beams intersection area, but not on the physics process. The luminosity represents therefore the number of collisions per unit time and cross-sectional area of the beams. The production cross-section for a certain process of interests depends instead on the energy scale, spanning several order of magnitude depending on the physics under consideration. Furthermore, for each unstable particle produced in the interaction, a wide range of decay channels open, each one characterized by a certain branching ratio (BR).
Certainly the statistical significance of various data analyses relies on the total number of events produced across several data taking runs,

$$N_{\text{ev}} = \sigma_{\text{ev}} L_{\text{int}}$$

being

$$L_{\text{int}} = \int_0^{T_{\text{run}}} dt\, L(t)$$

the *integrated luminosity*. Due to intrinsic inefficiencies of detectors (either because one or more detector subsystems are temporarily unavailable, or simply because the detector is out of data taking for some reason) the integrated luminosity recorded by the experiments is lower than the total luminosity delivered by the machine. Both delivered and recorded integrated luminosity values are therefore carefully tracked as a function of time during on-line activities.

[1] Actually, a non-zero crossing angle is required to identify the position of the interaction.

Parameter	Design value (pp interactions)
Circumference	27 km
beam energy	7 TeV/beam
centre-of-mass energy $\sqrt{s}$	14 TeV
Number of bunches per beam	2808
Number of particles per bunch	$\approx 10^{11}$
Bunch radius at interaction point	15 μm
Bunch length	53 mm
Bunch crossing rate	40 MHz
Luminosity	10^{34} cm^{-2} s^{-1}

Table 1.1: LHC nominal design parameters [Evans 2008].

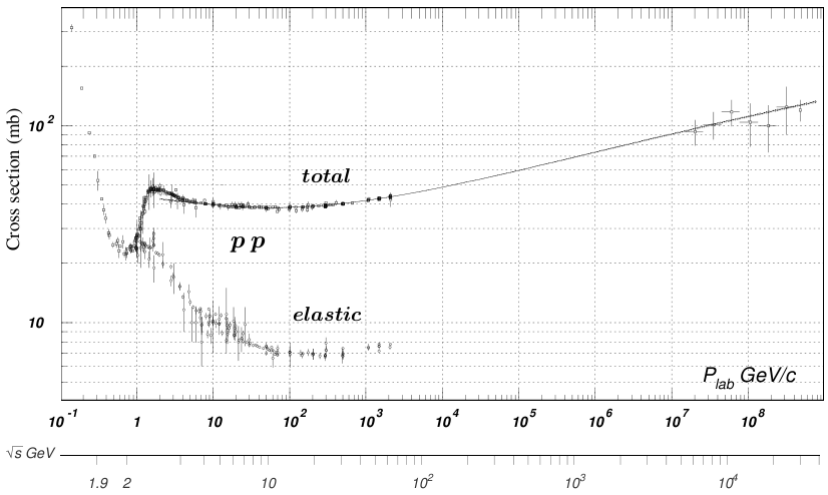


Figure 1.2: Total and elastic proton-proton cross section as a function of the centre-of-mass energy [Amsler 2008]. The foreseen total cross section at nominal LHC ($\sqrt{s} = 14$ TeV) is about 100 mb. Note that experimental data are available only from cosmic ray measurements.

High luminosities and high centre-of-mass energies in a hadron collider are therefore fundamental requirements in order to discovery or confirm physics processes with predicted small cross sections. At the LHC, protons are accelerated from 450 GeV up to 7 TeV. One beam circulates clockwise and the other one counterclockwise in separate but close orbits. They are forced to collide in specific regions around which the experiments are located. Protons are grouped into 2808 bunches, each one containing about 10^{11} particles. Protons constituting a bunch are confined in a cylindrical volume of 30 μm diameter and 5.3 cm length. According to nominal parameters, the accelerator has been designed to deliver proton-proton interactions every 25 ns at nominal beam energy of 7 TeV/beam and luminosity up to 10^{34} cm^{-2} s^{-1} [Evans 2008]. Indeed, due to the complexity and unprecedented operating conditions of such a giant machine, actual LHC performance from the first commissioning phase to physics data taking periods did not yet reach nominal values and continuously evolved in time in terms of collision frequency, luminosity and beam energy. Nevertheless, following considerations assume nominal design values. A more detailed description of effective LHC operations in 2009-2013 and its foreseen evolution through 2020 is remanded to Section 1.6. Most important nominal LHC design parameters are summarized in Table 1.1.

Figure 1.2 shows the total proton-proton cross section as a function of the centre-of-mass energy. At the design value $\sqrt{s} = 14$ TeV, the total proton-proton cross section is expected to be roughly 110 mb (1 b = 10 fm $\times$ 10 fm). Main contributions to the total cross-section are about 60 mb from non-diffractive processes, 10 mb from diffractive processes and 40 mb from elastic scattering. The total inelastic cross section is therefore about 70 mb. This implies that at the nominal LHC luminosity 10^{34} cm^{-2} s^{-1}, general-purpose detectors ATLAS and CMS have to cope with a total interaction rate of the order of 10^9 inelastic events/s. Indeed, interesting physics processes have cross sections orders of magnitude smaller than the total inelastic one. As an example, typical cross sections that involve the production of a Higgs boson at the LHC are of the order of a few tens pb. Most of particles created in each beams interaction do not contain any interesting new physics signatures, but only contribute to a huge background of particles with low transverse momentum p_T. Commonly referred to as *minimum bias*, this leads to an overwhelming background compared to the expected interesting physics channels.
The event rate is further increased by the fact that multiple proton-proton hard-scatter interactions can occur in the same bunch crossing, resulting into multiple primary vertices (PV) reconstructed by tracking algorithms. Commonly referred to as *pileup* (PU), this increases the overall number of particles emerging from the interaction region at each beams intersection. As a matter of fact, major challenges for ATLAS and CMS experiments arise from particle tracking under high PU conditions. As an example, Figure 1.3 shows an event display recorded by the CMS experiment during the 2012 data taking period with more than 20 reconstructed primary vertices in the same bunch crossing. The effective number of interactions for each beams intersection randomly varies and depends on instantaneous bunch-by-bunch luminosities. Figure 1.4 presents the typical distribution for the mean number of primary vertices per bunch crossing reconstructed in CMS. As a first approximation the PU distribution roughly fit a Poisson distribution. Detailed procedures have been deployed to compute PU distributions from delivered and recorded instantaneous luminosity measurements[2]. The average PU is proportional to the total proton-proton cross section, to the luminosity and inversely proportional to the bunch crossing rate. Hence the average PU increases by increasing the instantaneous luminosity. At LHC design luminosity 10^{34} cm^{-2} s^{-1} a mean of about 20 inelastic interactions superimpose in the same bunch crossing. As discussed later in the chapter, the foreseen luminosity upgrade by one order of magnitude up to 10^{35} cm^{-2} s^{-1} at HL-LHC will therefore increase the average PU up to 200, introducing unprecedented experimental challenges.

[2]The bunch-by-bunch recorded luminosity is usually determined by means of very-forward hadron calorimeters.

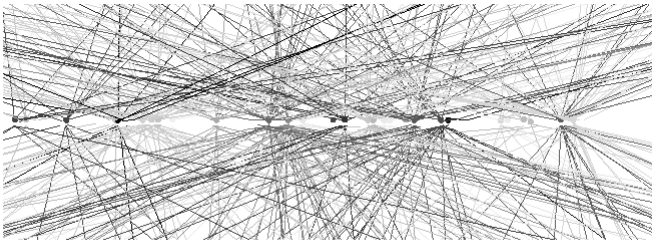


Figure 1.3: Event display recorded by the CMS experiment during the 2012 data taking period at $\sqrt{s} = 8$ TeV with more than 20 reconstructed primary vertices in the same bunch crossing.

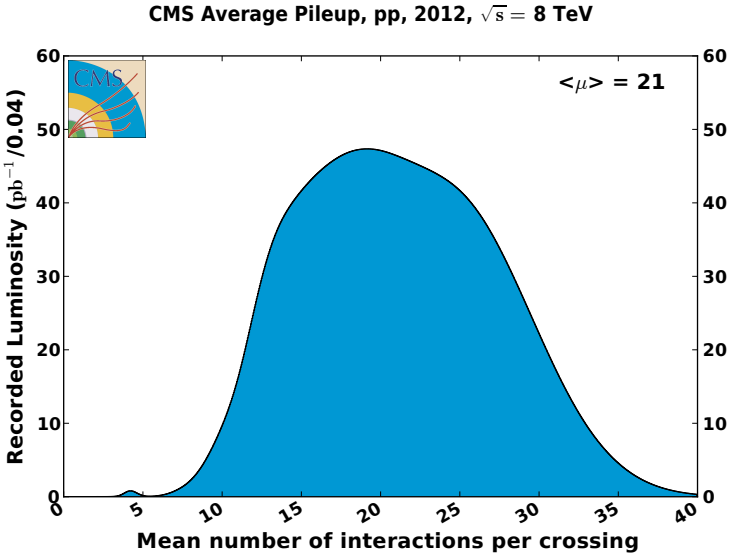


Figure 1.4: Distribution for the mean number of reconstructed primary vertices per bunch crossing (average pileup) in CMS during the 2012 data taking period. CMS public results.

Charged particles are detected by silicon tracking systems built around the interaction point. As discussed, at nominal LHC design luminosity, about 10^3 charged particles from more than 20 overlapping proton-proton interactions emerge on average from the interaction region every 25 ns. Silicon pixel detectors are placed closest to the beam line. Assuming an innermost barrel pixel detector located at 5 cm radius, this leads to a nominal maximum charged particle flux[3] of ≈ 50 MHz/cm^2 [Rossi 2006]. A detector technology featuring high granularity, fast response, low occupancy[4] and radiation hardness is therefore essential, justifying the extensive usage of hybrid silicon pixel detectors in the innermost regions of LHC experiments. Certainly by increasing sensor granularity to reduce the occupancy increases the number of readout electronics channels, requiring the implementation of highly specialized, low-power, radiation-hard readout electronics.
A few remarkable considerations are related to *trigger* aspects. As previously discussed, nominal LHC operating conditions lead to a total event rate in ATLAS and CMS experiments which is of the order of 1 GHz. Since it is impossible to store and process the large amount of data associated to all proton-proton collisions delivered by LHC, a drastic reduction of the event rate has to be performed. In practice, only about 100 events/s can be stored for subsequent data analysis, corresponding to 250 MB/s. Furthermore, most of the events are minimum bias and do not contain any interesting physics signature. A selection criteria must be therefore envisaged to filter out non interesting events, reducing the event rate to about 100 kHz that can be written to permanent storage. The goal of the trigger is to perform such a required huge data reduction by a factor 10^7. As described shortly thereafter, trigger requirements and output data rate introduce major implications in the design of suitable pixel readout electronics.
Trigger systems implemented at LHC experiments are complex, multi-step selection criteria. The time available to accept or reject an event is limited to the nominal bunch crossing time of 25 ns. According to a common terminology adopted for both ATLAS and CMS experiments, the event rate is reduced in two steps called Level-1 (L1) trigger and High-Level Trigger (HLT). The L1 trigger is mainly hardware-based. In order to increase flexibility, L1 hardware components are implemented in commercial FPGA technology wherever possible, but also custom-designed ASIC solutions are used where speed, density and radiation tolerance requirements must be satisfied. The L1 trigger information is basically extracted from coarsely segmented data provided by calorimeters and external muon detectors. The HLT is instead a software-based trigger system executed in dedicated computing farms equipped with thousands commercial CPUs. It processes all the events accepted by the L1 trigger, accessing the complete read-out detector raw data.
The L1 trigger inspects each bunch crossing. Nevertheless, the trigger decision, commonly referred to as L1-Accept (L1A), is communicated to all sub-detectors only after a certain amount of time or the order of μs. The L1 *trigger latency* is defined as the time between a given bunch crossing and the arrival time of the L1A to sub-detectors. Only events associated to an L1A must be readout (*zero-suppression*) thus requiring a dedicated *trigger matching* logic to accomplish this task. At present, the nominal trigger latency adopted in CMS is 3.2 μs, corresponding to 128 bunch intersections assuming a 40 MHz collision rate.
The necessity of a trigger introduces therefore fundamental requirements for the design of the readout electronics. Events information must be in facts continuously registered while waiting for the L1 trigger decision, requiring temporary data storage (*buffering*) features. Each event must be therefore properly associated to the corresponding bunch crossing (*time-stamping*) starting from a 40 MHz machine clock distributed to all detector sub-systems.
In case of pixel detectors, data buffering and retrieval capabilities must be embedded into the digital readout architecture of the Front-End ASIC, as there is no time to move the large amount of hit raw data anywhere else. As discussed later, this introduces additional challenges in perspective of the design of a new pixel ASIC suitable for HL-LHC.

[3] Also referred to as *track rate*.

[4] For a given detector subsystem, the occupancy is defined as the number of busy channels at a time with respect to the total number of available readout channels. It is usually expressed as a percentage.

1.3 The CMS experiment at the LHC

The Compact Muon Solenoid (CMS) detector [Chatrchyan 2008] operating at the CERN LHC has been conceived as a multi-purpose experiment to study proton-proton and lead-lead collisions at the TeV energy scale. In the following, a brief summary of the the CMS detector design is presented. A perspective view of the CMS apparatus is shown in Figure 1.5, completed by a transverse view in Figure 1.6. The overall layout follows the typical arrangement of a hermetic detector at a symmetric collider experiment, covering as much as possible the volume around the interaction region. Main components are therefore an innermost tracking system surrounded by electromagnetic and hadronic calorimeters and external muon detectors, arranged in a cylindrical geometry completed by endcaps structures.

The fundamental aspect in the detector design is the choice of a solenoidal[5] configuration of the magnetic field for tracking measurements. At the core of the CMS detector sits in fact a 13 m long, 6 m diameter, 4 T superconducting solenoid[6], which provides a large 12 Tm bending power which ensures good momentum resolution for charged particles with transverse momenta up to 1 TeV. The flux of the magnetic field is returned by a by a 10'000 tonnes iron yoke, composed of 5 barrel wheels and 6 endcap disks. The yoke also provides mechanical support for the entire system. The choice of a solenoidal configuration led therefore to a very compact design, allowing calorimeters and the inner tracking system to be installed inside the free bore of the magnet coil, resulting into a strong improvement in the detection and energy measurement of electrons and photons.
The overall tracking volume occupies a cylinder of 5.8 m length and 2.6 m diameter inside the 4 T homogeneous magnetic field provided by the solenoid. The CMS physics program requires excellent track reconstruction and vertexing performance. Efficient and precise reconstruction of charged particles tracks with transverse momenta above 1 GeV in the pseudorapidity range $|\eta| < 2.5$ are of primary importance for CMS. Moreover, precise reconstruction of secondary vertices and measurements of impact parameters are fundamental in the identification of short-living particles which are produced in many of interesting physics channels. Together with electromagnetic calorimeter and and the muon system, tracker information is used to identify electrons and muons respectively. Furthermore, tracking information is heavily used in the high level trigger of CMS. Hence a detector technology featuring high granularity, fast response and radiation hardness is essential in order to cope with the intense particle flux emerging from the interaction region at each bunch crossing. The main challenge in the design of the tracking system was to develop detector components able to operate in the LHC harsh environment for a long lifetime. High granularity, speed and radiation tolerance requirements led to a tracker design entirely based on silicon detector technologies.
A schematic cross section of the CMS inner tracking system is shown in Figure 1.7. The tracker is composed of an innermost Silicon Pixel Tracker (SPT) and an outer Silicon Strip Tracker (SST). The SPT is a system of 3 barrel layers equipped with hybrid pixel detectors at radii 4.4 cm, 7.3 cm and 10.2 cm, closest to the interaction region. The volume between 20 cm of radius and 116 cm is occupied instead by the SST, which is composed of 10 barrel layers of silicon microstrip detectors. Each barrel layer is completed by endcaps disks on each side, consisting of 2 disks in the pixel tracker and 3 plus 9 disks in the strip tracker. With a total active silicon area of about 200 m^2, the CMS tracker is the largest full-silicon tracking system ever built, with 1440 pixel-modules and 15'148 strip modules, corresponding to 66 million pixels and 9.3 million silicon strips.
A more exhaustive description of the current layout of the CMS pixel detector and its Front-End electronics is remanded to Section 1.4 and 1.5 respectively.

[5] In contrast, the ATLAS (A Toroidal LHC ApparatuS) detector uses a toroidal configuration, as suggested by the name of the experiment.

[6] Despite 4 T is the nominal value, a magnetic field of 3.8 T has been adopted during the first 2009-2013 data taking period.

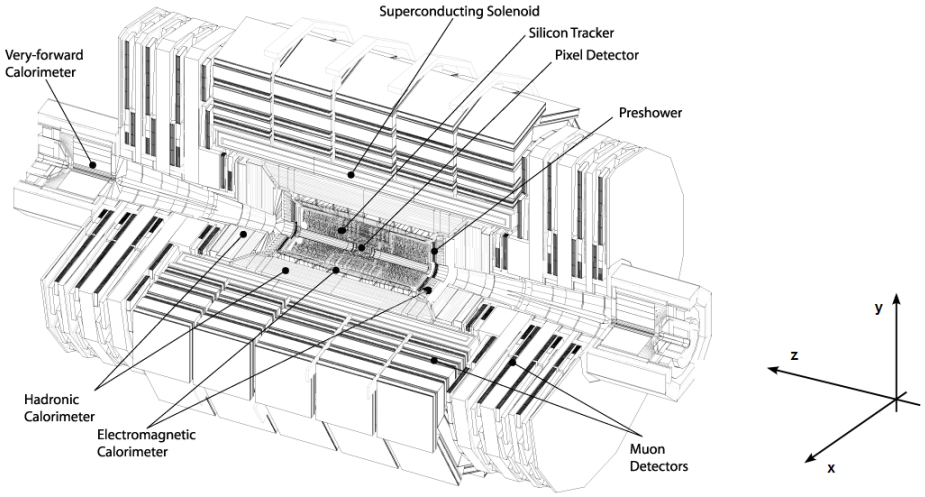


Figure 1.5: Perspective view of the CMS detector. The global coordinate system adopted by the collaboration has the origin centered at the nominal collision point inside the experiment. The $z-$axis points along the beam line, the $y-$axis points vertically upwards and the $x-$axis points radially inward to the centre of the LHC ring. The azimuthal angle ϕ is measured from the $x-$axis in the $x-y$ plane and the radial coordinate in this plane is denoted as r. The polar angle θ is measured from the $z-$axis. Pseudorapidity is defined as $\eta = -\ln[\tan\theta/2]$ and is extensively used to quote angular distributions.

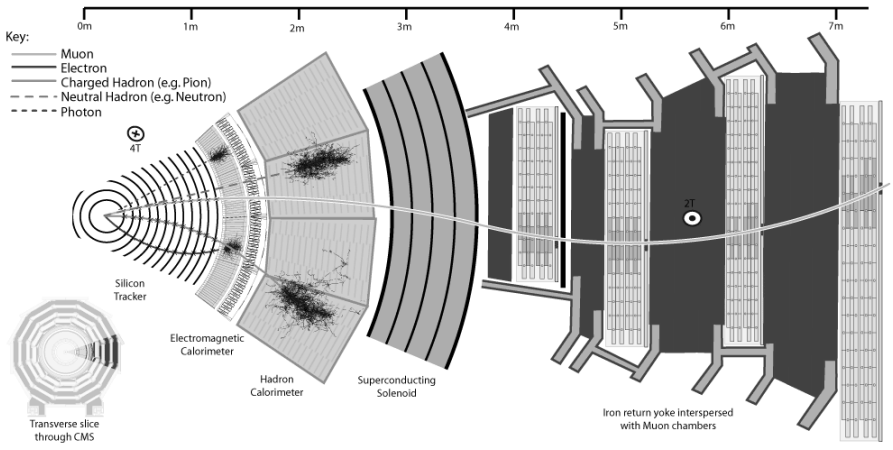


Figure 1.6: Transverse slice of the CMS detector and signatures for different particles crossing the volume. From inward to outward: all-silicon tracking system, ECAL, HCAL, superconducting solenoid and external muon chambers hosted by the iron yoke for the return of the magnetic field.

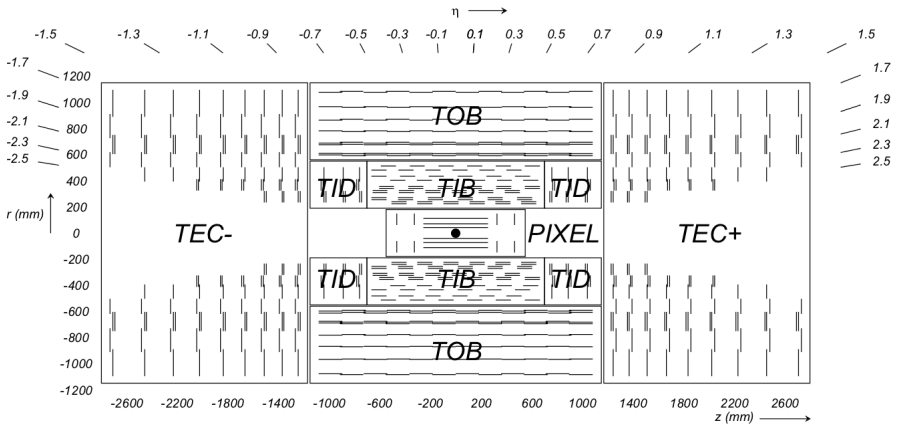


Figure 1.7: Schematic cross section of the CMS inner tracking system, 5.8 m length times 2.6 m diameter. Each line represents a detector module. The Silicon Pixel Tracker (SPT) is composed of 3 barrel layers and 2 endcap disks each side. The Silicon Strip Tracker (SST) is partitioned into two different sub-systems, referred to as Tracker Inner Barrel (TIB) and Tracker Outer Barrel (TOB). The TIB extends in radius towards 55 cm and $|z|$ <65 cm. It is composed of 4 barrel layers supplemented by 3 endcap disks at each end, referred to as Tracker Inner Disks (TID). The TIB system is surrounded by the TOB, consisting of 6 barrel layers completed by 9 lateral disks, referred to as Tracker End Caps (TEC).

The tracking volume is surrounded by an electromagnetic calorimeter (ECAL) to achieve accurate measurements of energy and position of electrons and photons, which are fundamental objects for the CMS physics program. The need for a good energy resolution and di-photon mass resolution ($\approx 1\%$ at 100 GeV) led to the choice of a homogeneous calorimeter composed of finely segmented lead-tungstate ($PbWO_4$) crystals with coverage up to $|\eta| < 3$, featuring fast response and necessary radiation tolerance. The scintillation light is detected by silicon avalanche photodiodes (APD) in the barrel region and vacuum phototriodes in the endcap regions. A preshower system is installed in front of the endcaps for $\pi^0 \to \gamma\gamma$ rejection.
The calorimetry system is then completed by a hadronic calorimeter (HCAL) for the measurement of the energy and direction of jets and reconstruction of missing transverse-energy contributions. A large geometric coverage is therefore of primary importance for this purpose. Coverage up to $|\eta| < 3$ is provided by a brass/scintillator sampling calorimeter surrounding ECAL and coupled to wavelenght-shift optical fibres which convert the scintillation light detected by hybrid photodiodes (HPD). Coverage up to $|\eta| < 5$ is obtained by an iron/quartz-fibre calorimeter which Cherenkov light is detected by standard photomultiplier tubes (PMT).
The iron return yoke of the magnet is large enough to host the external muon identification system, composed by 4 stations of aluminium Drift Tubes (DT) in the barrel region and Cathode Strip Chambers (CSC) in the endcap region, both complemented by Resistive Plate Chambers (RPC) which ensure redundancy and robustness for the muon trigger system.
The overall dimensions of the CMS detector are 21.6 m length and 14.6 m diameter, with a total weight of 12'500 tonnes and geometric coverage up to $|\eta| < 5$.

1.4 Current layout of the CMS silicon pixel tracker

The current layout of the CMS silicon pixel tracker is depicted in Figure 1.8. The pixel detector is placed closest to the interaction region, covering a pseudorapity range $|\eta| < 2.5$ and matching the acceptance of the outer silicon strip tracker. The pixel tracker consists of three barrel layers (BPIX) complemented by two endcap disks on each side (FPIX). Barrel layers are 53 cm long and are located at mean radii 4.4 cm, 7.3 cm and 10.2 cm. Endcap disks are placed at $\pm$ 35.5 cm and $\pm$ 45.5 cm from the nominal interaction point, extending in radius from about 6 cm to 15 cm. The total pixelated active area is about 1 m^2 and occupies a cylindrical volume of about 93 cm length and 30 cm diameter. This layout ensures 3 tracking points over almost the full $|\eta| < 2.5$ range. Figure 1.9 shows the geometrical acceptance as a function of pseudorapidity η. The barrel pixel detector covers the central region and the endcap disks the forward region. The transition barrel/forward occurs at $|\eta| \approx 1.5$.
The pixel tracker is instrumented with 285 μm thickness hybrid silicon pixel detectors. In order to achieve comparable spatial resolutions in both $r\phi$ and z directions, an almost-square pixel size of 100 μm $\times$ 150 μm has been adopted. Hit reconstruction algorithms heavily rely on the charge information measured for each pixel cell. The effect of charge-sharing among adjacent pixels induced by the large Lorentz drift in the 4 T magnetic field is in fact exploited to enhance the spatial resolution with a charge interpolation [Chatrchyan 2014]. The detector modules are therefore deliberately not tilted in the barrel layers. In the endcap disks instead, they are arranged in a turbine-like geometry to introduce charge-sharing. Nominal spatial resolutions are about 10 μm in the transverse coordinate and 20 μm in the longitudinal coordinate.
The operating conditions for the pixel tracker are challenging. As already discussed, at the LHC design luminosity of 10^{34} cm^{-2} s^{-1} on average 10^3 charged particles from more than 20 overlapping proton-proton interactions are created every bunch crossing. The position of the pixel detector close to the interaction region implies very high track densities and particle fluences that require both radiation-hard sensors and readout electronics. At nominal luminosity the innermost barrel layer placed at at radius $\approx$4 cm has to withstand a charged particle flux of the order of 100 MHz/cm^2, which in turn leads to pixel rate of about 10 kHz/pixel by adopting a 100 μm $\times$ 150 μm pixel size. This ensures a low occupancy of the order of 10^{-4} per bunch crossing, below 1%. The nominal charged particle flux decreases to 600 kHz/cm^2 at radius 22 cm and 30 kHz/cm^2 at 115 cm. Nominal radiation levels and particle fluences in different radial layers of the CMS tracker for an integrated luminosity of 500 fb^{-1} (corresponding to $\approx$10 years of nominal LHC operations) are summarized in Table 1.2. As one can see, assuming LHC design values the innermost barrel layer placed at radius 4.4 cm experiences a particle fluence of the order or 3×10^{14} n_{eq}/cm^2 per year.
Radiation hardness requirements led to the choice of pixel sensors implemented as heavily doped n^+ electrodes into a high resistance $n-$type substrate. Despite the higher costs due to the double sided processing, the $n-$on$-n$ concept was chosen as the collection of electrons ensures a high signal charge with moderate bias voltages after high hadron fluences, below 600 V. Extremely high operating voltages can be therefore avoided. This reduces the issues of leakage currents and high voltage breakdowns in a highly miniaturized environment. Most important CMS pixel sensor specifications are summarized in Table 1.3.
According to the hybrid pixel architecture, each pixel cell is bump-bonded to a full-custom ASIC. The overall readout electronics employed in the CMS tracking system has been designed in a commercial 0.25 μm CMOS technology by following special layout design rules which ensure the required radiation hardness [Snoeys 2000]. As a result, the lifetime of the silicon pixel tracker is limited by the radiation damage to silicon sensors. The innermost barrel layer has been designed to survive at least 2 years at nominal LHC luminosity, whereas the expected lifetime extends to more than 10 years for the third layer. Nevertheless, due to readout limitations in the current pixel Front-End ASIC, a replacement of the entire pixel tracker is already planned for 2016-2017, as discussed later in the chapter. In the following, the present Front-End ASIC coupled to CMS pixel sensors is described.

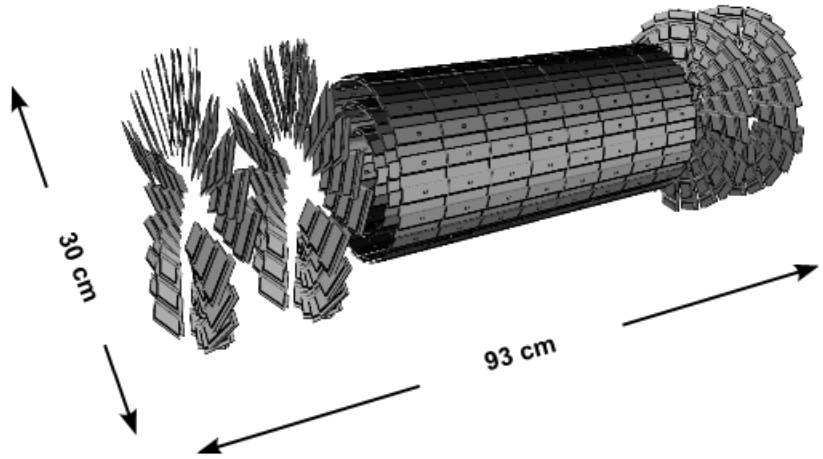


Figure 1.8: The current layout of the CMS Silicon Pixel Tracker (SPT) consists of three 53 cm long barrel layers (BPIX) at radii 4.4 cm, 7.3 cm and 10.2 cm complemented by two endcap disks on each side (FPIX) at $z = \pm$ 35.5 cm and $\pm$ 45.5 cm, extending in radius from about 6 cm to 15 cm. With a total active area $\sim$1 m^2 the CMS pixel tracker occupies a cylindrical volume of about 93 cm length and 30 cm diameter around the nominal interaction point.

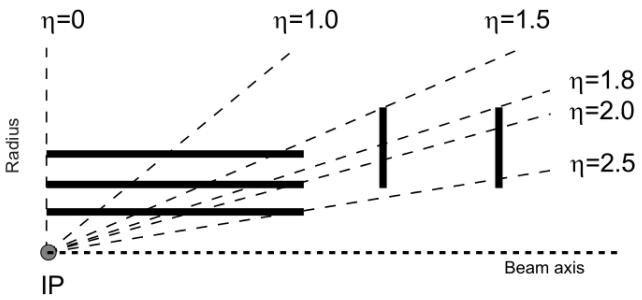


Figure 1.9: Geometrical coverage of the current CMS pixel detector in terms of pseudorapidity around the nominal interaction point [Chatrchyan 2008]. The three barrel layers cover the central region and the endcap disks the forward region. The transition barrel/forward occurs at $\eta \approx 1.5$. The current layout ensures 3 tracking points over almost the full $|\eta| < 2.5$ range.

radius [cm]	charged particle flux [cm^{-2} s^{-1}]	dose [kGy]	fluence of fast hadrons [cm^{-2}]
4	10^8	840	32×10^{14}
11		190	4.6×10^{14}
22	6×10^6	70	1.6×10^{14}
75		7	0.3×10^{14}
115	3×10^5	1.8	0.2×10^{14}

Table 1.2: Hadron fluence and radiation dose in different radial barrel layers of the CMS tracker assuming nominal LHC operating conditions and 500 fb^{-1} integrated luminosity, corresponding to $\approx$ 10 years operations [Chatrchyan 2008]. The fast hadron fluence is a good approximation to the 1 MeV neutron equivalent fluence. The innermost pixelated layer has to withstand a nominal charged particle flux of the order of 100 MHz/cm^2 and $\approx$ 1 MGy Total Ionizing Dose (TID) in 10 years.

Parameter	Specification
sensor technology	n^+-on$-n$
pixel size	100 μm ($r\phi$) $\times$ 150 μm (z)
sensor thickness	285 μm
Lorentz shift at 4 T	$\approx$ 59 μm (barrel region)
nominal reverse bias	-300 V
bump-bonding	15-20 μm diameter indium bumps
sensor capacitance	80-100 fF per pixel
nominal leakage current	10 nA per pixel

Table 1.3: Summary table for most important CMS pixel sensor specifications [Chatrchyan 2008]. The choice of an almost-square pixel size along with the charge-sharing among adjacent pixels lead to a nominal resolution of 10 μm in $r\phi$ and 20 μm in z.

1.5 State of the art hybrid pixel ASIC in CMS

CMS pixel sensors are made of 4160 cells arranged into 52 columns × 80 rows. Each sensitive cell covers an area of 100 μm × 150 μm. Signals generated by charged particles traversing the detector are processed and readout by a full-custom ASIC bump-bonded to the sensor. The initial concept of the chip dates back more than 15 years ago. Originally implemented in a dedicated 0.8 μm radiation tolerant process, the latest version of the ASIC, named PSI46V2, has been produced in a commercial 0.25 μm CMOS technology[7] officially supported by CERN [Barbero 2004, Erdmann 2005, Gabathuler 2005, Kastli 2006].
Despite the usage of a commercial solution, the integrated electronics has been made radiation-hard by design (RHBD) using special layout techniques, such as enclosed-layout transistors (ELT) and guard rings structures in order to ensure necessary radiation tolerance [Anelli 1999, Snoeys 2000]. According to a common terminology recently adopted within the LHC pixel ASIC community, the current CMS pixel readout chip (ROC) is a first generation pixel ASIC [Garcia-Sciveres 2013].

A complete view of a PSI46V2 chip layout is presented in Figure 1.10. Most important ASIC specifications are summarized in Table 1.4. The overall dimensions are 7.8 mm × 9.8 mm. The chip consists of a 7.8 mm × 8 mm core pixel matrix and a 7.8 mm × 1.8 mm chip periphery. The core active area contains 52 columns × 80 rows of 100 μm × 150 μm pixel cells bump-bonded to a sensor. As usually performed in most of pixel ASICs, two adjacent columns form a double-column in order to share power distribution, bias, data buses and common services.
As already discussed, a pixel ASIC operating at the LHC has to provide on-chip temporary data storage (buffering) and time-stamping capabilities in order to transfer off-detector only zero-suppressed (triggered) data. On the one hand, each pixel cell has to process and register the signals produced by particles in the sensor. On the other hand, the bunch crossing information, the pixel address and the amount of collected charge of all channels must be stored during the whole L1 trigger latency, sending out data only for those bunch crossings for which an L1A has been distributed to sub-detectors. Hit information generated in each single pixel unit cell (PUC) are therefore sent to the chip periphery where are temporary stored on buffers while waiting for a first level trigger decision. In particular, the PSI46V2 employs a full-analogue readout of the charge information of each pixel. As discussed shortly thereafter, charge information is retrieved in analogue form using a sample-and-hold (S/H) circuit placed in each pixel cell, transferred to analogue buffers in the chip periphery and buffered until the L1 trigger decision is taken. The actual off-chip data readout is performed in analogue form as well, using 40 MHz serial analogue links. A digitization of the charge information is performed only in counting rooms using VME modules equipped with 10-bit A/D converters. One remarkable aspect is the totally absence of synthesized logic in the pixel ASIC. All digital components have been implemented as full-custom solutions without the support of any automated synthesis and place-and-route engines.

A schematic block diagram of a pixel unit cell is presented in Figure 1.11, completed by a layout view in Figure 1.12. Each pixel cell is partitioned into an analogue part and a digital part.
The signal originated in a pixel sensor is transferred through the bump-bonding and enters a two-stage analogue Front-End system composed of a charge sensitive amplifier (CSA) and a shaper. Alternatively, test charge calibration signals can be injected through a selectable 4.8 fF injection capacitor directly connected to the CSA input node. The overall analogue Front-End chain has been optimized for a sensor capacitance of 80-100 fF.

[7] Almost all full-custom electronic systems currently installed at LHC experiments have been implemented using such a 0.25 μm CMOS technology radiation-hardened by design using special layout techniques.

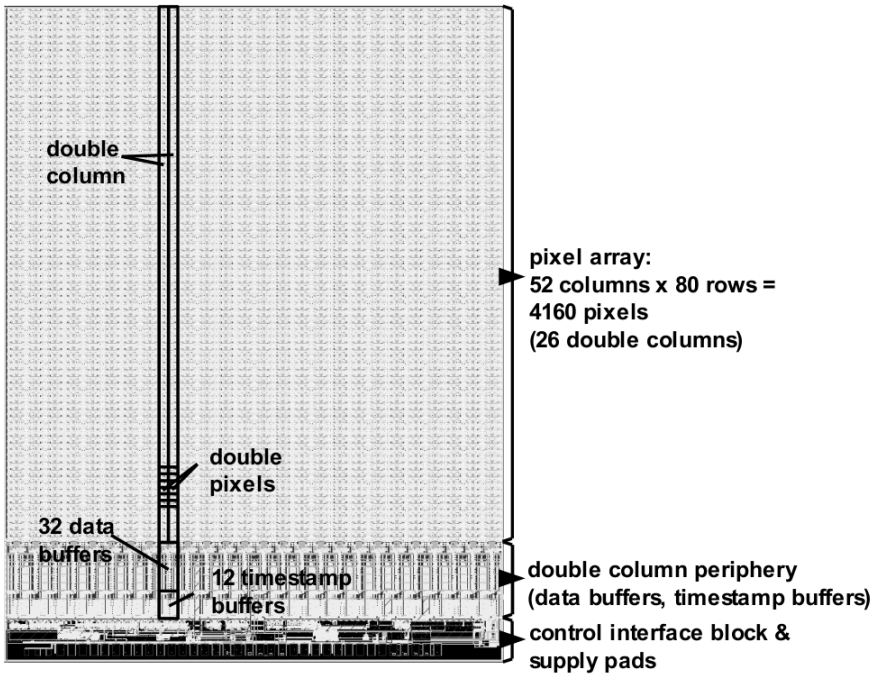


Figure 1.10: Full size layout view of a PSI46V2 readout chip (ROC) currently employed in the CMS silicon pixel tracker [Gabathuler 2005]. The active matrix of 52 × 80 pixels has a size of 8 mm × 7.8 mm. Pixel columns are grouped into 26 double-columns. The chip periphery has a length of 1.8 mm and hosts analogue buffers, time-stamp buffers, I/O interfaces, control D/A converters and voltage regulators.

Parameter	Specification/value
pixel size	100 μm × 150 μm
chip size	7.9 mm × 9.8 mm
fabrication technology	CMOS 0.25 μm (RHBD)
number of pixels	52 columns × 80 rows
readout	full-analogue
nominal charge threshold	3 ke^-
readout speed	40 MHz
supply voltages	1.5 V (A) and 2.5 V (D)
power consumption	≈ 30 μW/pixel

Table 1.4: PSI46V2 ASIC specifications summary table.

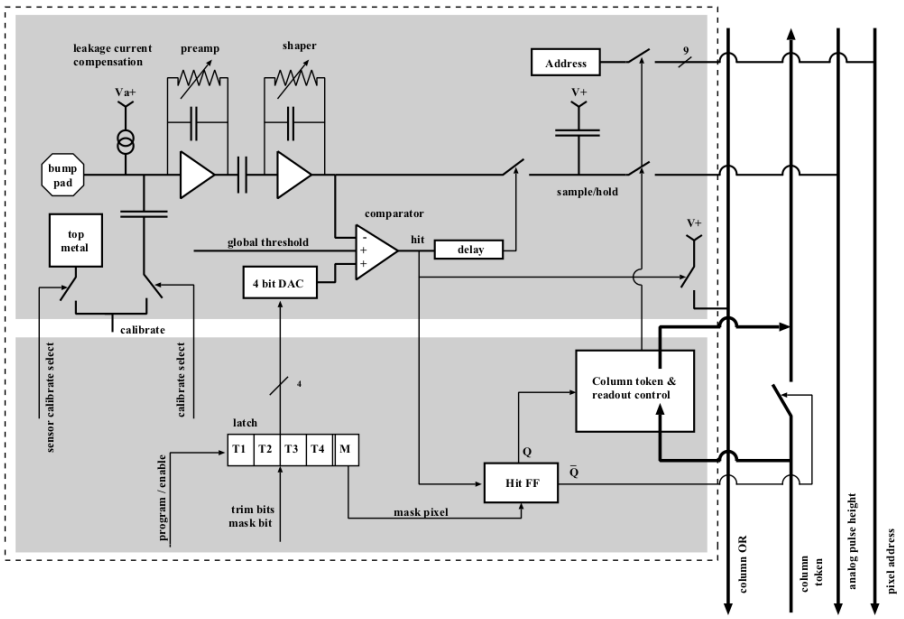


Figure 1.11: Schematic block diagram of a pixel unit cell (PUC) in the current CMS pixel readout chip PSI46V2 [Kastli 2006].

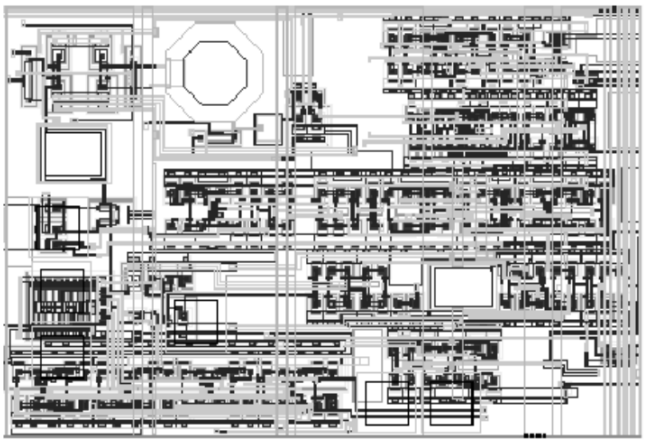


Figure 1.12: PSI46V2 pixel unit cell layout, 150 μm $\times$ 100 μm.

AC coupling is performed between CSA and shaper stages. It removes DC offsets caused by the sensor leakage current and it is part of the closed-loop gain of the shaping stage. The gain is given by the ratio between the coupling-capacitance and the feedback-capacitance. The nominal operating point of both amplifiers sits in the middle between the power and ground rails and it is independent of voltage drops across the chip because the analogue power rails have been kept symmetric and well separate from other power rails. This minimizes systematic variations of thresholds and feedback transistors. Passive feedback resistors are implemented with PMOS transistors working in linear region. Furthermore, both gate and substrate potential have been made adjustable in order to control the resistance of the feedback resistors. A programmable current source compensates the sensor leakage current, nominally 10 nA per pixel.
The shaper output is fed to a continuous-time voltage comparator. A global threshold is distributed to all pixels, generated at the chip periphery by a D/A converter. In order to compensate for channel-to-channel threshold variations, each pixel cell includes a local 4-bit D/A converter to trim the threshold. The overall threshold dispersion before correction is approximately 300 e^- RMS, reduced to 80 e^- after trimming. Furthermore, a mask bit allows to disable noisy pixels. Due to high energy particle irradiation, information stored in the chip can be corrupted by Single Event Upset (SEU) effects. In order to protect trim and mask storage cells, a capacitor is inserted in parallel to the classical cross-coupled inverters structure extensively adopted to achieve memory in CMOS digital circuits. At nominal LHC luminosity the SEU rate has been estimated to be less that 3×10^{-2} Hz. The occupancy of each pixel is monitored online and pixels that show significant changes with respect to calibrations are reprogrammed.
If the analogue pulse is above the threshold, the shaper output signal is sampled onto a capacitor and the pulse amplitude is stored for later readout. Charge information is therefore retrieved in full-analogue form. At the same time, a digital signal is sent to the chip periphery via an asynchronous column fast-OR bus to inform the chip periphery of the presence of a hit information. The pixel becomes insensitive and waits to be read out. No clock or bunch crossing time-stamp is distributed over the pixel array, resulting into a bare minimum digital switching activity in the core pixel array. The chip periphery synchronizes the wired-OR with the LHC machine clock and latches the current bunch crossing number in a time-stamp buffer whenever a hit is found. The state of the pixels that have a hit at that time is frozen and their data is subsequently collected. A token passing from pixel to pixel controls the data transfer. When a token arrives to the pixel cell, the pulse amplitude previously stored on the sampling capacitor is sent to the periphery together with the pixel address, where they are associated to the bunch crossing and stored in a second buffer. The token flag is then passed on and the pixel resumes data taking. All 26 double-columns are readout in parallel. Pixels in the same double column are read clockwise from the bottom left to the bottom right position. The time necessary to drain a double column depends on the number of hit pixels. Finally, a double-column logic at the chip periphery controls the data transfer, stores the hit information in analogue buffers for the whole trigger latency and performs the trigger verification. A schematic block diagram of the logic is presented in Figure 1.13. The trigger matching is performed as follows. The data buffer consists of 32-units, each made of a marker bit to indicate the beginning of a new event and to synchronize data and time-stamp, one analogue storage cell and 9 digital storage cells for the pulse height and the pixel address respectively. The oldest entry in the time-stamp buffer is continuously compared to the an 8-bit counter delayed with respect to the bunch crossing by a programmable amount of time corresponding to the trigger latency (Search Bunch Crossing counter, SBC). Time-stamp numbers and analogue data can be stored in the corresponding buffers for a maximum trigger latency of 3.2 μs, corresponding to 128 bunch crossing at 40 MHz bunch interaction rate. In case of agreement, the trigger is checked. If an L1A is present, the system enters in readout mode and the double column stops data acquisition to prevent overwriting of ready data. Otherwise, time-stamps and data buffers are cleared.

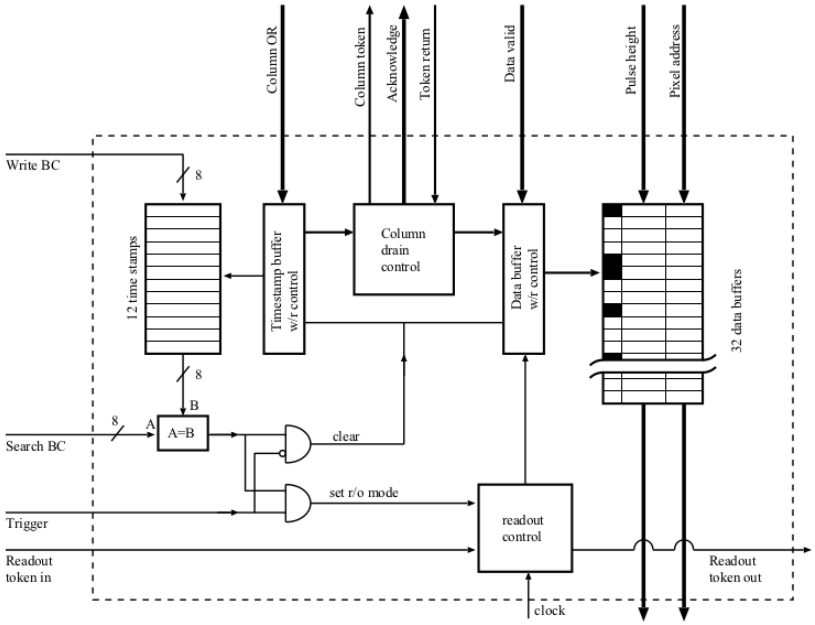


Figure 1.13: Schematic view of a double-column periphery logic [Kastly 2006].

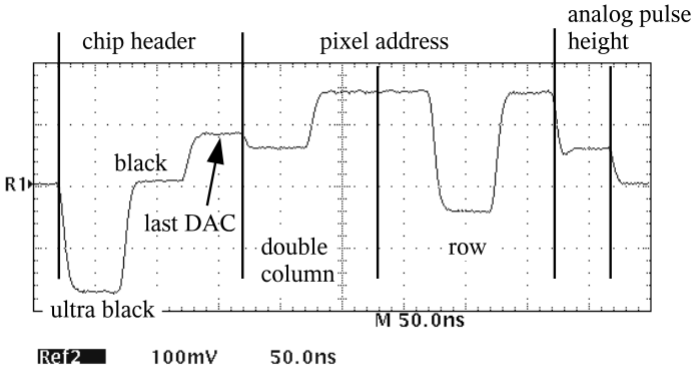


Figure 1.14: Full-analogue readout sequence of a ROC with one pixel hit [Kastly 2006]. Both the pixel address and the charge information are DC analogue levels.

Data losses can occur when one of the buffers is completely filled up. In case of a full data buffer the double-column gets reset. If the time-stamp buffer is full instead, data acquisition is paused until the next buffer cell is freed. Most of the chip periphery area is therefore occupied by double-columns buffers. As discussed later in Section 1.7, the number of buffers in the chip periphery currently represents the main limitation of the PSI46V2 chip in perspective of a first LHC luminosity increase foreseen before 2020. An improved version of the ASIC has been already implemented in order to sustain a luminosity twice the nominal value, up to 2×10^{34} cm^{-2} s^{-1}.
The chip periphery also hosts necessary slow controls and supply blocks such as such bandgap references and voltage regulators. The chip requires two external supply voltages, 1.5 V for the analogue sections and 2.5 V for the digital parts. The total power consumption is 120 mW per chip, corresponding to about 29 μW per pixel and a total power density of the order of 250 mW/cm^2. As discussed later in the chapter, this parameter will become a severe constraint for any future pixel ASIC design. Voltage regulators are programmable, hence the voltages can be set for each chip separately. The overall chip configuration uses a serial programming interface, based on a modified I^{2}C protocol running at 40 MHz. More than 30 D/A converters are used to configure and calibrate the chip, resulting into a very flexible system.
Zero-suppressed data are sent off-chip using a full-analogue serial readout running at 40 MHz (alternatively, it can be switched to 20 MHz) through optical links that connects a group of pixel chips to the remote electronics placed in counting rooms. Since the readout uses only analogue differential signals, the pixel address in digital form is encoded into 6 analogue discrete levels by means of D/A converters. Pixel address and signal pulse height can be transferred in 6 clock cycles. As already mentioned, digitization of pixel information is performed only in counting room.
A sample readout sequence for a single pixel hit is shown in Figure 1.14. The readout sequence starts with a header of three clock cycles. A large negative signal level well outside of the range of pixel data (ultra-black) followed by a zero differential level (black) indicates the begin of the data stream. In the third clock cycle of the chip header a DC level inversely proportional to the value of the to the most recently programmed DAC (last-DAC) is sent. Header information is followed by the double-column address (2 clock cycles), the row address (3 clock cycles) and the analogue pulse height for each hit pixel.

Barrel layers and the endcap disks of the CMS pixel tracker are composed of pixel detector modules. Each module contains a variable number of PSI46V2 readout chips bump-bonded to a common silicon sensors (8 or 16 in the barrel, 22 or 23 in the endcap disks). The overall CMS silicon pixel tracker consists of about 66 million pixels and 1400 detector modules.
The serial readout of triggered data for a group of PSI46V2 chips part of the same module is controlled by an auxiliary custom-designed ASIC implemented in the same radiation-hardened by design 0.25 μm CMOS technology, the Token Bit Manager (TBM) [Bartz 2005]. As depicted in Figure 1.15, a readout token is passed from chip to chip, connecting each of them to the analogue readout bus. The TBM is also responsible for the formatting of the serial data stream. Furthermore, it distributes L1 triggers and 40 MHz clocks to each ROC in the module. An arbitrary number of PSI46V2 chips can be chained and sequentially read out in a single token scan.
The basic structure of pixel module in the innermost barrel layes is shown in Figure 1.16. The overall dimensions are 26 mm ($r\phi$) $\times$ 66.6 mm (z). Two basestrips made of 250 μm thick silicon nitride provide the mechanical support for the entire module. The pixelated sensitive layer has a total active area of 64 mm $\times$ 16 mm, with silicon sensor of 280 μm thickness. As mentioned, 16 or 8 PSI46V2 readout chips are bump-bonded to the sensor. On the opposite side of the module there is a flexible and low mass 3-layers printed circuit board, the High Density Interconnect (HDI), which distributes signals and power to the electronics. At the centre of the HDI is placed the TBM chip. The nominal power dissipation is 2 W/module.

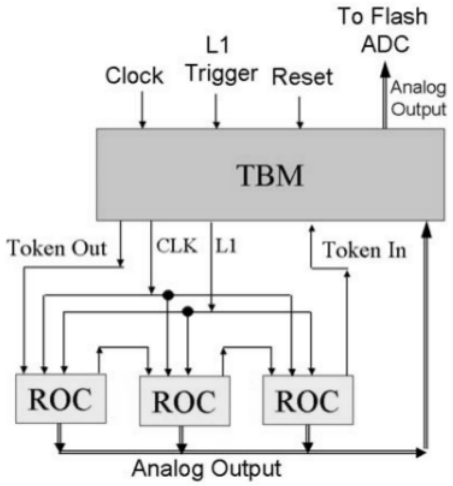


Figure 1.15: Schematic block diagram of a readout chain consisting of a TBM chip and a group of pixel ROCs [Bartz 2005].

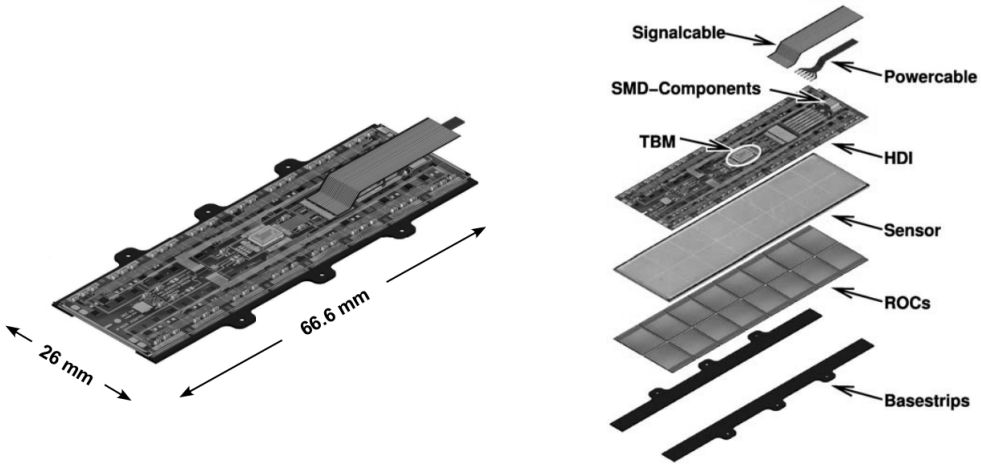


Figure 1.16: Perspective view of a CMS barrel pixel module. Each module contains 16 or 8 PSI46V2 readout chips bump-bonded to 280 μm thick silicon sensors with 100 μm $\times$ 150 μm pixel size. On the opposite side a flexible HDI hosts the TBM chip that controls the full-analogue serial readout of the ROCs. All signals are transferred using an impedance matched Kapton/copper cable.

1.6 LHC operations in 2009-2014 and upgrade scenarios

The LHC machine has been designed to reach nominal 14 TeV centre-of-mass energy and luminosity 10^{34} cm^{-2} s^{-1} at 40 MHz bunch crossing rate. Indeed, given the unprecedented complexity of such a giant machine, nominal values have not yet been reached. After the preliminary commissioning phase, LHC operations continuously evolved in time in terms of bunch crossing rate, beam energy and delivered instantaneous luminosity. Furthermore, several upgrade scenarios and technical plans have been investigated since the initial concept of the accelerator, following the expected evolution of machine performance and involving both short-term and long-term improvements and upgrades in the forthcoming 10-15 years.
The first LHC physics program, referred to as *Run* 1, includes collisions delivered between first operations in November 2009 up to a first 2-years long shutdown (LS1) started on February 2013. The original schedule of the accelerator was delayed by about one year due to an accident on September 2008 just a few days after the first proton beams were circulated in the main ring. Most likely the cause of the problem was a faulty electrical connection between two magnets, causing a loss of approximately 6 tonnes of liquid helium, which was vented into the LHC tunnel. Vacuum conditions in the beam pipe were lost and a total of 53 superconducting magnets were damaged. Most of 2009 was therefore spent to repair the LHC machine and reviewing the damage caused by the quench incident. Only on November 2009 first proton-proton collisions at limited 450 GeV centre-of-mass energy were delivered to experiments. The energy was then gradually increased up to the unprecedented value of 1.18 TeV/beam, beating the previous record of 0.98 TeV/beam held by the Tevatron machine at Fermilab, US.
After necessary commissioning, at the end of March 2010 two proton beams in the LHC collided at the unprecedented centre-of-mass energy of 7 TeV, marking the begin of the effective LHC physics research program. The LHC has been operated through the rest of 2010 at the same beam energy of 3.5 TeV/beam and 150 ns bunch spacing time. The instantaneous luminosity was increased either by increasing the current intensity of the beam or increasing the number of bunches per beam, targeting to reach 10^{32} cm^{-2} s^{-1} before the end of 2010. In such a first low luminosity phase the LHC experiments started their physics programs by measuring large cross section processes, confirming the Standard Model with a large number of recorded events despite the low luminosity. The early LHC physics represented a fundamental benchmark for LHC experiments, validating overall detectors performance and allowing a fine tuning of reconstruction algorithms and necessary calibration constants with real data. The first proton run officially ended on November 2010, reaching a total integrated luminosity of about 47 pb^{-1}. The first run with lead ions started on 8 November 2010 and ended on 6 December 2010.
At the end of the heavy ions run, the LHC was shutdown for a 3-months technical maintenance and restarted on March 2011. Another milestone was reached on April 2011, when the LHC delivered an unprecedented peak luminosity of 4.67×10^{32} cm^{-2} s^{-1} beating once again the Tevatron record of 4×10^{32} cm^{-2} s^{-1}. With a further luminosity increase up to $\approx 3 \times 10^{33}$ cm^{-2} s^{-1} ATLAS and CMS experiments reached 5 fb^{-1} of collected data in October 2011, reporting first hints of a signal in the search for a SM Higgs boson. With the real possibility of a discovery before the foreseen first 2-years long shutdown in 2013-2014, it was decided to increase the beam energy up to 4 TeV/beam, slightly changing the schedule of the LHC. On April 2012, the first collisions were delivered at the new unprecedented centre-of-mass energy of 8 TeV. On July 4th 2012 ATLAS and CMS collaborations announced the discovery of new particle with mass ≈ 125 GeV/c^2 consistent with a Higgs boson. Without doubts, such a discovery represents the major achievement of the first LHC physics run. A further 11 fb^{-1} integrated luminosity has been delivered through summer 2012, with instantaneous peak luminosities approaching 7×10^{33} cm^{-2} s^{-1}.

Parameter	2010	2011	2012	design value
beam energy [TeV]	3.5	3.5	4	7
bunch spacing [ns]	150	75-50	50	25
peak luminosity [cm^{-2} s^{-1}]	2.1×10^{32}	3.7×10^{33}	7.7×10^{33}	1×10^{34}
number of bunches	368	1380	1380	2808
number of protons/bunch	1.2×10^{11}	1.45×10^{11}	1.7×10^{11}	1.15×10^{11}
maximum pileup	4	17	37	20

Table 1.5: Summary table for most important performance parameters during LHC operations in 2010-2013 and comparison with nominal design values. LHC machine performance public results.

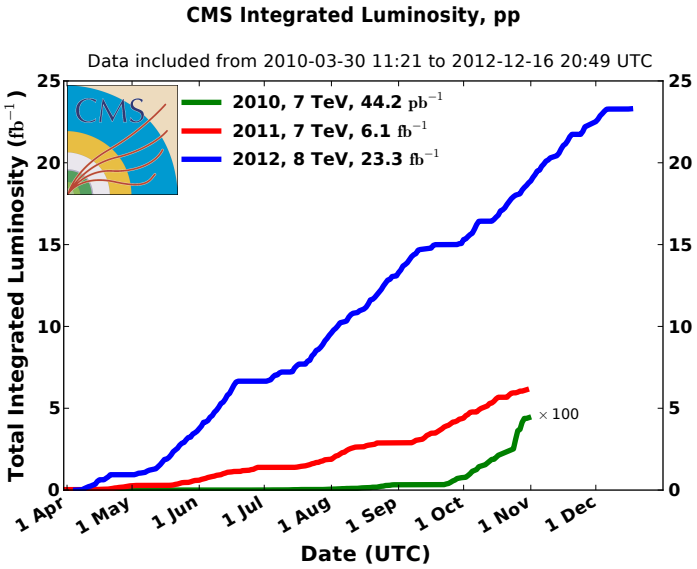


Figure 1.17: Integrated luminosity versus day delivered to the CMS experiment during stable beams and proton-proton collisions in 2010-2011-2012. CMS public results.

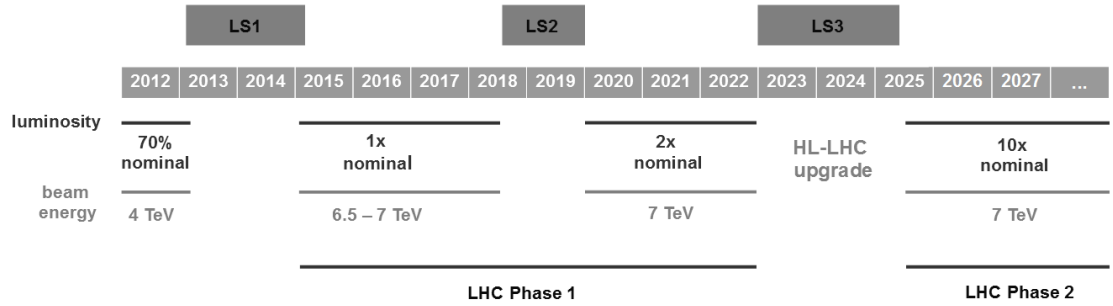


Figure 1.18: Foreseen evolution of LHC performance in terms of beam energy and luminosity through 2020 and beyond.

Electron cloud effects in the beam pipe have been identified as a major performance limitation for the LHC when operating at nominal 40 MHz bunch crossing rate, introducing complex issues related to beams instability [Rumolo 2011, Iadarola 2014]. Proton-proton collisions delivered to experiments for data taking during 2011 used bunch trains with either 75 ns or 50 ns bunch spacing, whereas a bunch spacing of 50 ns has been adopted through 2012. Only in dedicated special runs at the end of 2012 LHC operated at nominal 25 ns. Most important performance parameters for LHC operations during 2010-2013 and a comparison with nominal design values are summarized in Table 1.5. The total integrated luminosity versus day delivered to the CMS experiment during stable beams and proton-proton collisions in 2010-2012 is presented in Figure 1.17. By the end of the 2012 both ATLAS and CMS collaborations collected more than 20 fb^{-1}.
All machine and data taking activities have terminated on February 2013, successfully concluding the so called *Phase*-0 of the LHC schedule. After more than 3 years of operations, a massive program of technical maintenance for LHC magnets and its injector chain has been accomplished in the first 2-years long shutdown (LS1) in 2013-2014, targeting after restart to reach nominal luminosity 10^{34} cm^{-2} s^{-1} and 14 TeV centre-of-mass energy in proton-proton collisions without the need of major hardware modifications. In the same period, LHC experiments gained from machine inactivity for necessary consolidation, maintenance, technical repairs and improvements. With the end of LS1, the LHC is entering in its *Phase*-1 schedule. At the time of writing, the machine has just restarted, circulating first particle beams at low energy. After commissioning, particle collisions at new unprecedented beam energy of 6.5 TeV/beam are expected on June 2015, marking the begin of *Run* 2. Despite the plan is to operate with a nominal 25 ns bunch spacing, further operations at 50 ns cannot be ruled out at this time.
The LHC will remain the most powerful accelerator ever built in the world for at least two decades. In order to fully exploit the discovery potential of this machine, the foreseen LHC research program will extend over more than 20 years, investigating and validating the mathematical consistency of the Standard Model at the TeV energy scale and searching for evidences of new physics signatures. According with the expected increase of the machine performance in the forthcoming years, current planning for the LHC and its injector chain foresee a series of two additional long shutdowns, which will require major hardware modifications for both the machine and its experiments. A long-term plan for the evolution of the machine has been discussed during a Review of LHC Injector Upgrade Projects (RLIUP) workshop held at the end of October 2013, identifying the strategy that should lead to achieve an unprecedented integrated luminosity of 3000 fb^{-1} in proton-proton collisions. Based on this input, the long-term schedule depicted on Figure 1.18 was established on December 2013.

The LHC *Phase*-1 nominally extends up to 2022. Through 2015-2017 the machine will reach nominal centre-of-mass energy and luminosity. After attaining design operating conditions, the accelerator is expected to deliver a total integrated luminosity of about 40 fb^{-1}/year.
A second long shutdown (LS2) is foreseen to start on July 2018, lasting for at least 18 months. With first major hardware upgrades to LHC injectors, after LS2 the luminosity delivered to ATLAS and CMS experiments will be gradually increased beyond the nominal value, up to 2×10^{34} cm^{-2} s^{-1} or higher before 2020. The average pileup per bunch crossing will more than double, up to 40-50, thus increasing track densities and radiation levels in ATLAS and CMS beyond nominal values. As a matter of fact, this is an operating scenario for which these general purpose detectors have not been designed. Under these conditions, the ability of experiments to benefit from the higher luminosity delivered by the LHC will be seriously compromised. Several upgrade projects that will introduce first important hardware modifications in ATLAS and CMS detectors already started more than 5 years ago. Following the accepted terminology, LHC detectors will install most of their own *Phase*-1 *upgrades* during LS2. Indeed, as discussed in the next section, the CMS silicon pixel tracker will be already replaced in an extended 19-weeks winter technical stop between 2016-2017 in order to cope with twice the nominal luminosity, which otherwise would degrade overall tracking performance at unacceptable levels.

In order to further extend the machine discovery potential, a third extended long shutdown (LS3) is expected to start in 2022-2023, lasting more than 2 years. With a challenging upgrade of the LHC itself, after this period the accelerator will increase the instantaneous luminosity up to 10^{35} cm^{-2} s^{-1}, one order of magnitude higher with respect to the nominal value.
The new machine configuration, referred to as High Luminosity LHC (HL-LHC), has the goal to deliver an astonishing total integrated luminosity of 3000 fb^{-1} in 10 years after LS3. With such integrated luminosity, ATLAS ans CMS experiments will have a concrete chance to observe extremely rare processes predicted by the SM, as $H \rightarrow \mu^+\mu^-$ and $H \rightarrow Z\gamma$, which by having fixed the mass of a potential Higgs boson at ≈ 125 GeV/c^2 can be detected and precisely measured with 3000 fb^{-1}. With the commissioning of HL-LHC, the machine will enter in its *Phase*-2 schedule. The overall concept, design and installation of HL-LHC requires about 10 years to be accomplished. The upgrade will rely on a number of key innovative technologies and technological challenges, such as cutting-edge 13 T superconducting magnets, very compact and ultra-precise superconducting crab cavities, new technology for beam collimation and 300 m long high-power superconducting links with zero energy dissipation [Rossi 2012].
During LS3, ATLAS and CMS experiments[8] will perform their *Phase*-2 *upgrades*, requiring major hardware and software transformations in order to withstand the new unprecedented luminosity. With a luminosity of 10^{35} cm^{-2} s^{-1} and nominal 25 ns bunch spacing the pileup will increase to impressive values, resulting into extreme track densities and radiation levels in the innermost regions of tracking systems, one order of magnitude larger than original design values. Ongoing simulation studies suggest average values of 140 proton-proton interactions per bunch crossing or higher, up to 200. This can be appreciated in Figure 1.19, which shows a comparison between two simulated bunch crossing events under *Phase*-1 and *Phase*-2 pileup scenarios.
New HL-LHC operating conditions will therefore impose significant upgrades for the ATLAS and CMS inner tracking systems, demanding the installation of completely new silicon pixel detectors able to cope with unprecedented event rates and radiation levels. A replacement of outer trackers will be required as well and further upgrades will involve all other sub-detectors, trigger systems, DAQ systems and on-line softwares.

A brief overview of the already scheduled evolution of the current CMS silicon pixel detector in the forthcoming years is presented in the next section. According to the aim of this work, new ASIC requirements for pixel upgrades at HL-LHC are extensively discussed in Section 1.8 instead.

[8] In contrast, ALICE and LHCb will implement their major detector upgrades already during LS2.

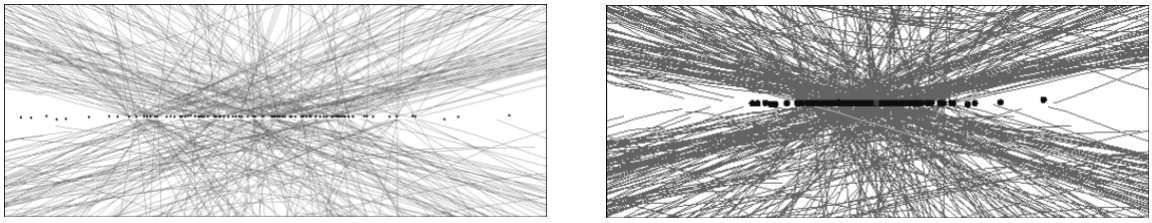


Figure 1.19: Comparison between two simulated bunch crossing events under LHC *Phase*-1 (left) and *Phase*-2 (right) pileup scenarios. Track densities and radiation levels foreseen at HL-LHC introduces unprecedented challenges in the design of new silicon pixel detectors that will be placed closest to the interaction region.

Parameter	LHC Phase0	LHC Phase1	LHC Phase2
luminosity	10^{34} cm^{-2}s^{-1}	2×10^{34} cm^{-2}s^{-1}	10^{35} cm^{-2}s^{-1}
bunch crossing	75-50 ns	50-25 ns	25 ns mandatory
average pileup	20	50	140 or higher
particle flux	50 MHz/cm^2	200 MHz/cm^2	500 MHz/cm^2
data rate	200 MHz/cm^2	600 MHz/cm^2	1-2 GHz/cm^2
TID (10 years)	1.5 MGy	3.5 MGy	10 MGy
signal threshold	2.5-3 ke$^-$	1.5-2 ke$^-$	1 ke$^-$ or below
L1 trigger latency	2-3 μs	2-3 μs	6-20 μs

Table 1.6: Comparison among different LHC operating conditions across machine commissioning phases. The foreseen upgrade to HL-LHC will introduce unprecedented rate and radiation levels, 10 times larger than original design values.

1.7 Evolution of the CMS pixel detector through 2020

The original design goal of the LHC was to operate at 10^{34} cm^{-2}s^{-1} luminosity and 25 ns bunch spacing. This is the nominal operating scenario for which the overall CMS detector has been designed for. As discussed, after LS1 the evolution of machine performance foresee a continuous increase of the instantaneous luminosity, aiming to reach 2×10^{34} cm^{-2}s^{-1} or higher before 2020, twice the initial design value. With such a luminosity the average pileup will more than double, increasing the number of particles emerging from the interaction region at each bunch crossing. Thereby CMS must be prepared to operate for the rest of this decade with an average number of multiple proton-proton interactions per bunch crossing of 50 or higher.
Tracking and vertexing capabilities of the current pixel detector are essential for the CMS physics research program. Nevertheless, the detector was not designed to perform effectively in such pileup scenario. The silicon pixel tracker is in fact optimized to record efficiently and with high precision 3 space-points near the interaction region in the $|\eta| < 2.5$ range under the assumption of nominal LHC operating conditions. Due to data losses in the present PSI46V2 read out chip, the current pixel system cannot withstand the evolution of machine performance throughout the whole *Phase*-1 LHC commissioning. Furthermore, the innermost barrel layer needs a replacement due to radiation damage on sensor modules. In order to run efficiently at luminosities beyond original design specifications, a pixel detector upgrade is therefore required.
The goal of the CMS pixel upgrade program is to replace the present pixel detector with one that can maintain or possibly improve high tracking efficiency and low fake rates at luminosities up to 2×10^{34} cm^{-2}s^{-1} and average pileup 50 or higher.
Based on the past CMS experience in planning, constructing and commissioning of the present pixel detector, upgrade activities have started more than 5 years ago, requiring physics performance studies, detailed simulations for the new pixel detector configuration and the design of improved readout electronics. An in-depth description can be found in [Chatrchyan 2012].
The upgrade of the CMS pixel detector, referred to as *Phase*-1 *pixel upgrade*, foresee a replacement of the current pixel detector during an extended winter technical stop between 2016-2017. Thanks to the modular design of the CMS detector in fact, this can be accomplished in a short amount of time without introducing disturbance to the volume occupied by the outer silicon strip tracker[9].
A few modules of the new detector have been already inserted into the CMS forward region in 2014 during the LS1, in order to ensure that the new data acquisition will be fully integrated into CMS framework. In order to cope with much higher track densities due to increased pileup, the upgraded pixel detector exhibits a new geometry. The main focus has been on lowering the material budget and adding more tracking points. Moreover, significant changes have been introduced to the current PSI46V2 pixel readout chip in order to reduce data loss.
As shown in Figure 1.20 and 1.21, in the barrel region the number of layers will be increased from 3 to 4. A fourth layer will be added at a mean radius of 16 cm. The current innermost layer will be replaced and brought closer to the interaction region, decreasing its mean radius from 4.4 cm to 3 cm. This will also require a replacement of the current beam pipe. For purely geometrical reasons a reduction in the radius of the innermost pixelated layer will further improve the impact parameter resolution. In addition, a third endcap disk on each side will also be added, along with a new arrangement of detector modules. The design comprises of thinner detector modules and a lightweight mechanical support structure. Large efforts have been made to reduce the material budget, move material from services out of the tracking region.
Sample performance simulation plots for the tracking efficiency and the fake under different pileup conditions and different detector configurations are presented in Figure 1.22.

[9] In contrast, a replacement of the entire outer silicon strip tracker is planned only during the third long shutdown (LS3) due to its much higher complexity.

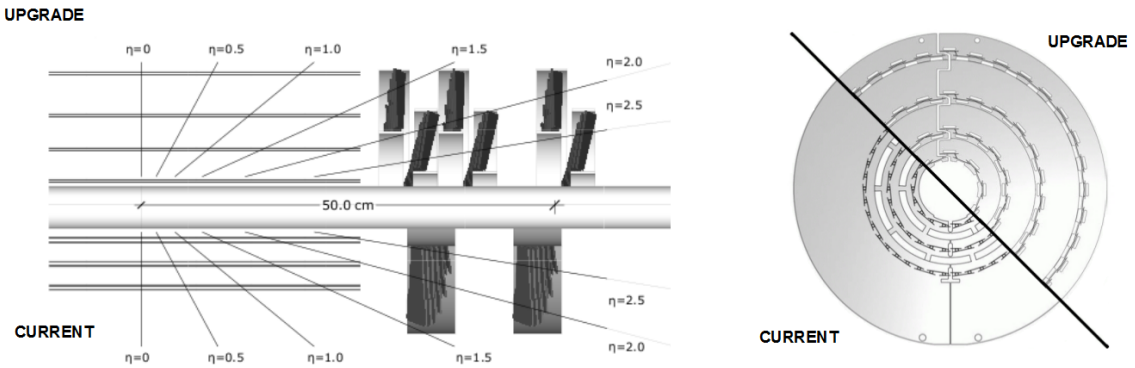


Figure 1.20: Comparison between the current 3-layers/2-disks (bottom) and the *Phase*-1 upgrade 4-layers/3-disks (top) CMS pixel detector layout.

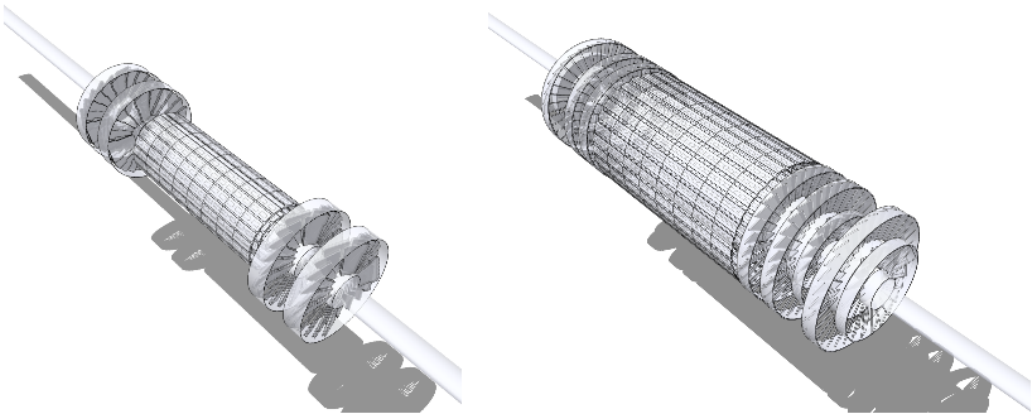


Figure 1.21: Transverse views of current and upgraded CMS pixel detector layouts.

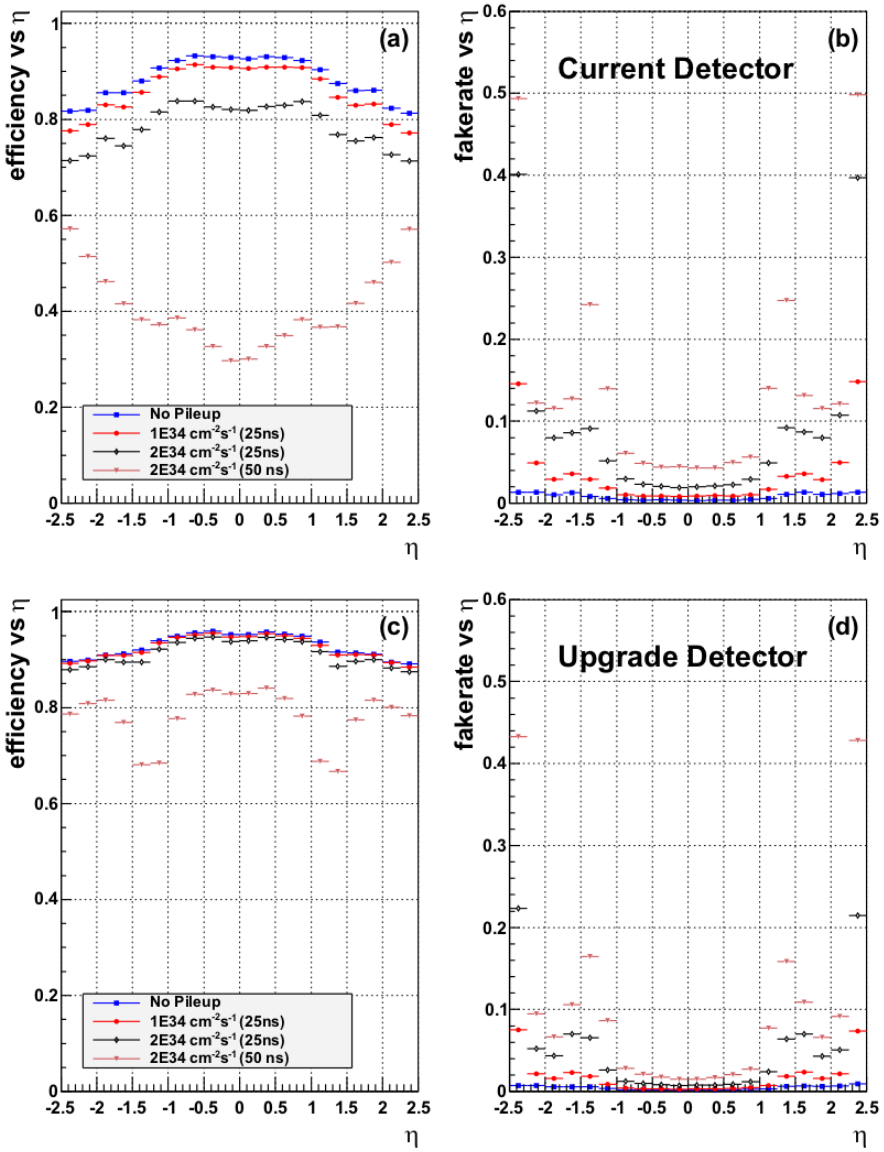


Figure 1.22: Simulated track reconstruction efficiency and fake rate as a function of pseudorapity η for the current pixel detector (a-b) and the upgraded detector (c-d) under different pileup conditions (25, 50 and 100). Simulated data have been obtained from a *ttbar* MC sample [Chatrchyan 2012].

Given its reduced radial position and increased pileup, the innermost barrel layer will experience a much higher charged particle flux, up to 200 MHz/cm^2. Hence the radiation damage will more than double, up to 3.5 MGy TID assuming 10 years LHC operations.
Sensor technology and pixel size will remain unchanged. Hence the pixel rate will increase to about 600 MHz/cm^2. As the number of pixel hits per unit time increases, the number of storage buffers cells for time-stamps and pixel data has to be increased as well. Significant modifications have been therefore implemented to the current pixel readout chip in order to cope with new data rate specifications. The PSI46V2 has been designed in fact to be efficient at nominal 10^{34} cm^{-2} s^{-1} LHC luminosity, but it is inappropriate for a peak luminosity of 2×10^{34} cm^{-2} s^{-1}.
Readout limitations are not inherent to the core double-column readout architecture, but resides in data loss mechanisms due to buffering capabilities and speed of data links at the chip periphery. Such data loss mechanisms have been extensively investigated with detailed high-level data flow simulations for the present ROC. Simulations demonstrated that under new LHC operating scenarios the hit recording efficiency would drop below 60% for the innermost barrel layer.

In order to efficiently operate at increased data rates, an improved version of the present ROC has been developed [Kastli 2013]. The new chip, named PSI46DIG, is implemented in the same 0.25 μm CMOS technology with radiation-hardened layout design. The core of the chip architecture is mostly unaltered. Performance in the analogue Front-End have been improved, reducing the nominal lowest achievable in-time charge threshold from 3.5 ke^- to 1.5-2 ke^-. Furthermore, pixel cells do not include a dedicated leakage current compensation circuit anymore.
Major modifications have been implemented instead at the chip periphery, introducing a digital readout of pixels information. As shown in the schematic block diagram in Figure 1.23, the new ROC uses an 80 MHz, 8-bit Successive Approximation Register (SAR) A/D converter that digitizes the sampled analogue pulse-height information. Hence charge digitization is performed at the chip periphery, allowing for a faster readout into the digital domain. Buffering is still performed in analogue form using capacitor arrays. An analogue information is therefore stored in the chip periphery during the trigger latency. Nominal L1 trigger rate (100 kHz) and trigger latency (3.2 μs) remained unchanged. Nevertheless, the number of buffers has been increased to avoid data loss that affect the current ROC. The number of time-stamp buffers has been doubled, from 12 to 24. The number of analogue buffers has been increased from 32 to 80. Only hits associated to an L1A are digitized. Digitized values and pixel address (row address and double-column address) are fed to a First-In First-Out (FIFO) buffer controlled by an improved version of the current TBM chip. A digital serial readout is performed using a serializer running at 160 Mbit/s through LVDS data links, replacing the previous full-analogue serial readout at 40 MHz in the PSI46V2 chip. A 160 MHz clock is therefore generated from the external 40 MHz distributed to the pixel module using a Phase Locked Loop (PLL). The PLL also generates the 80 MHz clock required by the SAR ADC. Beside higher speed and increased rate capabilities, the digital readout removes the need of complex multi-level decoding of analogue signals performed in the current chip.
Additional changes and improvements have been made for ease of the system. As an example, a power-up reset circuit guarantees that the system starts up in a well defined, low-power state.

The PSI46DIG offers significant improvements with respect to the present CMS pixel readout chip. It was designed to guarantee high tracking performance for the entire *Phase*-1. The increased usage of digital components to perform a full-digital readout required the design of advanced digital and mixed-signal integrated circuit blocks, such as a SAR ADC, a PLL, a serializer and LVDS output drivers. Following the accepted terminology, these new features makes the PSI46DIG a *second generation* pixel ASIC. In the next section, ASIC requirements for a pixel readout chip able to withstand extreme rates and radiation levels foreseen after the installation of the HL-LHC upgrade are discussed. This will require the design of a new *third generation* pixel ASIC for CMS.

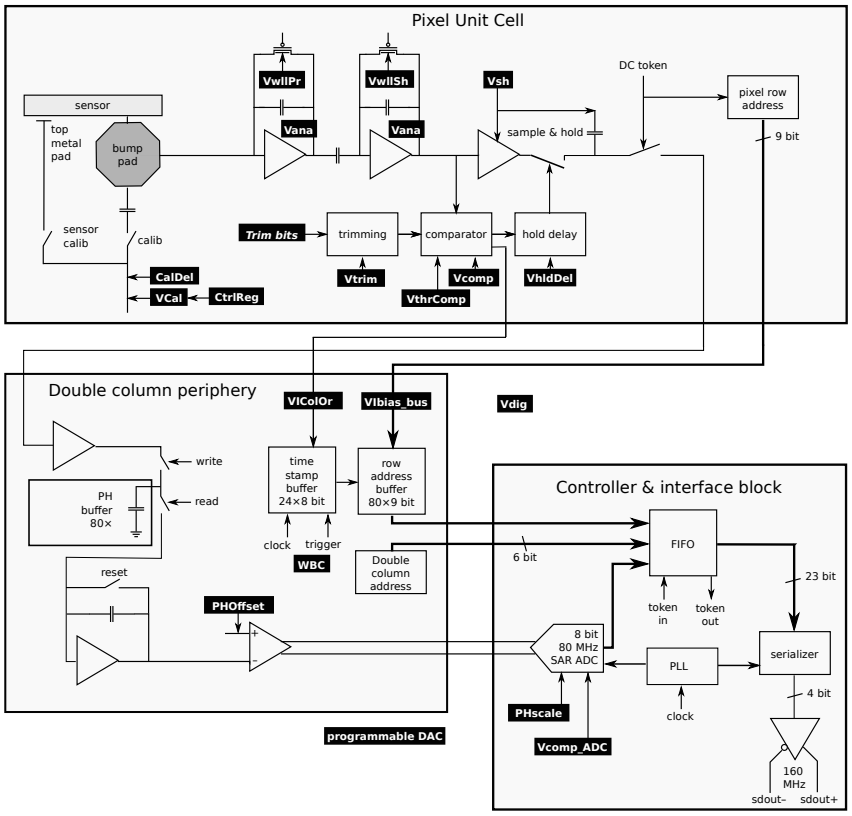


Figure 1.23: Schematic block diagram of the PSI46DIG chip. In the new ROC, charge digitization is performed by an 8-bit SAR ADC and the previous full-analogue readout at 40 MHz has been replaced with a digital readout at 160 Mbit/s through LVDS links.

Parameter	PSI46V2	PSI46DIG
ROC size	7.9 mm × 9.8 mm	7.9 mm × 10.2 mm
pixel size	100 μm × 150 μm	100 μm × 150 μm
charge readout	full-analogue	digital, 8-bit
readout speed	40 MHz	160 Mbit/s
time-stamp buffer size	12	24
data buffer size	32	80
double-column readout speed	20 MHz	20 MHz or 40 MHz
PLL for clock multiplication	no	80 MHz/160 MHz
in-time threshold	3.5 ke^-	$<$ 2 ke^-
leakage current compensation	yes	no
data loss at max. particle flux	3.8% at 120 MHz/cm^2	1.6% at 150 MHz/cm^2

Table 1.7: Comparison between PSI46V2 and PSI46DIG readout chip specifications [Gray 2013]

1.8 Pixel ASIC requirements for HL-LHC

The *Phase*-1 pixel detector upgrade will guarantee necessary tracking performance for the entire *Phase*-1 LHC commissioning. As already discussed, after a third long shutdown (LS3) foreseen for ∼2022-2023 the discovery potential of the machine will be further increased. With the installation of HL-LHC the accelerator will be able to deliver proton-proton collisions with an instantaneous luminosity up to 10^{35} cm^{-2}s^{-1}, one order of magnitude higher with respect to the original design value. With a luminosity 10-times larger and a centre-of-mass energy of 14 TeV, the nominal bunch collision rate of 40 MHz will introduce an unprecedented pileup, resulting into extreme track densities and radiation levels. The number of multiple proton-proton interactions per bunch crossing can potentially increase up to 200. Ongoing simulation studies suggest that taking into account of luminosity levelling[10] CMS and ATLAS detectors will experience a pileup of 140-150. In a worst-case scenario with 50 ns bunch spacing (only envisaged if the machine cannot get sufficient luminosity at 25 ns bunch spacing due to already mentioned electron cloud effects) the pileup will further increase, up to 300. Performing efficient tracking and vertexing with such an impressive number of overlapping interactions is a challenge.
At 10^{35} cm^{-2}s^{-1} luminosity the expected charged particle flux will increase to about 500 MHz/cm^2, one order of magnitude higher than the design value of 50 MHz/cm^2. Aiming to collect 3000 fb^{-1} total integrated luminosity in 10 years, the radiation damage at HL-LHC will be of the order of 10 MGy Total Ionizing Dose (TID), corresponding to 2×10^{16} (1 MeV) n_{eq}/cm^2.
Such operating conditions will require the installation of a new pixel detector able to withstand unprecedented track densities and radiation levels. Furthermore, the outer tracker will be replaced with major modifications during LS3. *Phase*-2 pixel upgrade activities already started in CMS [Chatrchyan 2012]. At the time of writing, a technical proposal that will be followed by a detailed technical design report for the CMS *Phase*-2 pixel detector upgrade are under preparation. The design of a new pixel ASIC able to operate at HL-LHC introduces major challenges on several fronts.

Physics performance and detector configuration are currently under investigation within CMS pixel upgrade simulation groups. The *Phase*-2 pixel detector layout is assumed to be very similar to the 4-layers upgraded configuration discussed in Section 1.7. Major modifications will be introduced in the number of endcap disks in order to improve the forward coverage up to $|\eta| < 3.5$ or above. Ongoing geometry studies foresee a number of 7-10 disks each side.
A pixel detector featuring higher granularity and spatial resolution will be essential in order to cope with the unprecedented pileup, increasing two-track separation performance. The usage of silicon sensors with smaller pixel size is therefore required. Despite it is assumed that hybrid pixel detectors bump-bonded to the new readout chip will be employed, a sensor choice for the *Phase*-2 pixel upgrade has not been yet finalized. A reach variety of sensor solutions was developed to substantially increase the radiation tolerance with respect to the $n-$on$-n$ approach extensively adopted in current pixel detectors at LHC. At present, both thin planar sensors and 3D sensors with a thickness in the range of 100 μm to 150 μm are being evaluated by the CMS pixel sensor community [Ravera 2013]. A final choice will be assessed only upon extensive sensor simulations and beam test results. The usage of different sensor technologies at different radii cannot be excluded. Nevertheless, realistic estimates for the sensor capacitance value (including strays from the bump bonding and from adjacent pixels), for the leakage current and for the minimum detectable charge are essential input specifications for the design of an optimized analogue Front-End chain in the readout chip.

[10] Proton beams circulating in the LHC can be moved relative to each other in order to modify the cross sectional area available for interactions, thus increasing or diminishing the instantaneous luminosity. This is referred to as luminosity levelling.

The choice of a pixel size is the most critical point and represents a delicate optimization between physics performance requirements, sensor technology, bump-bonding technology and integration density offered by the CMOS fabrication technology adopted for the implementation of the readout chip. On the one hand, due to charge sharing the optimal resolution in planar sensors is achieved if the pixel size is comparable to the Lorentz drift. Therefore the optimal segmentation depends on the sensor thickness. On the other hand, 3D sensors are insensitive to the magnetic field. Hence the spatial resolution is determined by the particle incidence angle, pixel size and effective charge threshold. Ongoing simulation studies for the new CMS pixel detector indicate that pixel cells of 50 μm $\times$ 50 μm or 25 μm $\times$ 100 μm are required to ensure high tracking and vertexing performance at HL-LHC, 1/6 of the current pixel size 100 μm $\times$ 150 μm. Thereby on-pixel analogue and digital components must fit a maximum area of 2500 μm^2. With a proper choice of the bump-bonding pattern, a readout chip designed with 50 μm $\times$ 50 μm pixel unit cells (PUCs) can also be bump-bonded to silicon sensors with different pixel aspect ratios.

Each pixel cell must incorporate an analogue Front-End chain to perform signal amplification, shaping and hit discrimination. The foreseen usage of thinner sensors in order to increase the radiation tolerance determines reduced charge signals. A minimum detectable charge of 1 ke^- has been constrained, requiring the design of very low-noise and low-threshold analogue Front-End electronics. According to the topic of this work, an in-depth description of all design specifications defined for the analogue Front-End system (input capacitance, leakage current, etc.) is remanded to Chapter 2, describing the practical implementation of an analogue Front-End chain suitable for the CMS *Phase*-2 upgrade.

Track densities foreseen at HL-LHC introduce extreme data rates. Detailed simulation studies for the expected pixel occupancy under high pileup scenarios shows in fact that the new pixel ASIC will have to handle hit rates[11] per unit area of the order of 1.5-2 GHz/cm^2 in the innermost layer. The new pixel chip will exhibit a much larger IC format, of the order of 2 cm $\times$ 2 cm. A $\sim$4 cm^2 pixel ASIC must internally handle a raw data rate $R = 4$ cm^2 $\times$ 1.5 GHz/cm^2 $\times$ B, where B is the number of bits needed to store hit information (pixel address, time stamp and charge information). For instance, if 16 bits are required, the data rate is 100 Gbits/s. The design of a new readout chip able to internally handle the pixel information from thousands pixel cells with such data rates is a challenge. Under these conditions, data transfers in true analogue form towards the chip periphery as in PSI46V2 and PSI46DIG chips are no more feasible.
As already discussed, only triggered events must be readout, requiring buffering, time-stamping and trigger matching functionalities. Hit information must be stored until the arrival of a trigger determining which event data to read out. Upon preliminary trigger studies, at HL-LHC the CMS L1 trigger latency is expected to increase at about 6.4 μs or above. Worst-case scenarios indicate up to 20 μs. Furthermore, the L1A trigger rate is expected to increase due to reduced rejection power in calorimeters at increased pileup. Worst-case scenarios indicate trigger rates up to 1 MHz, 10-times larger than the nominal 100 kHz value.
On-chip data buffering during a 20 μs L1 trigger latency corresponds to an information storage need of 2 Mbits per chip. In order to accommodate local fluctuations in the data rate without overflow, the amount of memory needs to be a factor 6 to 8 times larger, getting to a required buffering capability of up to 16 Mbits per chip. Assuming no on-chip data reduction, for an operating scenario with 1 MHz L1A trigger rate the required readout bandwidth per pixel chip is of the order of 3.6 Gbits/s. A much higher readout speed is therefore required, demanding for the usage of high-speed data transmission circuits.

[11] hit rate = occupancy $\times$ 40 MHz

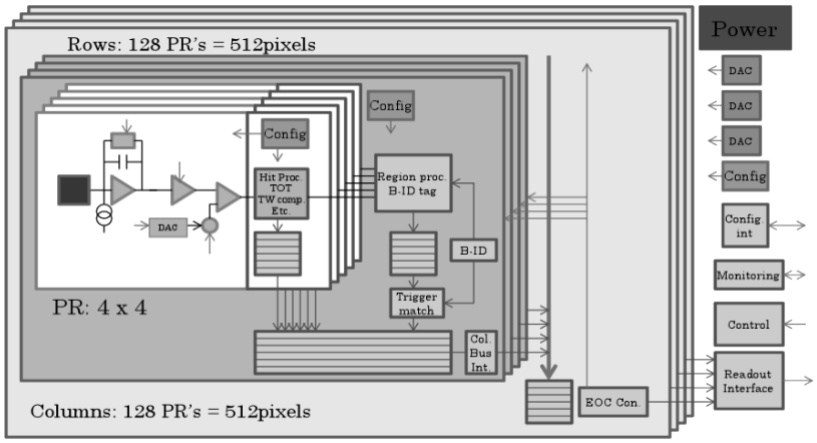


Figure 1.24: Schematic block diagram for a possible global architecture of a third generation pixel ASIC for CMS [Christiansen 2013]. With a hierarchical organization, pixel cells are grouped in pixel regions (PR), pixel regions in columns and column pairs in a full-matrix. Charge digitization, buffering and trigger matching are performed at the pixel level, minimizing data transfers towards the chip periphery.

Parameter	Specification
number of barrel layers	4
number of endcap disks	2×5
pileup	140-200
charged particle flux	500 MHz/cm^2
total radiation damage	10 MGy TID/10y
sensor technology	2D thin planar, 3D
pixel size	50 × 50 μm or 25 × 100 μm
chip size	$\sim$ 4 cm^2
hit rate	1-2 GHz/cm^2
hit time resolution	$<$ 25 ns
signal threshold	1-1.8 ke^-
charge resolution	4-8 bits
max. acceptable hit loss	$<$1%
L1 trigger rate	100-200 kHz
trigger latency	6.4-20 μs
power budget	$<$ 0.4 W/cm^2
hit memory per chip	16 Mbits
output bandwidth	3600 Mbits/s

Table 1.8: General ASIC specifications for the new CMS pixel readout chip.

1.9 IC technology choice

The usage of a modern CMOS fabrication technology represents the only means to address the challenges of reduced pixel size, 10-times higher hit rates and 100-times higher readout data rates demanded for a pixel ASIC suitable for HL-LHC pixel upgrades. The choice of a specific CMOS process represents a critical decision in perspective of a complex and long-term design project. Several considerations determined the baseline solution adopted by CMS for the implementation of the new pixel readout chip.

Technology qualification, prototyping, testing, design optimization and final large-scale production are tasks that extend over several years. The typical time scale required for the implementation of custom and highly specialized integrated circuit solutions is of about 5-10 years. A large majority of the required building blocks have to be optimized for the specific application according to performance of the selected CMOS process. Once a specific fabrication technology is chosen it is difficult to translate the designs into another technology platform halfway through a project. Hence the usage of a well-established technology with long-term availability and support for the entire duration of design and test activities is of utmost importance. Such a request is in contrast to mainstream technologies for commercial applications, which have relatively short lifetimes. This can be appreciated in the plot of Figure 1.26, that shows a comparison between technology nodes adopted by the HEP community and mainstream technologies in industry.
Given the extreme radiation levels expected at HL-LHC the the radiation hardness requirement is exacerbated. The process must exhibit a very high radiation tolerance for nearly a 10-years lifetime in the unprecedented HL-LHC radiation environment, up to 10 MGy and 10^{16} (1 MeV) n_{eq}/cm^2 for the innermost barrel layer.
The complexity of the new pixel ASIC is comparable to modern commercial integrated circuits developed in industry. A technology featuring high integration densities for both analogue and digital functions is therefore required. The technology must be appropriate for the integration of several thousands analogue Front-End amplifiers/shapers and hit discriminators together with a large amount of synthesized digital logic for data buffering, trigger matching and readout. The availability of a complete mixed-signal design kit is essential. Furthermore, low-power operations are a vital constraint in order to instrument large areas. Last but not least, fabrication costs must be affordable by the HEP community, from small prototype circuits to full-size readout chips and large-scale engineering runs.

In the last two decades, commercial deep-submicron CMOS processes have been successfully used to implement radiation-tolerant custom integrated circuits for radiation detection in particle physics experiments. Most of current Front-End ASICs operating at LHC experiments have been designed in a CMOS 0.25 μm process. As already mentioned, such a technology has been proved to be radiation tolerant only by following special radiation hardening layout techniques for all MOS transistors in both analogue and digital domains. However, an enclosed-layout transistor (ELT) occupies about twice the silicon area in comparison to an equivalent device with standard planar layout. Furthermore, the necessity of using full-custom layout solutions required the design and characterization of new analogue and digital libraries from scratch.
Thanks to its reduced gate oxide thickness, a commercial 130 nm CMOS technology officially supported by CERN showed a better radiation tolerance even without radiation hardening by design [Gonella 2007]. With increased integration density capabilities, reduced power dissipation (1.2 V core voltage) and better radiation tolerance such a technology is currently adopted for several short- and mid-term projects in the HEP community. Many of state of the art second generation pixel ASICs have been implemented using this 130 nm CMOS process [Ballabriga 2006, Llopart 2007, Calvo 2008, Garcia-Sciveres 2011, Rolo 2013].

The 130 nm technology node has been considered as a possible option for the high luminosity CMS pixel upgrade. It is however estimated not to offer enough logic density to fulfill all the requirements for a HL-LHC application. Radiation tolerance and higher integration density requirements led to the choice of a commercial 65 nm CMOS fabrication technology as a promising candidate for the implementation of the new pixel readout chip. Such a 65 nm CMOS process is mature and long-term available in industry. It is widely used for industrial and automotive applications that require availability over extended periods. Thereby it is expected to remain commercially available across the full time scale required for CMS *Phase*-2 pixel upgrade activities. The 65 nm technology is now officially supported by CERN, with a recently defined foundry frame contract and a tapeout mixed-signal design kit developed as a close collaboration between the CERN microelectronics group, the design tool supplier, the foundry access service and the manufacturer. It is also a technology node considered reasonably affordable by the HEP community.
The chosen 65 nm CMOS process offers about 4-times higher logic density compared to the 130 nm node and about 30-times compared to 0.25 μm with enclosed layout devices. Hence 65 nm CMOS is well suited to fit considerably more logic at the pixel level, offering a concrete chance to move buffering and trigger matching at the pixel level, as demanded for the future pixel ASIC. Thanks to technology scaling speed performance significantly increase, with a single MOS transistor intrinsic frequency of the order of GHz. It is also a technology with excellent low-power characteristics for both analogue and digital functions. The 65 nm technology has also shown to be feasible to achieve low-noise performance for analogue Front-End amplifiers [Manghisoni 2011]. Prototypes of analogue Front-End electronics have been already successfully implemented in 65 nm CMOS, as reported in [Gaioni 2011, Valerio 2012, Mekkaoui 2013, Havranek 2014].
This technology is being extensively tested for radiation tolerance. The usage of MOS transistors with thin gate oxides is a fundamental requirement for the radiation tolerance. The chosen 65 nm CMOS is the highest technology node for which thin SiO_2 oxide gate dielectrics are used by all producers, whereas for higher density nodes manufacturers moved to thicker high-K dielectrics. The chosen 65 nm technology has already been proved to have excellent radiation tolerance up to a total dose of 3 MGy [Bonacini 2011]. Ongoing radiation qualification studies are now exploring radiation effects in 65 nm up to 10 MGy. Initial but very preliminary indications have shown that PMOS devices with inappropriate sizing experience a significant degradation of certain parameters above the 3 MGy level. Significant additional work is required to develop an understanding of such effects. In addition to single-transistor structures, detailed characterizations must be performed on logic cell libraries. Radiation-induced Single Event Upsets (SEU) in 65 nm CMOS technology are handled with same techniques already adopted in previous technologies, using triple modular redundancy (TMR) and Hamming encoding for most critical digital functions. The measured SEU cross-section for 65 nm is slightly better than previous technology generations, though with an increased risk of multiple bit upsets in neighbouring elements.

Better performance in terms of speed, logic density, low-power consumption and radiation tolerance make the 65 nm technology node a favoured candidate with respect to the previous 130 nm CMOS process extensively adopted for several *Phase*-1 electronics upgrades at LHC experiments and other projects within the HEP VLSI community. All design activities have started assuming 65 nm as a baseline choice. At the time of writing, the chosen 65 nm technology represents the most advanced technology node for the implementation of full-custom solutions for radiation detection and measurements in particle physics and medical applications. As discussed in the next section, research activities on 65 nm CMOS technology are now part of the official international RD53 collaboration and of the Italian INFN CHIPIX65 project.

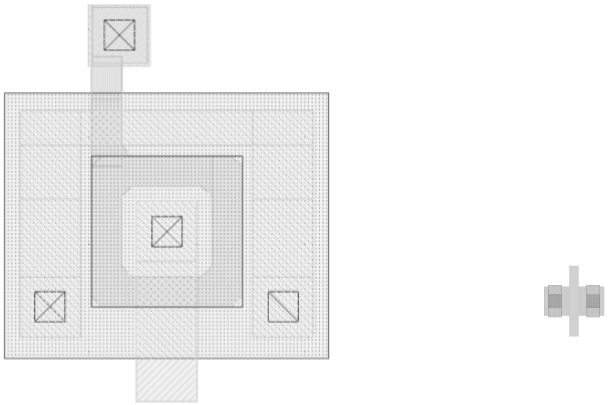


Figure 1.25: Full-custom enclosed layout transistor (ELT) in a 0.25 μm CMOS technology and comparison with a minimum size planar device from the 65 nm library.

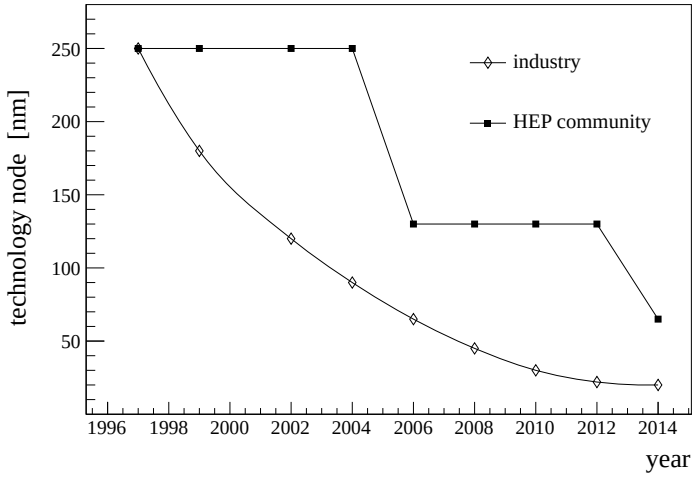


Figure 1.26: Mainstream technology nodes adopted by the HEP VLSI community and comparison with the industry trend in the last decade.

1.10 New pixel ASIC research communities on 65 nm CMOS

Unprecedented radiation levels, pileup and track densities foreseen at HL-LHC are common to both ATLAS and CMS experiments. As shown in Table 1.9, most of general ASIC requirements and specifications for ATLAS and CMS *Phase*-2 pixel upgrades are similar. Furthermore, both ATLAS and CMS pixel ASIC communities identified 65 nm CMOS technology as the baseline choice for the development of their new readout chips.
A significant amount of preliminary work is necessary to assess the overall design foundation and characterize the radiation tolerance of a new technology, requiring large efforts and man power.
Following a joint ATLAS and CMS workshop held at CERN in November 2012, it turned out that a cross-experiment research framework would have been the most efficient strategy to resolve most of common challenges in using and qualifying the chosen 65 nm process.
A first letter of intent followed by a detailed collaboration proposal have been presented to the CERN LHC Committee (LHCC) in June 2013 [Christiansen 2013]. This led to the constitution of a new 3-years research-and-development program, named RD53, officially supported by CERN for the development of pixel readout integrated circuits for extreme rates and radiation at HL-LHC.
The collaboration is extended to other non-LHC experiments and groups interested in using the 65 nm CMOS technology for pixel detectors, such as the Linear Collider Detector (LCD) project within the Compact Linear Collider (CLIC) international collaboration.
A cross-experiment collaboration does not imply that ATLAS and CMS will adopt the same pixel ASIC for their own *Phase*-2 upgrades. Indeed, the overall design foundation using 65 nm CMOS, tests and radiation qualification efforts are independent of the specific implementation of final pixel readout chips that ATLAS and CMS experiments will adopt. As summarized in Table 1.10, RD53 research activities have been organized in dedicated working groups, covering technology qualification and all design aspects related to the implementation of a pixel readout chip of unprecedented complexity.

With a strong Italian component from seven INFN units (Bari, Milano, Padova, Pavia, Perugia, Pisa and Torino) the RD53 collaboration involves about 20 institutes around the world, with about 100 members, of which about 50% are ASIC designers. RD53 will over the next years develop the methods and design foundation needed to produce next generation pixel integrated circuits that can deliver the higher performance required under the extreme operating conditions at HL-LHC.
Given the foreseen importance of 65 nm CMOS in the forthcoming years and the strong INFN component involved in the RD53 collaboration, Italian CMS and ATLAS groups participating to the project have submitted in July 2013 a detailed proposal to INFN/CSN5 to finance a new 3-years research program on such a technology node [Demaria 2013], leading to the approval of the CHIPIX65 project in October 2013. With 35 members experts on the field, of which 20 ASIC designers, the CHIPIX65 project involves a substantial fraction of INFN expertise on integrated circuit design. This makes CHIPIX65 a unique opportunity for an efficient propagation across INFN of CMOS 65 nm technology and constitutes the largest collaboration on a microelectronics project ever made across INFN.

The overall research activity presented in this thesis is framed in the context of RD53 collaboration and the CHIPIX65 INFN project. Next chapters will present personal contributions on the design and test of both analogue and digital integrated circuits in 65 nm CMOS technology as part of the RD53 and CHIPIX65 research programs. These studies have provided the necessary first steps towards the design of a future complete hybrid pixel ASIC demonstrator suitable for the long-term CMS pixel detector upgrade.

Parameter	ATLAS	CMS
number of barrel layers	4 (option 5)	4
number of endcap disks	2×6	2×5
chip size	~ 4 cm^2	~ 4 cm^2
pixel area	2500 μm^2	2500 μm^2
pileup	140-200	140-200
charged particle flux	500 MHz/cm^2	500 MHz/cm^2
hit pixel rate	~ 1 GHz/cm^2	1-2 GHz/cm^2
10 years TID	10 MGy	10 MGy
hit time resolution	25 ns	25 ns
charge digitization	4-5 bits	4-8 bits
max. acceptable hit loss	<1%	<1%
L1 trigger rate	100-200 kHz	100-200 kHz
trigger latency	6.4 μs	6.4-20 μs
power budget	0.3 W/cm^2	< 0.4 W/cm^2
hit memory per chip	16 Mbits	16 Mbits
output bandwidth	3600 Mbits/s	3600 Mbits/s

Table 1.9: Comparison between ATLAS and CMS pixel ASIC specifications [Christiansen 2013]. Due to very similar requirements, ATLAS and CMS pixel ASIC communities proposed a new RD collaboration as the most efficient solution to address common challenges in the design of readout integrated circuits for their *Phase*-2 pixel detector upgrades.

Working group	Research activities
Radiation	qualification of 65 nm CMOS technology up to 10 MGy TID and 10^{16} n_{eq}/cm^2, evaluation of digital libraries after irradiation
Top-level	definition of a top-level architecture and chip integration strategies (power distribution, clock distribution, etc.)
Simulation	development of a full-chip simulation framework, definition and optimization of digital architectures
I/O	definition of readout and control interfaces, I/O protocols
Analogue	evaluation and design of different analogue Front-End solutions
IP blocks	design and optimization of general-purpose analogue, digital and mixed-signal blocks (PLLs, ADCs, DACs, references etc.)

Table 1.10: RD53 working group activities.

Part II
Research activity

Chapter 2

Synchronous pixel Front-End design in 65 nm CMOS technology

Within RD53 and CHIPIX65 collaboration frameworks, ASIC designers of the CMS Torino Tracker group have been primarily involved in the development of pixel analogue Front-End electronics in 65 nm CMOS technology. The basic layout of the chosen pixel cell architecture was then used to assemble two small pixel matrices. They were put on silicon as part of a dedicated chip in the first October 2014 CHIPIX65 submission, along with further analogue Front-End solutions designed by other INFN groups. In this chapter, an in depth description of the initial concept, simulation and physical implementation of the proposed Front-End chain is given. The assembling of pixel matrices and top-level integration issues for the full chip will be addressed in Chapter 3 instead.

Keywords: Front-End amplifier, ENC, CSA, Krummenacher feedback, leakage current compensation, TOT, hit discrimination, synchronous comparator, latch, positive feedback, offset, autozeroing, VCO, asynchronous logic

2.1 Introduction

Large design efforts have been dedicated to the realization of a first pixel Front-End prototype in 65 nm CMOS, with the additional challenges related to the usage of a completely unknown fabrication technology in Torino. Certainly the primary goal has been to meet all CMS-specific performance requirements and general guidelines defined within the RD53 collaboration for the pixel upgrades at HL-LHC. Power consumption and pixel area constraints, as well as low-noise and low-threshold specifications were therefore assumed as constant design benchmarks. Nevertheless, the choice of the Front-End architecture was essentially driven by the aim of explore innovative solutions for both the hit discrimination and a charge encoding performed at the pixel level, taking advantage of increased speeds offered by a 65 nm technology node.
As a result, beside a more traditional implementation of a single-stage charge sensitive amplifier (CSA) with triangular shaping, a discrete-time solution for the hit discriminator is proposed in this work, aiming to demonstrate the feasibility of synchronize the analogue Front-End with the machine activity in CMS. A track-and-latch voltage comparator has been therefore adopted. Thanks to the usage of a positive feedback stage coupled to a low-gain differential amplifier, precise and fast voltage comparisons can be obtained with low-power dissipations, allowing to discriminate very low charge-induced signals above the nominal threshold. The hit generation is synchronized with a 40 MHz master clock, sampling the CSA analogue output around its peaking time, hence providing a reliable solution that naturally overcomes time-walk issues in the time stamp assignment. This is a clear advantage with respect to continuous-time architectures.

Pixel-to-pixel threshold variations are corrected by means of capacitors using an autozeroed scheme, without the need of a local on-pixel D/A converter for digital trimming. This aspect introduces fundamental advantages in perspective of pixel operations in a harsh radiation environment, and efficient calibration schedules for the threshold adjustment can be defined according to machine operations. Finally, flexible and high-speed time-over-threshold (TOT) charge digitizations can be obtained by turning the latched comparator into a compact voltage-controlled oscillator (VCO) using asynchronous logic. As derived from transient simulations, 5-8 bit TOT measurements can be retrieved using locally self-generated clock signals up to GHz frequencies in 65 nm technology. The choice of a synchronous Front-End discriminator introduces therefore several advantages and very promising features that naturally fits into a bunched experiment.
After a detailed review of specification requirements defined for the the analogue pixel cell, the architecture of the Front-End chain is presented. Practical transistor level implementation and optimization of all analogue and mixed-signal blocks is then discussed, justifying most important design choices supported by circuit simulations. Both schematic-only and final post-layout results are presented, selecting from the large set of simulations performed to ensure circuits reliability under different working conditions across process, voltage and temperature (PVT) variations.
Personal contributions are in the design of the high open-loop gain inverting amplifier for the CSA stage, in the overall concept and validation of the proposed synchronous comparator/oscillator architecture and in the design and layout of the necessary on-pixel digital control logic to support discriminator operations for fast TOT-based charge encoding.

2.2 Performance requirements for the analogue Front-End

Input specifications for the design of a new pixel Front-End prototype in 65 nm CMOS have been derived from general performance requirements defined for the CMS *Phase*-2 pixel upgrade within the analogue working group of the RD53 collaboration [Re 2014]. Table 2.1 summarizes fundamental values and constraints. Without doubts most critical design parameters derive from the extremely high track densities and radiation levels foreseen at HL-LHC. Indeed, any realistic design necessary has to meet all CMS-specific pixel detector requirements, which derives from magnetic field configuration, sensors, tracking performance and physics requirements.
As a baseline choice, a pixel size of 50 μm $\times$ 50 μm has been adopted. The usage of square pixels is primarily motivated by the possibility of coupling first ASIC prototypes in 65 nm to sensors with different bump patterns and geometries. This maximizes the overall testability and offers a good compromise between the amount of electronics that can fit into a single pixel cell and nowadays hybrid sensor technologies as well as bonding technologies limitations. Nevertheless, a total pixel area of 2500 μm^2 is a realistic value in order to ensure necessary granularity and keep low the channels occupancy under the foreseen unprecedented track densities at HL-LHC. It is also assumed that the analogue Front-End can occupy a maximum area of 25 μm $\times$ 50 μm, half the total pixel size. This is actually a quite optimistic and relaxed constraint. As a matter of fact, once a pixel cell/pixel region digital architecture will be defined for the readout of the chip, most of the total available area in the pixel will be dedicated to digital electronics, targeting to move as much as possible data buffering and zero suppression from the chip periphery to the pixel level. Thus, in the near future the area constraint for the analogue part can be reduced to about 25-30% of the total pixel size. Certainly a definitive pixel cell geometry for the CMS *Phase*- 2 upgrade has to be defined upon tracking and vertexing performance simulation results according to physics requirements. As already mentioned in Chapter 1, ongoing simulation activities are focused on 50 μm $\times$ 50 μm or 25 μm $\times$ 100 μm pixel sizes.

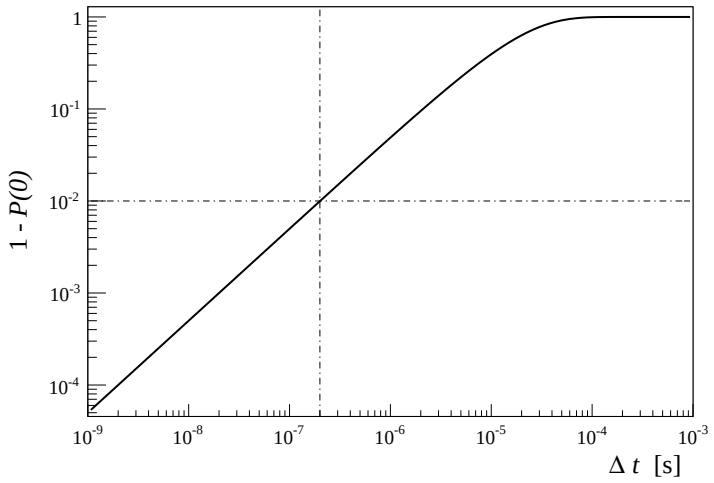


Figure 2.1: Pulse overlap probability as a function of time assuming an average hit rate of 50 kHz/pixel.

The pixel size determines the expected average event rate per pixel. Assuming a pixel area of 2500 μm^2, the foreseen hit rate of 2 GHz/cm^2 in the CMS innermost barrel layer leads to an average hit rate per pixel of 50 kHz. From a bare simulation point of view, this corresponds to a δ-like current pulse presented at the input of the Front-End amplifier on average every 20 μs. This introduces a further design constraint for the maximum duration of the largest signal of interest in order to minimize the probability of pulse overlaps, which at the end would result into a hit detection inefficiency. For a first rough estimation of the overall analogue Front-End processing time necessary to cope with the average event rate per pixel, a Poisson distribution approximation can be adopted [Spieler 2005, Rossi 2006]. In case the arrival time of the events is random with an average event rate r in fact, the expected number of events n within a certain observation time window Δt is given by

$$P(n) = \frac{(r\Delta t)^n e^{(-r\Delta t)}}{n!}$$

This is the case of most of radiation detectors in which the exact arrival time of the events is unknown. The probability to observe no signals during a processing time Δt is $P(0)$, hence the probability that any pulse overlaps occur during the same time is

$$1 - P(0) = 1 - e^{(-r\Delta t)}$$

Figure 2.1 shows the probability $1 - P(0)$ as a function of time for an average event rate of 50 kHz/pixel. In case a continuous-time shaping is adopted, the analogue output signal for the maximum signal of interest has to return to the baseline in less than 200 ns in order to keep the hit detection inefficiency due to pulse overlaps below 1%. This is a conservative design choice under the hypothesis of Poisson-distributed events. Detailed simulation studies for the expected hit rate distributions in CMS at HL-LC are therefore necessary in order to unambiguously constraint the maximum signal duration.

At present, an upper limit of 400 ns has been defined. Certainly the binary hit information must be available in less then 25 ns in order to properly associate the event with the beam crossing (time stamp), which in turn requires an in-time response below 25 ns for the minimum signal of interest. Hence the peaking time of the signal generated by the Front-End amplifier in combination with the delay of the discriminator must be less than 25 ns.

Power consumption is a key issue in the design of the future pixel ASIC. The overall power density cannot increase with respect to the CMS current pixel detector value of 0.4 W/cm^2 determined by the capabilities of the cooling system. Neglecting power contributions due to electronics at the chip periphery and assuming that most of the power is dissipated inside the matrix, the choice of a 50 μm $\times$ 50 μm pixel area introduces therefore a power budget constraint of the order of 10 μW per pixel. As a first guess, it is also assumed that the power dissipation in the pixel cell can be allocated in rough equal contributions between the analogue and the digital parts, resulting into a power budget of about 5 μW/pixel for the analogue Front-End. Within the RD53 collaboration, an upper limit of 6 μW/pixel has been proposed for the analogue part. As a matter of fact, this is the most severe and limiting constraint for the design of the on-pixel analogue processing chain. Assuming a supply voltage of 1.2 V, the total maximum current that can be absorbed by the analogue circuitry is reduced to 5 μA. In practice, no more than 3 μA DC bias currents can be used in the design of the Front-End charge amplifier/shaper and no more than 2 μA for the hit discriminator. For a more conservative design, an upper limit of about 3-4 μA for the total static supply current can be assumed. This reduces the effective number of stages that can be adopted for the signal processing. Without doubts, the design of CMOS analogue building blocks with adequate performance under such low current constraints becomes more challenging.

Most important input specifications for the design of the analogue Front-End in terms of signal polarity, input capacitance, charge resolution and dynamic range arise from sensor and physics requirements. Indeed, a sensor choice for the CMS *Phase*-2 upgrade has not been yet finalized. At present, both planar and 3D sensors with a thickness in the range of 100 μm to 150 μm are being evaluated. The foreseen usage of n-type sensors (collection of electrons) does not require the design of a bipolar solution. The overall Front-End system can be therefore optimized for negative-only input charges and sensor leakage currents. Pixel capacitance values proposed by the CMS sensor community are of the order of 100-150 fF. This range takes into account both different possible sensor technology choices, geometry of pixels (square or rectangular) and parasitic contributions associated with the bump bonds for the interconnection. A nominal total input capacitance of 100 fF is assumed for the overall optimization of the analogue Front-End.
The actual charge value integrated by the Front-End amplifier depends on the sensor thickness, charge sharing among pixels and radiation damage. The average number of electron-hole pairs per unit length generated in thin silicon sensors is about 100 pairs/μm for a minimum ionizing particle (MIP). Due to Landau fluctuations, the most probable value (MPV) is slightly reduced. According to the well know approximation [Bichsel 1988]

$$Q_{\mathrm{MPV}} = \left[53e^- + 4.5e^- \ln\left(\frac{t}{\mu\mathrm{m}}\right)\right] \frac{t}{\mu\mathrm{m}}$$

for a sensor thickness t of 100 μm it is expected that before irradiation a MIP most likely releases a charge signal of about 7 ke^- (1.12 fC), whereas the average value is about 10 ke^- (1.6 fC). At the maximum level of radiation damage foreseen at HL-LHC, the actual collected charge will be reduced by a factor 2-3. In the worst-case, a MIP charge of 4 ke^- is expected at highest radiation damage, under the assumption that charge collection efficiency can be recovered with the usage of larger reverse bias voltages and introducing charge multiplication mechanisms in the sensors. Certainly charge sharing due to magnetic field further contributes in reducing the effective amount of charge presented at the input node.

Thereby very low-threshold performance are demanded for the analogue chain, targeting to a minimum detectable charge of 1000 e^-. Furthermore, nominal leakage currents up to 10 nA after irradiation must be taken into account. In case charge multiplication is adopted, a worst-case value of 20 nA is assumed.
Charge digitization resolution and dynamic range specifications are determined by tracking and physics requirements. As discussed in Chapter 1, the CMS tracking system heavily relies on charge measurements provided by the silicon pixel tracker. Thus, in order to maintain or possibly improve CMS tracking performance at HL-LHC, energy loss measurements must be available from pixels. Furthermore, in perspective of pixel operations under unprecedented radiation levels, the charge information will be also of primary importance for monitoring radiation damage effects on sensors charge collection efficiency. In order to move as soon as possible the signal processing into the digital domain, charge digitization must be performed at the pixel level, transferring only digital information outside the pixel cell. The actual digitization resolution is still under investigation and is going to be defined from CMS *Phase*-2 pixel upgrade simulation groups. At present, guideline values are in the 5 to 8-bit range. Certainly the A/D conversion must be accomplished in less than 400 ns according to the maximum signal duration defined for the analogue Front-End. A linear response must be guaranteed up to 4 MIP charge, which in turn corresponds to about 30 ke^- MPV before irradiation. For signal charges larger than 4 MIP, the analogue output signal should restore to the baseline in less than 1 μs in order to avoid an excessive dead time, demanding for the usage of a dedicated recovery circuit for large signals up to 10 MIP full dynamic range. Note that more data buffering will be required to withstand the expected increase of the L1 trigger latency up to 20 μs. As a matter of fact, due to the limited effective amount of digital circuitry that can fit into a pixel cell or a pixel region, the chance to perform the zero suppression at the pixel level can become unrealistic if a large number of bits is demanded for the charge measurement. In case local buffering can not be implemented, charge encoded data necessary will be sent to the chip periphery, increasing power dissipation. Physics simulations play therefore a fundamental role also in defining any realistic digital pixel readout architecture.

Finally, the noise hit rate must be minimized according to minimum detectable charge and leakage current constraints. The discriminator threshold value must be set low enough in order to maximize the detection efficiency, but well above the noise floor in order to keep fake hit events at negligible levels. Careful transistor level optimization is therefore required to minimize the electronic noise at the output of the Front-End amplifier due to MOS devices thermal and flicker noise as well as shot noise introduced by the sensor leakage current. Assuming Gaussian distributed noise fluctuations superimposed to the nominal baseline, a threshold-to-noise ratio of about 7 for the minimum signal of interest leads to a noise occupancy below 10^{-6}, which is usually a good compromise between hit detection efficiency and fake rate [Spieler 2005, Rossi 2006]. With a minimum detectable charge of 1000 e$^-$, the upper limit for the Equivalent Noise Charge (ENC) referred to the input of the Front-End chain has been fixed to 150 e^- RMS for a 100 fF total input capacitance. Furthermore, pixel-to-pixel gain variations and discriminator threshold dispersions cannot degrade the overall detection performance. Hence a local fine adjustment of the threshold is required. Assuming again to deal with Gaussian effects, a specification for the RMS threshold dispersion after trimming σ_{th} can be obtained from ENC and minimum detectable charge requirements according to [Rossi 2006]

$$Q_{\mathrm{min}} \geq 5\sqrt{\mathrm{ENC}^2 + \sigma_{\mathrm{th}}^2}$$

As an upper limit, a residual threshold dispersion below 40 e^- RMS after tuning is required, whereas a maximum value for the threshold dispersion that can be tolerated before trimming has been set to 400 e^- RMS.

Additional design constraints and specifications are summarized in Table 2.1

Parameter	Spec requirement
fabrication technology	65 nm CMOS
hit rate	2 GHz/cm^2 in the innermost barrel layer
TID	1 Grad in 10 years
pixel size	50 μm $\times$ 50 μm
average event rate per pixel	50 kHz/pixel
maximum analogue area	25 μm $\times$ 50 μm
total power budget	10 μW/pixel
analogue supply voltage	1.2 V $\pm$ 60 mV DC variations
PSRR	$<$ 0 dB at all frequencies
analogue power consumption	$\leq$ 6 μW/pixel
hit time resolution	$<$ 25 ns
maximum input capacitance	150 fF (including strays)
nominal input capacitance	100 fF (including strays)
sensor signal polarity	negative input charges
nominal sensor leakage current	10 nA at highest irradiation levels
maximum sensor leakage current	20 nA at highest irradiation levels (worst case)
nominal operating temperature range	-30°C $<$ T $<$ 0°C
maximum operating temperature range	-30°C $<$ T $<$ 80°C
MIP deposited charge	4000 e$^-$ at highest irradiation levels
minimum detectable charge	1000 e$^-$
linear dynamic range	4 MIP (30 ke$^-$ before irradiation)
full dynamic range	10 MIP with fast recovery
charge digitization resolution	$\geq$ 5-bit in less than 400 ns
noise occupancy	$< 10^{-6}$
noise budget	ENC $<$ 150 e$^-$ RMS at 100 fF input capacitance
threshold dispersion after correction	$\sigma_{\text{th}} < 40$ e$^-$ RMS

Table 2.1: Input design specifications and constraints for the pixel Front-End chain [Re 2014].

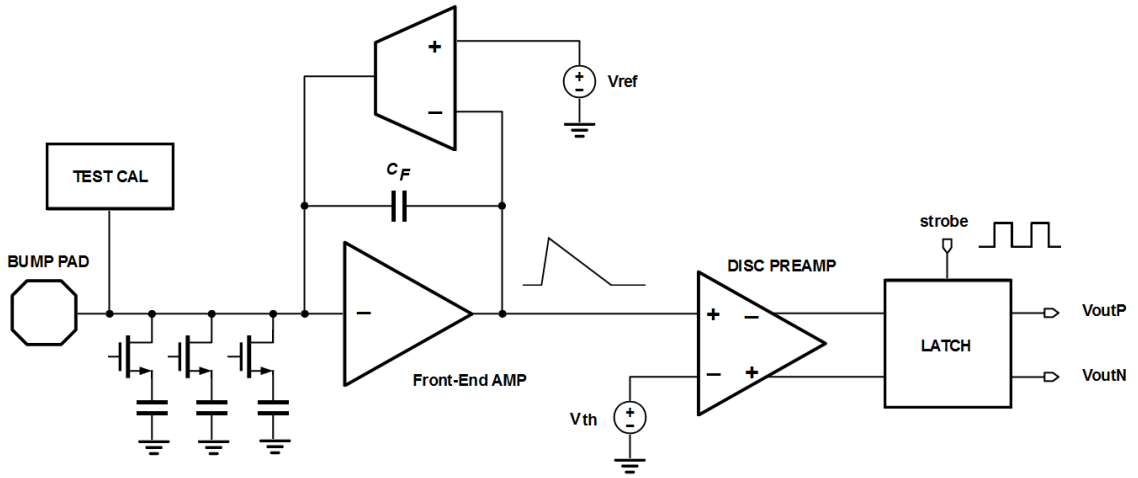


Figure 2.2: Simplified block diagram for the proposed analogue pixel Front-End chain architecture.

2.3 Analogue pixel cell architecture

A simplified block diagram of the proposed analogue pixel Front-End architecture is presented in Figure 2.2. Fundamental building blocks are briefly introduced hereafter, remanding to following sections an exhaustive and in-depth description of each block.
The input stage is a typical Charge Sensitive Amplifier (CSA) implemented as a single-ended, high open-loop gain inverting amplifier with capacitive feedback. Due to the severe power budget constraint defined for the analogue part, the usage of a dedicated filtering stage becomes quite unrealistic. As a result, a shaper-less architecture has been adopted, in which the CSA output drives directly the hit discriminator. Linear charge measurements are performed by means of the TOT technique, hence triangular shaping is adopted. A time-invariant active feedback network based on an auxiliary transconductance amplifier compensates leakage-induced baseline variations and discharges the feedback capacitance with a constant current after a charge signal has been detected. Signal amplification, pulse shaping and leakage current compensation are therefore accomplished in a compact way with a bare minimum amount of circuitry. For the sake of simplicity, CSA and feedback network stages will be referred to as Front-End amplifier in the following. A calibration circuit is used to inject a test charge at the CSA input node. Furthermore, test capacitors have been added to mimic different values of pixel sensor capacitance.
The most innovative component is the hit discriminator, implemented as a discrete-time voltage comparator. It uses a low gain differential amplifier coupled to a fast regenerative latch stage. An external clock signal is required to periodically enable/reset positive feedback in the latch, thus introducing synchronous Front-End operations. Depending on the comparator decision, two differential outputs settle to rail-to-rail complementary logic levels at each clock cycle and the hit generation becomes synchronized with the external clock. Local threshold adjustment is achieved by means of an autozeroed solution (not shown if figure), without the need of digital trimming using a local D/A converter. The comparator interfaces with a dedicated on-pixel digital control logic necessary to support latch operations as a local oscillator and perform high-speed TOT charge digitizations.

2.4 Front-End amplifier design

Charge sensitive amplifiers and shapers are essential components of any pulse processing chain for radiation detection. Given the importance and widespread use of these circuits, extensive theoretical analysis and analytical calculations can be found in reference literature through different detector and electronic technologies over the years [Radeka 1964, Kowalski 1970, Gatti 1986, Leo 1994, Spieler 2005, Rossi 2006, Iniewsky 2011]. Furthermore, a reach variety of CMOS pixel analogue Front-End solutions has been proposed in the last decades.
Indeed, only necessary basic concepts and mathematics applicable to the proposed architecture are reviewed, dedicating most of the attention to practical challenges and design choices related to the transistor level implementation of a Front-End amplifier in a 65 nm technology which targets previously discussed specifications. In the following, design and optimization of core charge amplifier, feedback network and test charge injection circuits are presented. Large design efforts and simulation time have been spent in the overall optimization of the Front-End amplifier in terms of minimum noise performance and closed-loop stability. A description of the discriminator is given is Section 2.5 instead, highlighting benefits and innovative aspects introduced by the usage of a discrete-time comparator.

Charge sensitive amplifier

At the core of the analogue Front-End resides a charge sensitive amplifier (CSA). As depicted in Figure 2.3, it uses a single-ended, high open-loop gain inverting amplifier with a capacitor C_F in feedback loop. For simulation purposes and analytical calculations, the pixel detector is modeled with a lumped capacitance C_D connected at the input node and referred to a small signal ground. The effective capacitance value incorporates all pixel capacitance contributions, including neighbours and bump bond parasitics. A shunt constant current source I_{leak} in parallel with C_D models the sensor leakage current. Finally, a time-dependent current pulse generator $i_{in}(t)$ is used to mimic the current signal induced at sensor electrodes during the charge collection process, according to the Shockley-Ramo theorem [He 2001]. For next considerations it is also assumed that a simple high-value noise-less resistor R_F connected in parallel to C_F provides proper DC path for biasing the input transistor of the core amplifier without contributing to the signal processing. In the final implementation input bias and capacitance discharge are provided by the feedback network. As discussed later in the chapter, for large enough input charges the feedback circuit provides a constant current discharge of the feedback capacitance, suitable for amplitude measurements with the time-over-threshold technique. For small input charges instead the feedback circuit behaves as a small-signal resistance R_F and the effects of its finite value on the signal processing will be reviewed. From design specifications, the foreseen usage of n-type sensors for the upgrade allows the Front-End to be optimized for negative only input charges (collection of electrons). A nominal total input capacitance of 100 fF is assumed for the overall CSA noise optimization, with an upper limit set to 150 fF and a noise budget ENC $< 150\ e^-$ at 100 fF. Furthermore, worst-case leakage currents up to 20 nA sinking the input node must be tolerated without compromising on noise performance and circuit functionality.

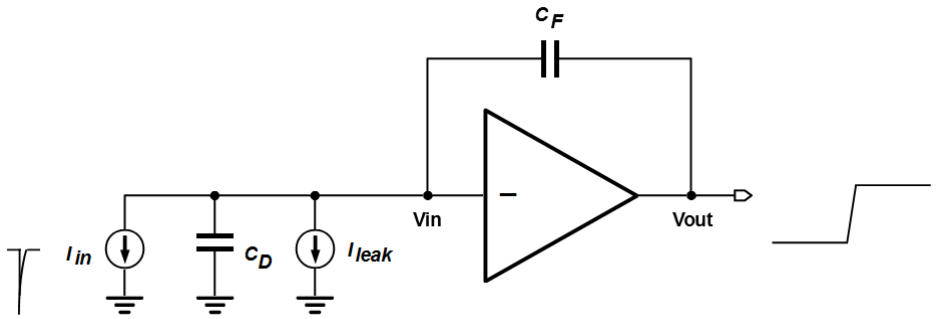


Figure 2.3: Sensor electrical equivalent model and CSA stage. A high value noise-less resistor is supposed to be connected in parallel with C_F in order to define the necessary input DC bias point without contributing to the signal processing.

As a first step, an estimate for the feedback capacitance and of the minimum required open-loop gain must be derived for practical design of the stage. At low frequencies, the core amplifier has a DC voltage gain $-A_o$ and $V_{out} = -A_o V_{in}$. Due to the finite value of A_o only a fraction of the total input charge

$$Q_{in} \;=\; \int_0^{\infty} dt\; i_{in}(t)$$

deposited in the sensor is integrated by the feedback capacitance C_F. The remaining component is lost in the detector instead, being collected over C_D. Neglecting the leakage current in fact, the Kirchoff current law (KCL) at the input node in the Laplace domain gives

$$I_{in}(s) + sC_D V_{in}(s) + sC_F \left[V_{in}(s) - V_{out}(s)\right] \;=\; 0$$

which in turn leads to the transfer function

$$\frac{V_{out}(s)}{I_{in}(s)} \;=\; \frac{1}{s}\left[\frac{A_o}{(1+A_o)\,C_F + C_D}\right]$$

As a result, in the time domain the output voltage is

$$V_{out}(t) \;=\; \frac{A_o}{(1+A_o)\,C_F + C_D}\int_0^{t} dt'\; i_{in}(t')$$

Assuming a δ-like input current pulse $i_{in}(t) = Q_{in}\delta(t)$ and neglecting for the moment the finite rise time due to the limited bandwidth of the core amplifier, the output signal is a voltage step with amplitude

$$\Delta V_{out} \;=\; -\frac{Q_{in}\;A_o}{(1+A_o)\,C_F + C_D} \;=\; -\frac{Q_{in}}{C_F}\left[1+\frac{1}{A_o}+\frac{C_D}{A_o C_F}\right]^{-1}$$

For an infinite gain amplifier, the formula reduces to the well-known result $-Q_{in}/C_F$ valid for an ideal integrator. In the term $(1 + A_o)\;C_F$ we can immediately recognize the contribution of the feedback capacitance referred to the input node due to Miller effect. Hence the effective total input capacitance is given by the parallel combination between the detector capacitance C_D and the Miller equivalent of C_F,

$$C_{in} \;=\; C_D + (1+A_o)\;C_F$$

whereas the input voltage V_{in} rearranged in terms of all other parameters becomes

$$V_{in} \;=\; \frac{Q_{in}}{C_{in}} \;=\; \frac{Q_{in}}{(1+A_o)\;C_F + C_D}$$

As expected, provided that A_o is sufficiently large $V_{in} \approx 0$ and the input terminal can be assumed as a virtual ground, resulting into a very low impedance node. This is a key requirement in order to minimize crosstalk currents due to charge injection effects through stray capacitances between adjacent pixels. The charge collection efficiency Q_F/Q_{in} represents the fraction of the total charge Q_{in} deposited in the sensor and the effective amount of charge integrated over the Miller feedback capacitance, $Q_F = [(1 + A_o)\;C_F]\,V_{in}$. Thus we can write

$$\frac{Q_F}{Q_{in}} \;=\; \frac{(1+A_o)\;C_F}{(1+A_o)\;C_F + C_D}$$

while a charge fraction $Q_D/Q_{in} = 1 - Q_F/Q_{in}$ is lost in the sensor. For a high charge collection efficiency the relationship $(1 + A_o)\;C_F \gg C_D$ must be guaranteed. Nevertheless, a large feedback capacitance value is not desirable since it would decrease the charge-to-voltage conversion gain, which in turn would result into weak output signals more susceptible to parasitics and crosstalk effects. Moreover, large capacitances necessary require significant pixel area. Assuming 100 fF nominal input capacitance, realistic feedback capacitance values lie the 2-15 fF range and can be implemented using precise Metal-Oxide-Metal (MOM) or Metal-Insulator-Metal (MIM) capacitors offered by the 65 nm library and easily integrable on a pixel cell without large area contributions. Indeed, the open-loop gain must be maximized, introducing therefore a trade-off among gain, noise and power dissipation constraints.

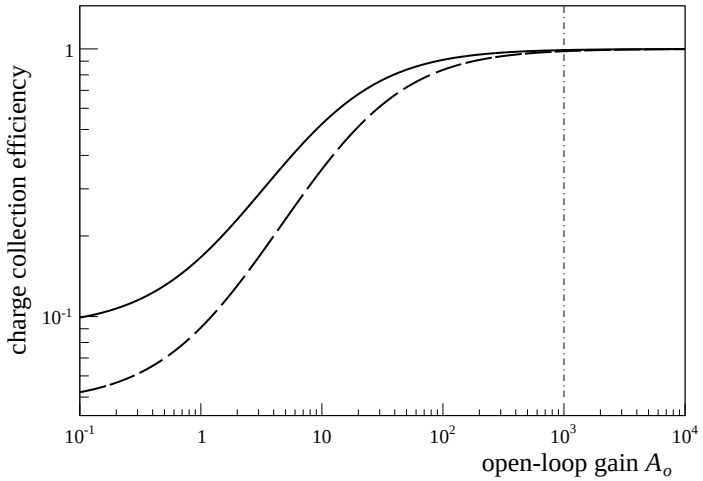


Figure 2.4: Charge collection efficiency Q_F/Q_{in} as a function of the CSA open-loop gain A_o assuming $C_D = 100$ fF (solid curve) of 200 fF (dashed curve) for 10 fF feedback capacitance C_F. For an efficient and sensor-independent readout $A_o \approx 10^3$ or above is required.

The plot in Figure 2.4 shows the Q_F/Q_{in} ratio as a function of the core amplifier open-loop gain for different values of detector capacitance C_D and assuming $C_F = 10$ fF. Thus, efficient and sensor-independent charge integrations can be obtained if $A_o \approx 10^3$ (60 dB) is satisfied. Note that previous formulas apply also when the open-loop gain drops due to amplifier saturation. If the CSA output exceeds the linear voltage swing, the charge collection efficiency degrades and the input node can not be considered anymore as a virtual ground. A significant amount of charge is shared between C_F and C_D and the input node becomes more susceptible to crosstalk. When $A_0 = C_D/C_F - 1$ the efficiency drops to 0.5 and the input charge is equally collected on feedback and detector capacitances.

In order to achieve a high voltage gain along with low-noise performance in a single stage, cascode amplifiers are the natural choice [Laker 1994, Johns 1996, Razavi 2000, Gray 2001, Allen 2002, Sansen 2006]. They represent a standard solution for the design of the input stage in most of CMOS Front-End systems. A transistor level schematic of the proposed charge sensitive amplifier is presented in Figure 2.5. It is a low-power and reliable architecture that has been extensively adopted also in other designs of the INFN Torino VLSI Design Laboratory and already implemented in higher technology nodes [Delaurenti 2006, Martoiu 2006, Kugathasan 2010, Di Pietro 2012].
The input branch uses a telescopic cascode M1-M4 with gain enhancement obtained by means of the auxiliary PMOS cascode current source M5-M6. Furthermore, to ensure proper impedance matching and necessary drive strength a voltage buffer M7-M8 implemented with a NMOS source follower is included. Due to asymmetric voltage swings in source followers, the usage of a NMOS configuration is optimized for the readout of n-type sensors. Finally, NMOS switches M9-M10 have been added such that three different feedback capacitance values $C_F = 2.5$ fF, 4 fF or 6.5 fF can be selected using a couple of configuration bits SEL_C2F and SEL_C4F and two integrating capacitors connected in parallel. This maximizes the testability in the first prototyping phase.

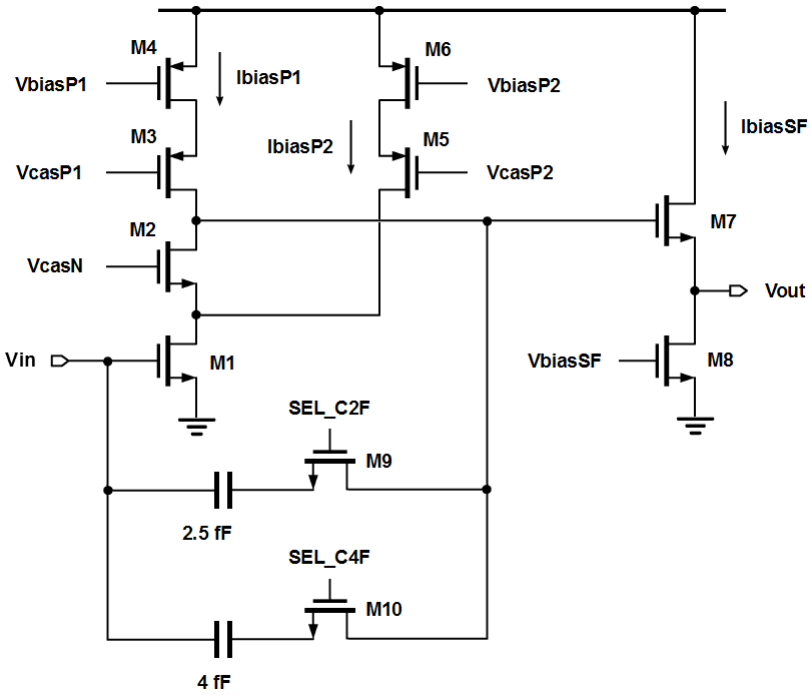


Figure 2.5: Charge sensitive amplifier transistor level implementation. For practical device sizing and optimization a high value noise-less resistor R_F is supposed to be connected between input and output nodes in order to define the necessary DC bias point for the input transistor without contributing to the signal processing. All NMOS devices have been implemented in Deep-N-Well (DNW) configuration to ensure substrate isolation and low-noise performance. The actual feedback network is discussed later in the chapter.

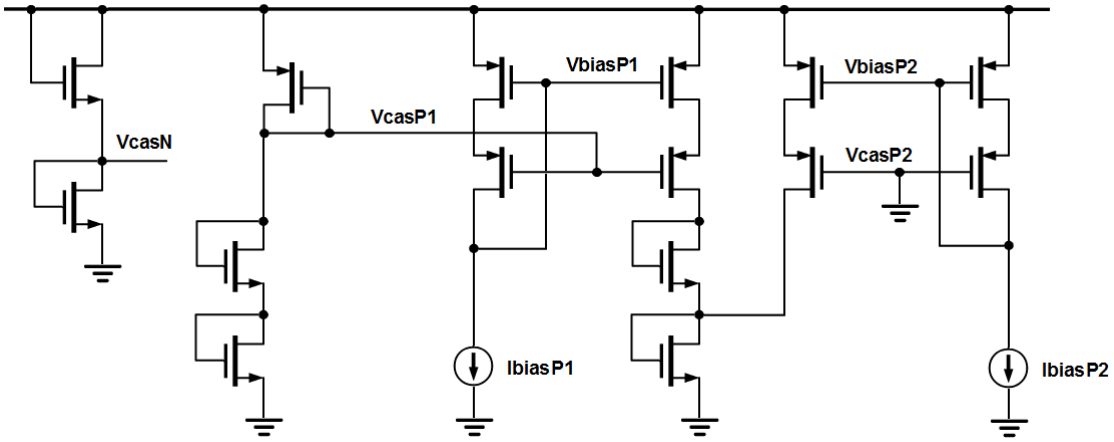


Figure 2.6: External bias network to generate cascode voltages and bias currents. MOS bypass capacitors added to filter the power supply rail are not shown.

As shown if Figure 2.6, replica wide-swing cascode current mirrors have been used to bias PMOS current source loads M3-M4 and M5-M6, whereas steady voltages for cascode devices are provided by simple supply voltage dividers using diode-connected MOS transistors. Note that in the final optimized design the cascode voltage $VcasP2$ has been tied to ground. The bias for the source follower is provided by a simple current mirror instead. Similar considerations apply also in the design of the feedback network and of the voltage comparator. If not otherwise specified, it is assumed that all bias currents are generated with current mirror techniques and optimized cascode voltages have been fixed with MOS voltage dividers. However, in the final chip all bias cells have been placed at the periphery, thus they not contribute to the compute of the on-pixel static power dissipation. The usage of dedicated PVT-independent references and biasing D/A converters will be addressed in a second iteration.

It is a matter of fact that biasing and preliminary device sizing of the circuit are essentially driven by the tight power budget constraint defined for the analogue pixel cell, which in practice limits the choice of bias currents to a few μA per stage. As already discussed, for a realistic and conservative design no more than 2.5-3 μA DC currents can be allocated in the overall Front-End amplifier and no more than 1-1.5 μA for the hit discriminator in order to keep the total analogue power consumption $\leq$ 6 μW/pixel at 1.2 V supply voltage. Actually, the nominal current demanded for the feedback network is low, below 100 nA. Most of the current can be therefore used to maximize the open-loop gain in the core amplifier. A total supply current of 2 μA has been selected for the input device M1, whereas additional 500 nA flowing in the output source follower are needed to ensure adequate drive strength capabilities. With a total DC current of 2.5 μA the static power dissipation in the charge sensitive amplifier stage settles to 3 μW, one half the maximum analogue power budget per pixel. The actual current fractions $IbiasP1$ and $IbiasP2$ provided by PMOS cascode loads will be discussed shortly thereafter in the text.

For a large gain, the transconductance g_{m1} of the input device must be maximized. The choice of the total bias current $I_{D1} = 2\ \mu$A fixes the gate-source voltage V_{GS1} (hence the DC level of the baseline) required to accept such a current according to the device aspect ratio $(W/L)_1$. In order to mitigate the degradation of the single-transistor intrinsic-gain g_m/g_{ds} due to technology scaling and short channel effects (SCE), the minimum channel length has been increased and a nominal design length $L \approx 3L_{min}$ has been adopted [Razavi 2000, Sansen 2006]. As shown in plot of Figure 2.7, the simulated transconductance-to-current ratio g_m/I_D for a diode-connected NMOS device[1] and realistic sizing for the input transistor suggests that M1 is biased at the centre of the moderate inversion region if a drain current of 2 μA is adopted [Tsividis 1999, Enz 2006, Jespers 2010]. The input device transconductance g_{m1} is therefore mainly determined by the bias current and exhibits only weak dependences on transistor dimensions, as shown in Figure 2.8. This can be better appreciated in the plot of Figure 2.9, reporting transconductance values as a function of the channel width W for a fixed 2 μA bias current and $L = 200$ nm. The usage of a wide input device does not increase significantly its g_{m1} but only contributes to increase the total gate capacitance, as shown in Figure 2.10. Assuming $(W/L)_1 = 50$, a bias current of 2 μA leads to $g_{m1} \approx 50\ \mu$S and requires $V_{GS1} \approx 360$ mV, with an underdrive voltage[2] of about 150 mV.
The input transconductance is basically fixed by the low-power requirement. The maximum gain achievable at low frequencies relies therefore on the value of the total small signal resistance R_o at the cascode amplifier output node. Indeed, thanks to current splitting a fraction $IbiasP2 = x\, I_{D1}$ of the total bias current that flows in M1 can be supplied by the additional PMOS current source M5-M6, thus reducing the current flowing in the left branch to $IbiasP2 = (1 - x)\ I_{D1}$ without changing the bias current I_{D1} chosen for the input device. As a result, a higher output resistance can be obtained, which in turn increases the gain. A common choice for the fraction x is in the 0.6-0.8 range. In the final optimized circuit $x = 0.75$, thereby a bias current $IbiasP1 = 500$ nA is mirrored in the main input branch M2-M4 while a current $IbiasP2 = 1.5\ \mu$A is injected at the drain of the input transistor by the auxiliary PMOS current source M5-M6. Note that PMOS loads are biased with two totally independent current mirrors, maximizing devices optimization.

An analytical expression for the DC voltage gain in terms of typical MOS transistor small signal parameters such as transconductance g_m, body-effect transconductance g_{mb} and output resistance $r_{ds} = 1/g_{ds}$ can be derived by applying cascode formulas to the circuit [Kugathasan 2010].
Referring to the simplified schematic of Figure 2.11, small signal resistances of PMOS cascode current source loads are given by

$$R_{casP1} = r_{ds3} + r_{ds4} + (g_{m3} + g_{mb3})\ r_{ds3}\ r_{ds4} \approx g_{m3}\ r_{ds3}\ r_{ds4}$$

$$R_{casP2} = r_{ds5} + r_{ds6} + (g_{m5} + g_{mb5})\ r_{ds5}\ r_{ds6} \approx g_{m5}\ r_{ds5}\ r_{ds6}$$

Since the small signal resistance R_{casP2} of the auxiliary branch goes in parallel with the output resistance r_{ds1} of the input device, the effective impedance of the NMOS cascode M1-M2 is

$$R_{casN} = [r_{ds1}||R_{casP2}] + r_{ds2} + (g_{m5} + g_{mb5})\ [r_{ds1}||R_{casP2}]\ r_{ds2}$$

[1] Since 65 nm CMOS was never used before in Torino, validation of the full-custom design flow and technology characterization in terms of SPICE simulation models provided by the foundry has been an essential task in the preliminary design phase. Basic studies performed with simple single-transistor testbenches extracted the most important large-signal characteristics and small-signal parameters (transconductance, g_m/I_D, output resistance, gate capacitance etc.) that are fundamental quantities and figures of merit for analogue integrated circuit design.

[2] For MOS devices in subthreshold operations, the overdrive voltage $V_{ov} = V_{GS} - V_{TH}$ is a negative quantity.

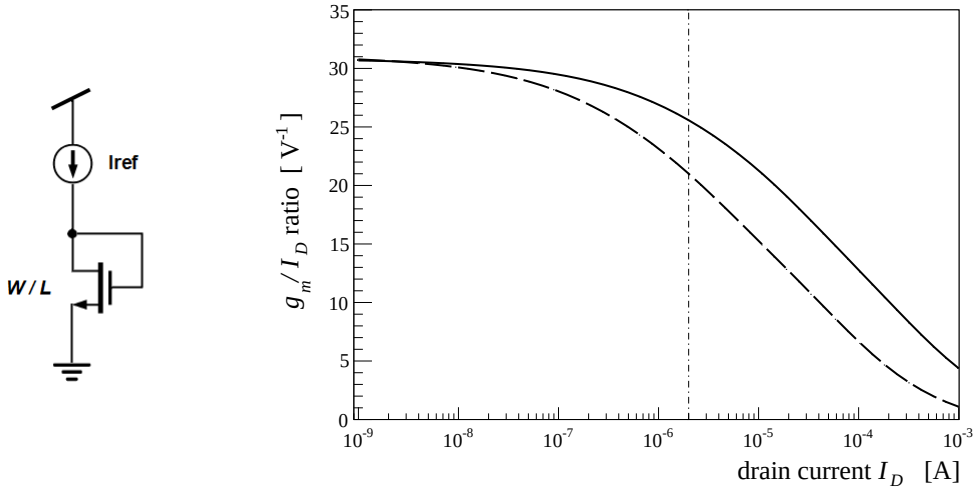


Figure 2.7: Simulated g_m/I_D ratio as a function of the bias current for a NMOS diode-connected device with $W/L = 50$ (solid) and $W/L = 10$ (dashed) assuming $L = 200$ nm. With a nominal bias current of 2 μA and realistic transistor sizing the input device is expected to work in moderate inversion.

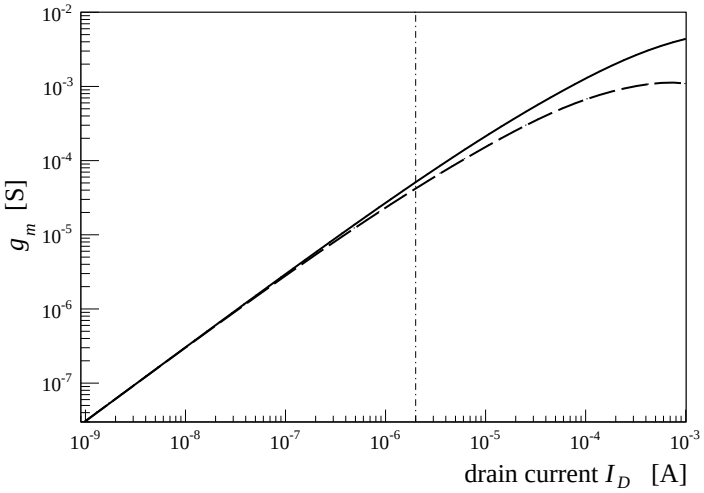


Figure 2.8: Device transconductance as a function of the bias current from the same testbench. The limited power budget basically determines a maximum achievable $g_{m1} \approx 50$ μS for the input device, without strong dependence on MOS transistor channel width W in moderate inversion.

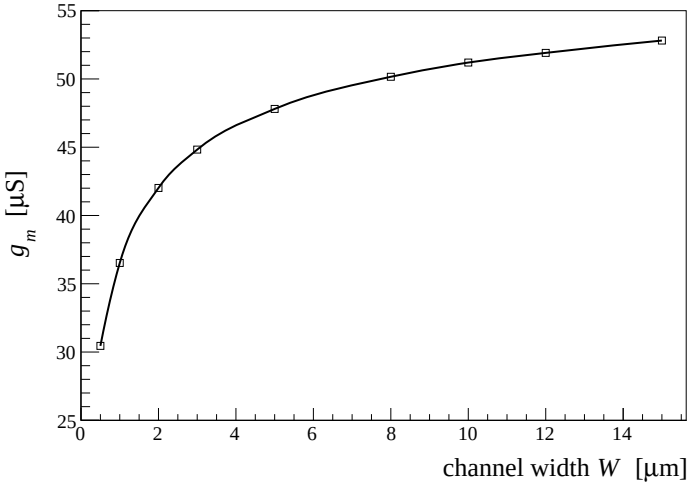


Figure 2.9: Device transconductance g_m as a function of the channel width W for $L = 200$ nm and assuming 2 μA bias current.

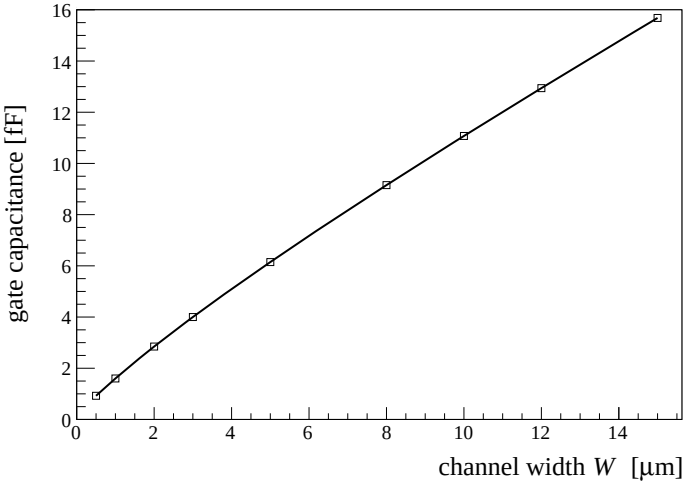


Figure 2.10: Total gate capacitance as a function of the channel width W for $L = 200$ nm and 2 μA bias current. The choice of a too large input device ($W > 10$ μm) does not significantly increases the gain of the input stage, but only results into a higher input capacitance.

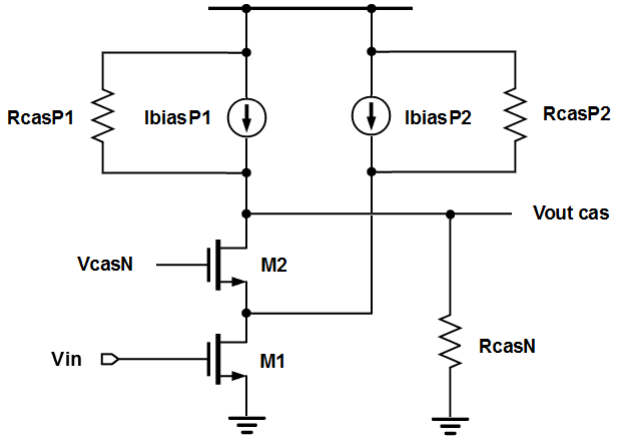


Figure 2.11: Small signal resistances that contribute in determining the total output impedance of the telescopic cascode amplifier.

The total impedance seen at the cascode amplifier output node is therefore

$$R_o \;=\; R_{casN} || R_{casP1}$$

Finally, the equivalent transconductance G_m of the NMOS input cascode roughly equals g_{m1} of the input device,

$$G_m \;=\; g_{m1} \left\{ \frac{[r_{ds1} || R_{casP2}] + (g_{m2} + g_{mb2})\; [r_{ds1} || R_{casP2}]\; r_{ds2}}{[r_{ds1} || R_{casP2}] + r_{ds2} + (g_{m2} + g_{mb2})\; [r_{ds1} || R_{casP2}]\; r_{ds2}} \right\} \;\approx\; g_{m1}$$

As a result, the small signal DC open-loop gain of the overall cascode stage is

$$|A_o| \;=\; G_m\; R_o \;\approx\; g_{m1}\; [R_{casN} || R_{casP1}]$$

The simulated AC open-loop gain and phase at the cascode amplifier output node for the final optimized circuit are presented in Figure 2.12. Figure 2.13 shows instead the small signal resistance at the same node. As one can see, at low frequencies the voltage gain of the cascode stage sets to 10^3, with 2 MHz bandwidth (BW) and 2 GHz GBW product. The effective value at the CSA output node *Vout* is slightly reduced due to the presence of the output source follower. The small signal resistance is 23 MΩ, probed with an AC current source connected at the cascode output node. Simulated values are in good agreement with small signal model parameters provided by DC operating point analysis.

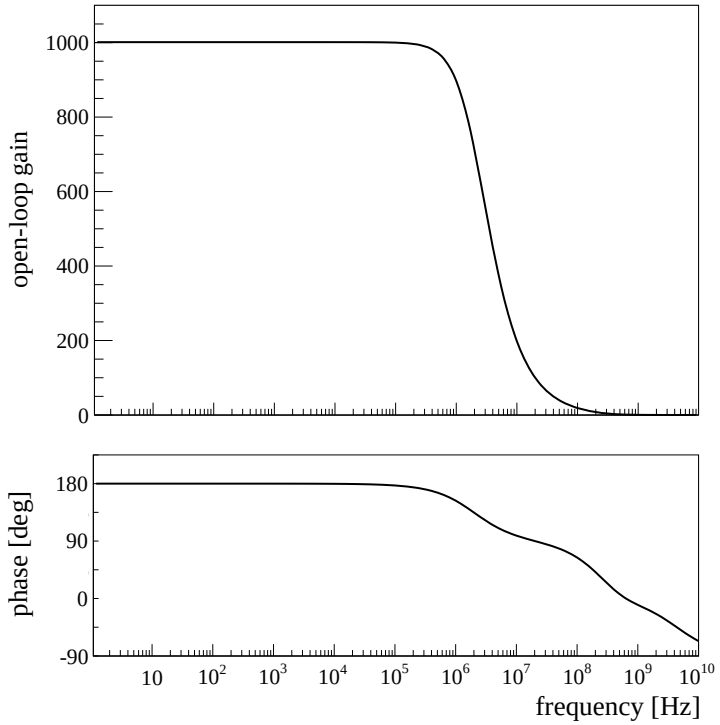


Figure 2.12: Simulated open-loop gain and phase as a function of frequency for the input cascode amplifier.

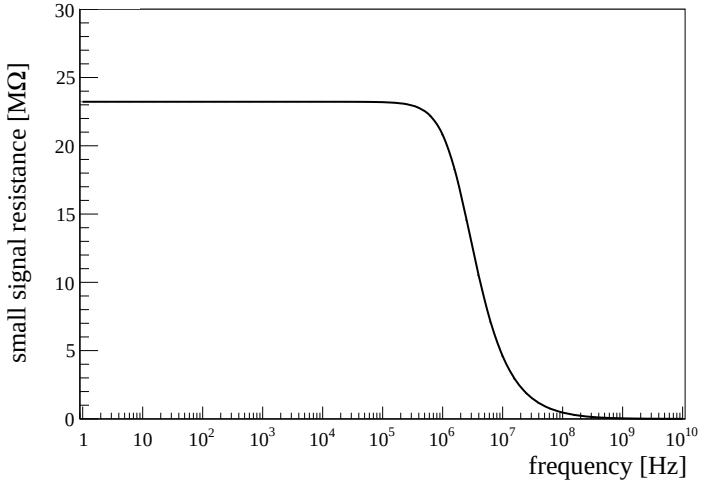


Figure 2.13: Total cascode output impedance as a function of the frequency.

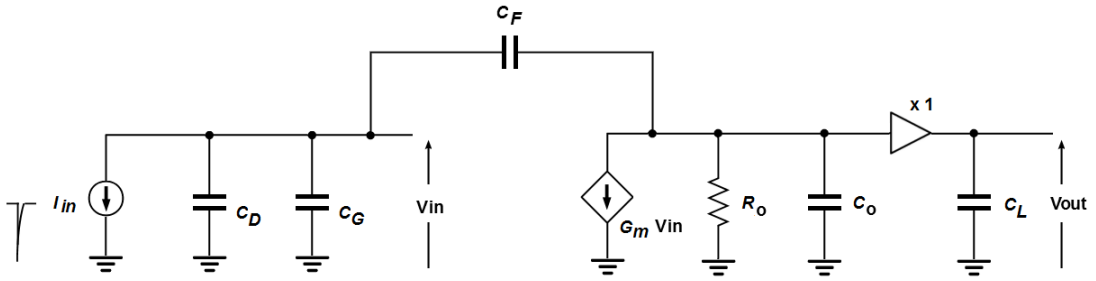


Figure 2.14: CSA small signal equivalent model. The lumped capacitor C_G includes all extrinsic capacitances seen at the gate of the input transistor. A single-pole frequency response is assumed for the core amplifier and the output source follower is modelled as a unity-gain buffer. This stage isolates the system from any additional extra capacitive load C_L due to other elements connected to the Front-End amplifier.

Due to the limited bandwidth of the core amplifier, the CSA output voltage has a non-zero signal rise time. This introduces fundamental considerations for the choice of the feedback capacitance value. As already discussed in fact, in order to properly assign a time stamp flag for an event at the LHC, the binary hit information provided by the Front-End discriminator must be available in less then 25 ns. In practice, this constraints the peaking time of the Front-End amplifier to be less than 25 ns in order to prevent time-walk issues in the time stamp assignment. Hence fundamental CSA parameters in determining the signal rise time must be considered.
A small signal equivalent model for the charge integrator is depicted in Figure 2.14. The additional lumped capacitor C_G ad the input node includes all extrinsic capacitances seen at the gate of the input transistor, mainly due to the gate-source capacitance $C_{GS} \approx C'_{ox}WL$. Thanks to the high impedance at the cascode output node, a single-pole frequency behaviour is a good approximation for the core amplifier, thus

$$A(s) = \frac{A_o}{1 + s/\omega_o} \qquad \text{and} \qquad \omega_o = \frac{1}{2\pi R_o C_o}$$

According to simulated values for BW and R_o, an output capacitance $C_o \approx 3.5$ fF is derived. An analytical expression for the signal rise time can be derived from the impulse response of the CSA small signal circuit in the time domain. By writing the KCL at both input and output nodes in fact we obtain the linear system

$$\begin{cases} I_{in}(s) + s(C_D + C_G)V_{in}(s) + sC_F\left[V_{in}(s) - V_{out}(s)\right] = 0 \\ \\ G_m V_{in}(s) + V_{out}(s)/R_o + sC_o V_{out}(s) + sC_F\left[V_{out}(s) - V_{in}(s)\right] = 0 \end{cases}$$

which can be solved to get the transfer function $H(s) = V_{out}(s)/I_{in}(s)$ in the Laplace domain. Despite circuit simplicity, the presence of a non-zero output capacitance immediately leads to more elaborate nodal analysis calculations [Peric 2004, Karagounis 2011].

Beside algebraic computations, due to the large DC open-loop gain $A_o = G_m R_o \gg 1$ the resulting transfer function simplifies to

$$\frac{V_{out}(s)}{I_{in}(s)} \approx \frac{1}{sC_F}\left[\frac{1 - sC_F/G_m}{1 + s\tau_r}\right]$$

where

$$\tau_r = \frac{1}{G_m}\left[(C_D + C_G + C_o) + \frac{(C_D + C_G)\, C_o}{C_F}\right]$$

is the time constant associated to the second pole. The transfer function exhibits therefore two poles and one zero. The zero at frequency G_m/C_F is a direct consequence of the coupling between input and output nodes through to the feedback capacitance. In fact, the fraction of input signal that goes through the feedback path is not inverted, thus resulting into a positive contribution in the numerator. Nevertheless, according to realistic values proposed for the circuit under analysis, $G_m \approx g_{m1} \approx 50\ \mu$S and $C_F \approx 10$ fF, the zero is actually placed at very high frequency, in the GHz range. Neglecting such a zero, the transfer function can be approximated as

$$\frac{V_{out}(s)}{I_{in}(s)} \approx \frac{1}{sC_F}\left[\frac{1}{1 + s\tau_r}\right]$$

As a result, in the time domain the output voltage for a δ-like input current $i_{in}(t) = Q_{in}\delta(t)$ reduces to a simple first-order low-pass filter step response,

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

Hence the *rise time* of the signal evaluated between 10% and 90% of its maximum value Q_{in}/C_F is given by $t_r = 2.2\ \tau_r$. The CSA rise time exhibits therefore dependences with all small-signal model parameters except the output resistance R_o, provided that a high gain at low frequencies is guaranteed. On the one hand, t_r strongly depends on the transconductance $G_m \approx g_{m1}$ of the input stage. A large value of G_m results into a fast signal rise time at the cost of a much higher power dissipation. On the other hand, both the total input capacitance $C_D + C_G$ and the load capacitance at the cascode output node C_o contribute in reducing the speed. Indeed, the effect of the feedback capacitance C_F really depends on the chosen value compared to other capacitance magnitudes, with a dominant contribution $\sim 1/C_F$ in case $C_F \ll C_o$. The feedback capacitance fixes the nominal charge-to-voltage gain of the CSA, with a maximum value Q_{in}/C_F for an ideal integrator. A large feedback capacitance reduces the signal rise time but diminish the gain as well, which is not desirable in order to avoid weak signals more susceptible to parasitics and crosstalk effects. A careful choice of the feedback capacitance C_F is therefore essential and has to take into account all input specifications and design trade-offs between sensor capacitance, minimum detectable charge, charge-to-voltage gain, power dissipation, signal rise time and available area.

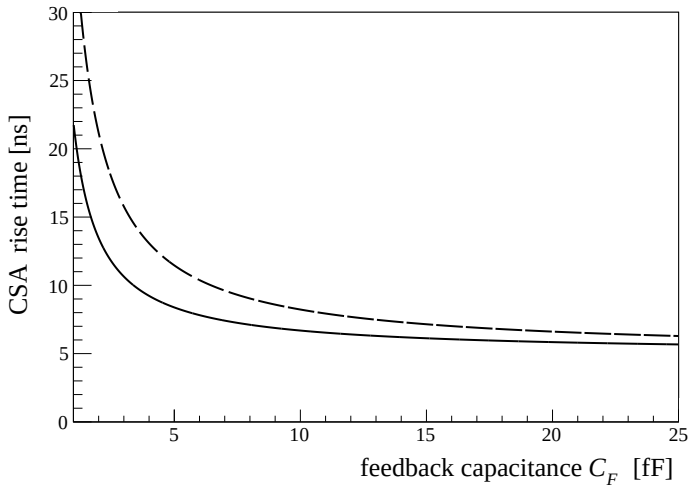


Figure 2.15: CSA rise time $t_r = 2.2\,\tau_r$ (single-pole approximation) as a function of the feedback capacitance C_F for 100 fF (solid) and 200 fF (dashed) detector capacitances. Small-signal parameters $G_m = 50\ \mu$S, $C_G = 10$ fF and $C_o = 3.5$ fF have been assumed.

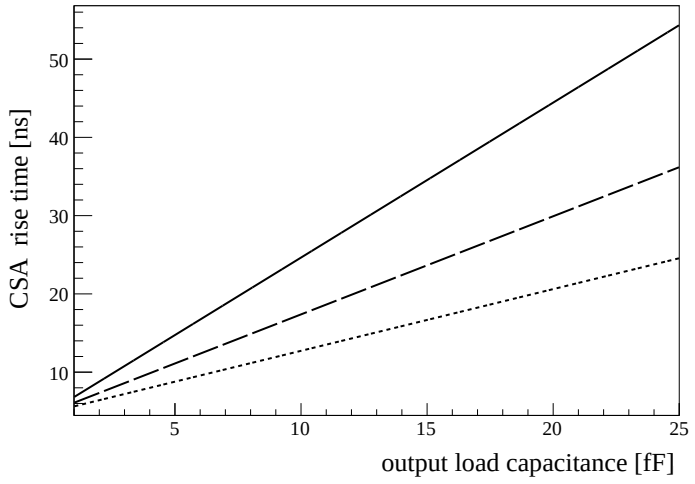


Figure 2.16: CSA rise time t_r as a function of the cascode output capacitance C_o for $C_F = 2.5$ fF (solid) 4 fF (dashed) and 6.5 fF (dotted) feedback capacitances. The usage of an output buffer is mandatory.

As already stressed, due to the severe power budget constraint defined for the analogue pixel cell the maximum available transconductance g_{m1} for the input transistor is essentially determined by the total 2 μA supply current allocated in the cascode stage. The nominal detector capacitance is 100 fF, whereas as derived from simulations the total gate capacitance for a realistic input device with $W/L \approx 50$ is of the order of 10 fF. Figure 2.15 presents the analytical expression of the rise time as a function of C_F for two different values of detector capacitance $C_D = 100$ fF and 200fF, assuming small signal parameters $G_m = 50$ μS, $C_G = 10$ fF and 3.5 fF load capacitance. As one can see, for $C_F > 10$ fF the rise time does not depend significantly on C_F, settling to about 6-8 ns. At smaller feedback capacitance values instead, the contribution $1/C_F$ dominates and the rise time rapidly increases. Assuming traditional continuous-time operations, for a reliable in-time response below 25 ns a conservative choice for the maximum peaking time of the signal fed to the Front-End discriminator is usually of the order of 15 ns or below. This is necessary since the hit discriminator has a non-zero delay that must be taken into account. Indeed, more relaxed peaking time constraints can be assumed in perspective of synchronous operations. By adopting a discrete-time comparator in fact, the nominal peaking time of the Front-End amplifier can be assumed to be 12.5 ns, half the bunch crossing. Nevertheless, since the binary hit information is always synchronized with an external 40 MHz master clock, peaking time delays of the order of 5-10 ns around such a nominal value do not affect the actual response of the discriminator, provided that the signal is found above the threshold. As a result, smaller values of the feedback capacitance can be adopted, thus maximizing the charge-to-voltage gain without compromising the in-time response demanded for a pixel system at LHC. Certainly a reduced closed-loop stability must be considered. As shown in the plot, a feedback capacitance $C_F = 4$ fF leads to a signal rise time $t_r \approx 10$ ns for 100 fF detector capacitance, with an ideal charge-to-voltage gain $1/C_F = 250$ mV/fC. This represents a good compromise between gain and speed. However, as already anticipated the CSA has been equipped with two selectable feedback capacitors of 2.5 fF and 4 fF, resulting into a flexible solution for the the first prototyping phase.
Benefits and necessity of an output buffer can be appreciated instead by inspecting the CSA rise time as a function of the load capacitance, as presented in Figure 2.16. According to analytical calculations, t_r linearly increases with the output capacitance C_o. As a matter of fact, a few tens fF load capacitance are enough to cause totally unrealistic values for the signal rise time. Due to the high impedance seen at the cascode output node any small variations in the value of the load capacitance introduce large differences in the open-loop bandwidth of the cascode, which in turn would result into a completely unreliable amplifier. When connected to the feedback network and to the Front-End discriminator the CSA output node must be able to drive a significant amount of load capacitance, up to 100 fF or higher, including layout parasitics from metal interconnections. The usage of an additional buffer stage is therefore mandatory to ensure necessary drive strength and protect the high impedance internal node from the remaining circuitry. This justifies the choice of a source follower output stage, at the cost of an extra 500 nA bias current in the output branch and a negligible attenuation of the signal amplitude. Since the signal swing in source followers is asymmetric, the usage of a NMOS configuration is optimized for positive swings of the output signal with respect to the baseline, hence for n-type sensors.

Transient simulations for the CSA impulse response assuming a nominal minimum detectable input charge of 1 ke^- and different feedback capacitance values C_F are presented in Figure 2.17. For the same testbench, the rise time as a function of the selected feedback capacitance is reported in Figure 2.18. The impulse response for different detector capacitance values C_D is presented instead in Figure 2.19. The rise time as a function of the detector capacitance is shown in Figure 2.20. As expected, t_r decreases by increasing C_F and linearly increases with C_D.
Finally, the impulse response for different input charges is presented in Figure 2.21, whereas the charge-to-voltage characteristic is shown in Figure 2.42. As one can see, gain linearity is guaranteed only for a limited range of input charges, up to 12$-$13 ke^-. Nevertheless, a CSA saturation is not a limiting factor since actual linear charge measurements are obtained by means of the TOT technique.

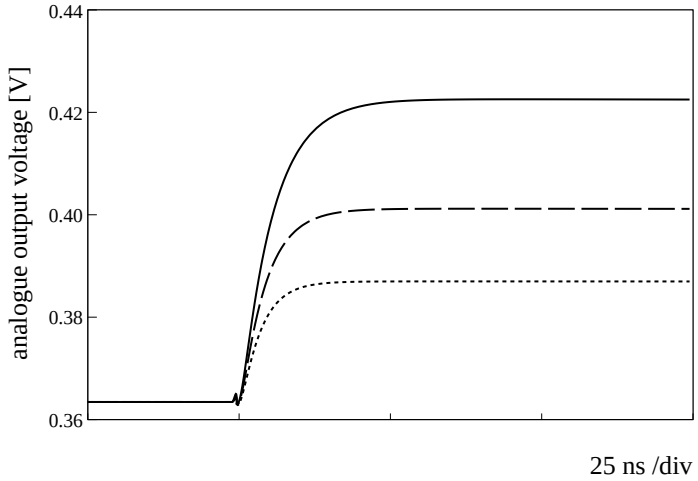


Figure 2.17: CSA impulse response for 2.5 fF (solid), 4 fF (dashed) and 6.5 fF (dotted) feedback capacitance, assuming 1 ke^- input charge and 100 fF detector capacitance.

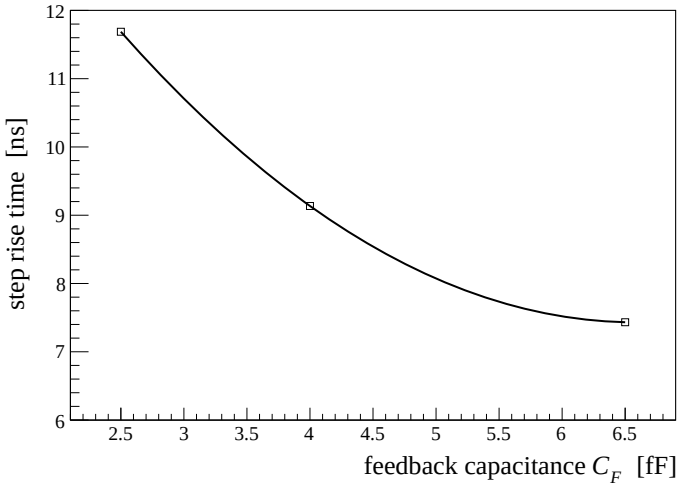


Figure 2.18: Step rise time as a function of the selected feedback capacitance.

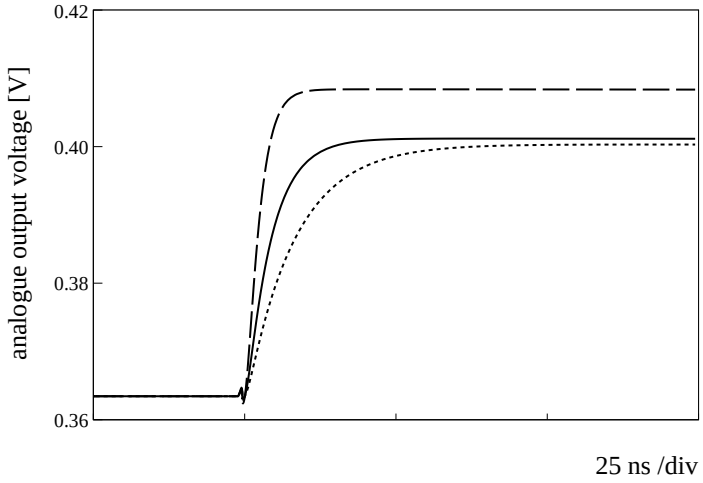


Figure 2.19: CSA impulse response for 50 fF (dashed) 100 fF (solid) and 200 fF (dotted) detector capacitance, assuming 1 ke^- input charge with 4 fF feedback capacitance.

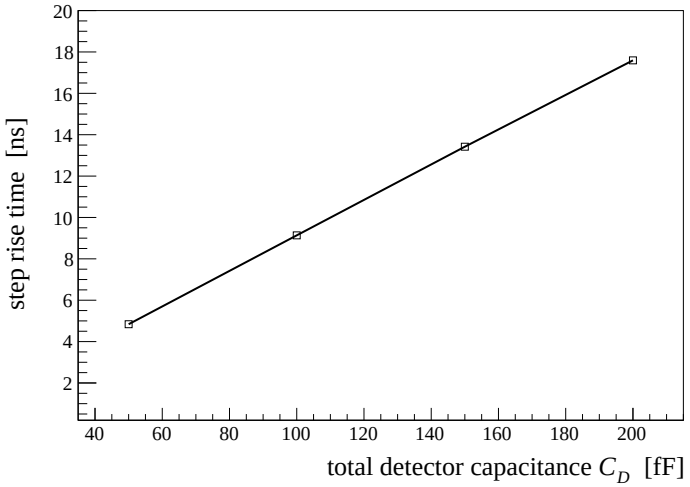


Figure 2.20: Step rise time as a function of the total detector capacitance C_D.

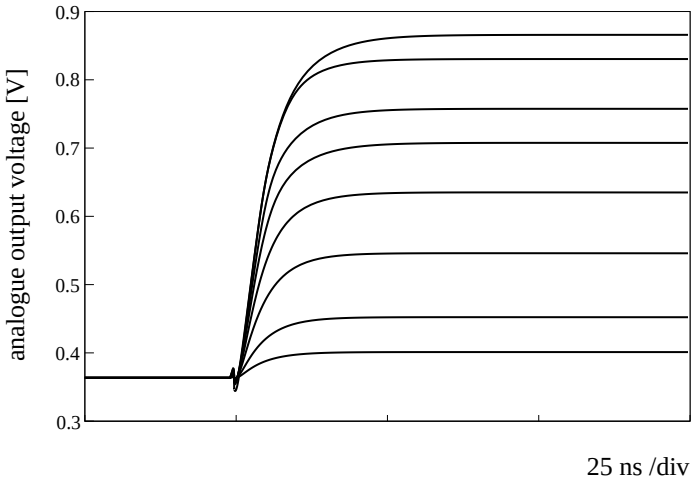


Figure 2.21: CSA impulse response for different input charges (1 ke^- – 30 ke^-) assuming 4 fF feedback capacitance and 100 fF detector capacitance.

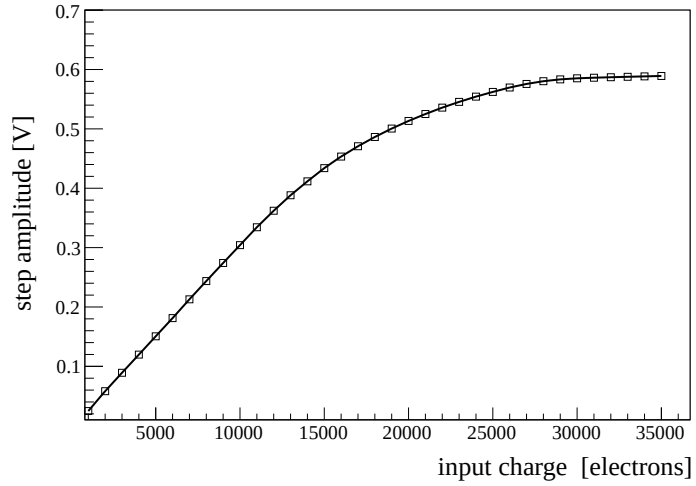


Figure 2.22: Step amplitude as a function of the input charge. Saturation of the core amplifier is not a limiting factor since actual linear charge measurements are obtained by means of the TOT technique.

After necessary circuit inspection analysis, practical transistor sizing strategy adopted for the core amplifier can be now derived from minimum noise requirements and matching considerations. As extensively demonstrated in literature in fact, in a properly designed charge sensitive amplifier the largest noise contribution is imputable to the input transistor, requiring a careful choice of its aspect ratio W/L [Sansen 1990, O'Connor 2002, De Geronimo 2005]. Noise performance are usually expressed in terms of the Equivalent Noise Charge (ENC), defined as the ratio between the total output-referred RMS noise voltage and the charge-to-voltage gain $A_Q = \Delta V_{out}/Q_{in}$ of the amplifier,

$$\mathrm{ENC} \;=\; \frac{\sqrt{< v_{n,out}^2 >}}{A_Q} \;=\; \frac{1}{A_Q}\left[\int_0^\infty d\,f\left(\frac{d < v_{n,out}^2 >}{d\,f}\right)\right]^{1/2}$$

being $d < v_{n,out}^2 > /d\,f$ the noise power spectral density (PSD) at the output node. The input charge Q_{in} is quoted as a number of electrons, hence the ENC equals the total output-referred RMS noise voltage divided by the signal amplitude for an input charge of a single electron. Such a definition is however quite cumbersome for practical ENC calculation during CAD simulations[3] and experimental measurements. For a given minimum detectable charge specification Q_{min} the signal-to-noise ratio (SNR) is Q_{min}/ENC. In order to keep low the noise occupancy, the SNR must be maximized in respect of low-power consumption requirements. As discussed, a minimum detectable charge of 1 ke^- has been defined for the pixel upgrade, requiring ENC < 150 e^- at 100 fF nominal detector capacitance. This ensures SNR > 6. Without doubts the choice of a shaper-less Front-End chain makes the ENC minimization more challenging. Since the maximum achievable charge-to-voltage gain for an ideal CSA (infinite open-loop gain) reduces to $1/C_F$, for a given ENC constraint the upper limit

$$\sqrt{< v_{n,out}^2 >} \;\leq\; \frac{\mathrm{ENC}}{C_F}$$

must be satisfied. Assuming a nominal value $C_F = 4$ fF, the above ENC constraint requires that noise-induced random baseline fluctuations (noise floor) must be kept below 6 mV RMS.
Different noise sources contribute in determining the total ENC value in a Front-End amplifier. On the one hand, MOS transistors exhibit frequency-independent (white) channel thermal noise and frequency-dependent flicker ($1/f$) noise [Johns 1996, Tsividis 1999, Razavi 2000]. On the other hand, the sensor leakage current introduces frequency-independent shot noise [Spieler 2005, Rossi 2006]. As already mentioned, in the core amplifier design it is assumed that a noise-less feedback resistor is used to properly set the DC operating of the input transistor. Thermal noise introduced by the active feedback network will be considered later. For statistically uncorrelated noise sources, the total ENC for the CSA can be therefore expressed as

$$\mathrm{ENC} = \sqrt{\mathrm{ENC}_{th}^2 + \mathrm{ENC}_{1/f}^2 + \mathrm{ENC}_{leak}^2}$$

[3] In SPICE simulations, the integral is numerically computed within a minimum and a maximum frequency specified for the noise analysis

Capacitances do not generate noise. Nevertheless, they determine the frequency response of the system, resulting into specific noise transfer functions to the output node in the frequency domain for each noise contribution. Therefore, both the feedback capacitance C_F and the total capacitance at the input node $C_D + C_G$ appear in the compute of the total output-referred RMS noise, hence in the ENC. According to the common terminology, CSA noise sources can be grouped in two components. A series component (series noise) includes MOS input transistor thermal noise and $1/f$ noise, as well as the input-referred thermal noise from current source loads. It is represented by a single noise voltage source connected in series with the CSA input node. A parallel component (parallel noise) includes instead the shot noise due to sensor leakage current and the thermal noise generated by the feedback resistance if considered, and is represented by a single input-referred noise current source placed in parallel with the input node. Short peaking time values (< 100 ns) increase the importance of the thermal series noise over the parallel one, without affecting the $1/f$ contribution [Rossi 2006, Rivetti 2015]. As a result, the dominant ENC contribution in the circuit under analysis derives from channel thermal noise.
Since the input device is biased in moderate inversion, simple hands noise calculations are not available. Analytical predictions necessary require the usage of the modern EKV MOS transistor model, which provides analytical equations for both large signal characteristics and small signal parameters valid in all device operating regions [Enz 1995, 2006]. In a preliminary design phase, dedicated CAD simulations and comparisons with EKV model predictions have been performed in order to characterize input-transistor thermal noise contributions in 65 nm CMOS technology [Monteil 2013]. Indeed, the complexity of the description of the MOS transistor dramatically increases with technology scaling and short channel effects. Accurate final noise performance can be derived only from post-layout simulations. Nevertheless, practical transistor sizing considerations can be derived as follows.

The thermal noise in MOS transistors has a frequency-independent spectral density that, when referred as a RMS voltage to the gate terminal, is proportional to the temperature T and inversely proportional to the transconductance g_m. For deep-submicron devices, the most general expression of the noise density is usually written as

$$\frac{d < v_{n,th}^2 >}{d\,f} = \alpha_{\mathrm{w}}\,\gamma\,\frac{4k_BT}{g_m}$$

where k_B is the Boltzmann constant, γ is a numerical coefficient which depends on transistor operating region (noise inversion factor) and α_{w} is a technology-dependent numerical correction introduced to take into account of a thermal noise increase that manifests in short channel devices (excess noise factor). In practice, most important thermal noise contributions in the telescopic cascode amplifier arise from the input transistor M1 and from PMOS current source loads M4 and M6, whereas the noise in cascode devices M2, M3 and M4 is negligible. Neglecting numerical constants, the total input-referred noise spectral density exhibits the dependence

$$\frac{d < v_{n,in}^2 >}{d\,f} \sim k_BT\left(\frac{1}{g_{m1}} + \frac{g_{m4}}{g_{m1}^2} + \frac{g_{m6}}{g_{m1}^2}\right)$$

where thermal noise contributions of PMOS current source loads have been properly referred to the input node of the amplifier.

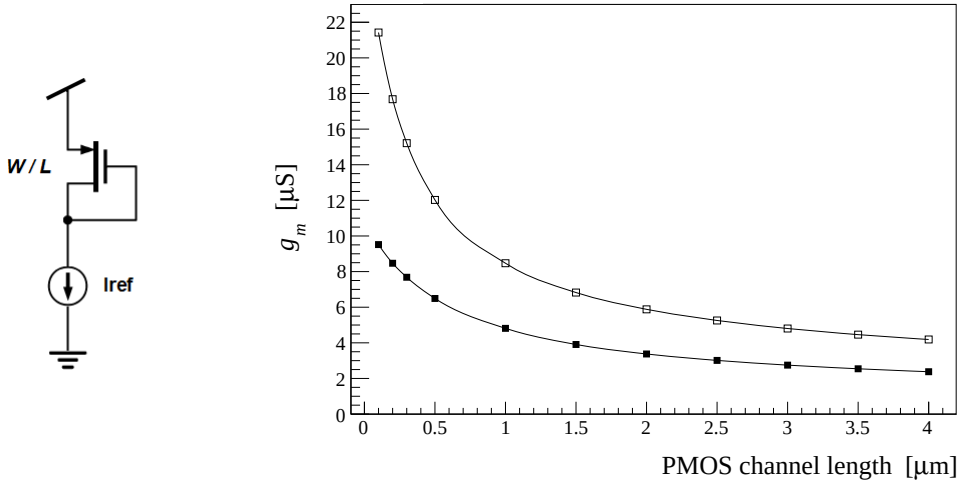


Figure 2.23: Simulated device transconductance g_m as a function of the channel length L for a PMOS diode-connected device with minimum channel width and 1.5 μA (empty points) or 500 nA (filled points) bias currents.

The thermal noise generated in PMOS devices must be therefore minimized. As one can see, this is accomplished if transconductance values g_{m4} and g_{m6} satisfy

$$\frac{g_{m4,6}}{g_{m1}} \ll 1$$

This requires a proper choice of device aspect ratios $(W/L)_4$ and $(W/L)_6$. Since $g_{m1} \approx 50$ μS is already determined from the maximum bias current allocated in the input transistor in respect of the limited power budget, in practice $g_{m4} \approx g_{m6} \leq 5$ μS ensures $g_{m4,6}/g_{m1} \leq 0.1$. The plot in Figure 2.23 presents simulated transconductance values as a function of the channel length L for a PMOS diode-connected device with minimum channel width and assuming 500 nA or 1.5 μA bias currents. As expected, the g_m decreases by increasing the channel length. PMOS transistors with $L > 1$ μm are therefore necessary in order to minimize the thermal noise from current source loads. Furthermore, a longer device offers a higher small-signal resistance r_{ds}, thus increases the impedance at the cascode output node.
Indeed, the usage of larger channel lengths introduces a fundamental additional benefit, because device operations are pushed from the moderate inversion to the strong inversion. According to modern MOS transistor modelling in fact, an inversion coefficient IC [Enz 2006, Sansen 2006] is used to discriminate among weak/moderate/strong inversion operating regions,

$$IC = \frac{I_D}{I_o\,(W/L)}$$

with I_o a normalization current (technology current) that only depends on physical constants, temperature and CMOS process parameters, but not on device aspect ratio. Conventionally, weak inversion operations (subthreshold conduction) are assumed for $IC < 0.1$ and strong inversion operations for $IC > 10$, whereas the moderate inversion region extends within $0.1 < IC < 10$. For a given bias current therefore, longer transistors have a higher inversion coefficient.

The fundamental advantage of a current source device working in strong inversion resides in reduced current mismatches $\Delta I_D/I_D$ against threshold voltage variations [Schneider 2010]. Hence, strong inversion operations must be always guaranteed in the design of current mirror structures. Since small bias currents flows in cascode branches, moderate channel lengths $L \approx$ 2-4 μm are sufficient to drive M4 and M6 in strong inversion and minimize the thermal noise.
Note that thanks to strong inversion operations, realistic device aspect ratios for M4 and M6 can be now derived from analytical calculations. Assuming in fact a square-law behaviour, the transconductance g_m can be expressed as

$$g_m \;=\; \sqrt{2\mu_p C'_{ox} \left(\frac{W}{L}\right) I_D}$$

For thermal noise minimization it is assumed that $g_{m4} \approx g_{m6}$, therefore

$$\frac{g_{m4}}{g_{m6}} \;=\; \sqrt{\frac{(W/L)_4\; I_{D4}}{(W/L)_6\; I_{D6}}} \;=\; \sqrt{\frac{(W/L)_4}{(W/L)_6} \left(\frac{1-x}{x}\right)} \;\approx\; 1$$

where current fractions $I_{D6} = x\, I_{D1}$ and $I_{D4} = (1-x)\, I_{D1}$ have been inserted. Since cascode branches are biased with two totally independent current mirrors, different channel lengths L_4 and L_6 can be adopted, whereas for a more compact layout the choice $W_4 = W_6$ is preferable. Under these assumptions we finally obtain

$$\frac{L_6}{L_4} \;\approx\; \frac{x}{1-x}$$

For a current fraction $x = 0.75$ driven by the auxiliary PMOS cascode current source, it turns out that $L_6 \approx 3\, L_4$. Assuming $L_4 = 1$ μm, analytical calculations predict $L_6 = 3$ μm.

As a matter of fact, definitive optimized values for all devices have been assessed upon CAD simulation results, verifying circuit reliability across mismatches and PVT variations. Nevertheless, all estimates that have been derived from detailed inspection analysis and design specifications are in good agreement with actual final values. As an example, the plot in Figure 2.24 presents the simulated ENC and the total RMS output voltage as a function of the input transistor channel width for the final optimized circuit and assuming 4 fF feedback capacitance and 100 fF sensor capacitance. The curve shows that an optimum value of the channel width $W_1 \approx 5$ μm minimizes the ENC, which settles to about 42 e^-. For a wider device the ENC increases at a negligible level. The same analysis is proposed in Figure 2.25 for all selectable capacitance values. Certainly a smaller feedback capacitance increases the charge-to-voltage gain, hence the ENC contribution. As a final choice, $(W/L)_1 = 8/0.2$ has been adopted for the input transistor.
Figure 2.26 shows the total output-referred noise power spectral density for nominal capacitance values. At very low frequencies one can recognize the dominant contribution of the $1/f$ noise of the input device. A simulation of the ENC as a function of the sensor capacitance C_D for the different values of feedback capacitance is presented in Figure 2.27 instead. Note that presented ENC values do not yet include noise contributions from the feedback network and refer to pre-layout simulations. A summary of most important performance parameters for the CSA is presented in Table 2.2. The final optimized transistor sizing adopted for the amplifier is summarized in Table 2.3.

Parameter	Value
DC open-loop gain	60 dB
cascode output resistance	27 MΩ
open-loop BW	2 MHz
SF drive capability	up to 200 fF
total DC supply current	2.5 μA
static power consumption	3 μW at 1.2 V supply voltage
feedback capacitance	selectable, 2.5 fF, 4 fF or 6.5 fF
maximum signal rise time	$\approx$ 20 ns
charge-to-voltage gain	$\approx$ 38 mV/ke^- at 4 fF nominal feedback capacitance

Table 2.2: Charge sensitive amplifier performance summary table (pre-layout).

Device	W/L [μm/μm]
M1	8/0.2
M2	2/0.2
M3	2.5/0.25
M4	0.5/2
M5	2.5/0.25
M6	0.5/3
M7	4/0.2
M8	0.5/6
M9-M10	1/0.08

Table 2.3: Core amplifier final optimized transistor sizing.

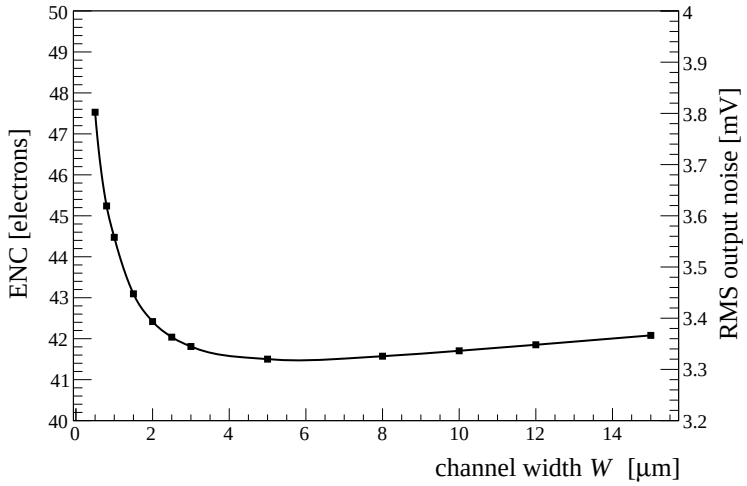


Figure 2.24: Simulated ENC as a function of the channel width W of the input transistor assuming 4 fF feedback capacitance and 100 fF detector capacitance.

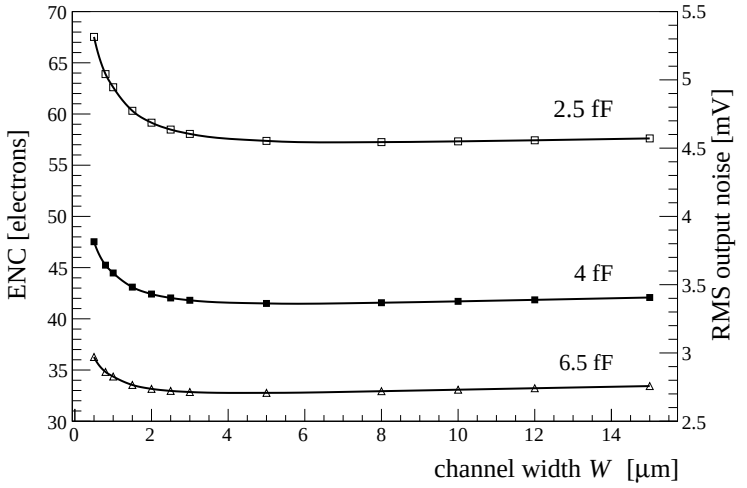


Figure 2.25: Simulated ENC as a function of the channel width for different values of feedback capacitance.

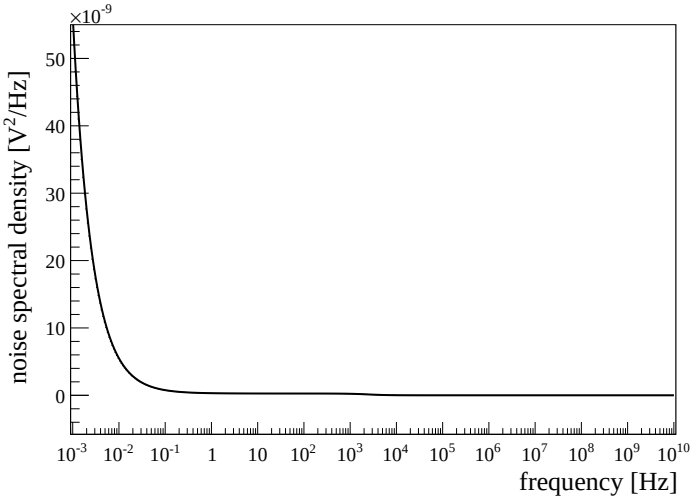


Figure 2.26: Simulated total output-referred noise spectral density. At very low frequencies one can recognize the of $1/f$ noise contribution.

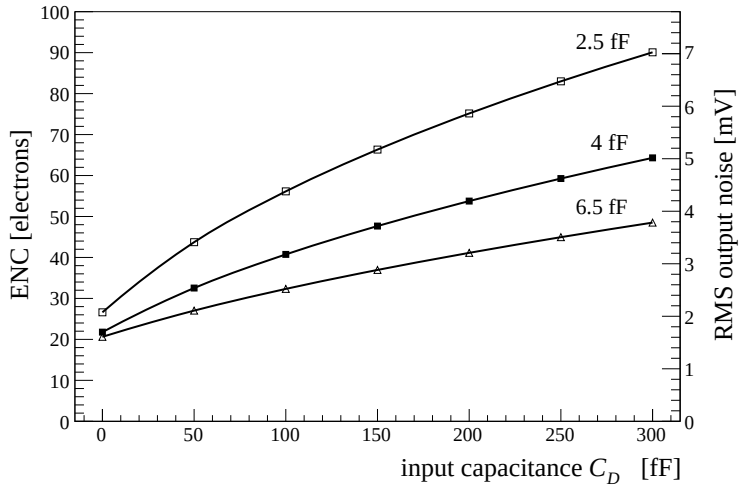


Figure 2.27: Simulated ENC as a function of the detector capacitance C_D for different selectable values of feedback capacitance.

Feedback network

The choice of a single-stage Front-End amplifier introduces relevant considerations in the design of the feedback network built around the core charge integrator. A feedback circuit is required to provide the necessary DC bias point at the input node and to discharge the feedback capacitance after a charge signal has been integrated. Due to the absence of a dedicated filtering stage, an optimized pulse waveform in terms of noise, linearity and stability must be already available at the CSA output for the subsequent hit discrimination. Furthermore, baseline shifts due to leakage currents in the pixel sensor must be compensated at the input node. The feedback network has therefore to address both shaping and leakage current compensation tasks, while keeping the circuit complexity to a bare minimum in order to meet area and low-power constraints defined for the analogue Front-End. According to CMS pixel upgrade requirements, the foreseen usage of n-type sensors (collection of electrons) does not require the design of a bipolar solution. The overall Front-End system can be therefore optimized for negative-only input sensor currents. Worst-case leakage contributions up to 20 nA after irradiation must be taken into account. Noise occupancy and SNR constraints require that ENC < 150 e^- RMS for 100 fF input capacitance. The choice of a specific pulse shaping strategy is essentially determined by charge measurement requirements and the expected event rate per pixel. On the one hand, each pixel cell has to provide a linear charge encoding up to 30 ke^- (4 MIP) with 1 ke^- minimum detectable signal. Charge digitization must be performed at the pixel level in order to move as soon as possible the signal processing into the digital domain and gain from speed and integration densities offered by a 65 nm CMOS technology. The actual necessary digitization resolution is still under investigation and is going to be defined from physics simulations. At present, guideline values are in the 5 to 8-bit range. On the other hand, assuming a 2 GHz/cm^2 hit rate for the innermost pixelated layer and 50 μm $\times$ 50 μm pixels, the expected event rate is 50 kHz per pixel, which in turn corresponds to a current pulse on average every 20 μs at the CSA input. Therefore, in order to keep the maximum inefficiency due to pulse overlaps below 1%, the Front-End amplifier output signal has to return to the baseline in less than about 200 ns when a 4 MIP charge is collected in the sensor. As already discussed, in perspective of synchronous operations in the Front-End discriminator the nominal peaking time is fixed to 12.5 ns. For signal charges larger than 4 MIP, the analogue waveform should recover to the baseline in less than 1 μs to avoid excessive dead times. Recovery capabilities for large charge signals must be therefore included in the feedback network design, demanding for the usage of a dedicated clipping circuit.

Different architectures have been proposed to retrieve charge information from radiation detection sensors. Among these, for a pixel system the time-over-threshold (TOT) technique has important advantages over a direct pulse-height measurement based on analogue sampling and subsequent A/D conversion. In fact, a simple binary architecture can be transformed into a charge measuring system without the need of any additional analogue and mixed-signal building blocks.
With the TOT technique, a pulse amplitude information is extracted from a time measurement. This is accomplished by measuring the time spent by the analogue waveform over a certain given threshold voltage. In practice, this reduces in determining the width of the digital output pulse generated by the Front-End discriminator if a signal exceeds the threshold value. A digital control logic is used to register leading-edge and trailing-edge transitions of the hit pulse, then a suitable counting logic placed at the chip periphery or at the pixel level provides the duration of the comparator response in terms of a TOT integer count. This is the same operating principle of counting-type A/D converters. A key advantage of TOT resides therefore in the simplicity of the circuit, which in turn leads to a low-power and area-effective solution to extract the information about the input charge. Indeed, the relationship between the input charge to be measured and the width of the encoded hit pulse depends on the type of signal processing adopted before the discriminator.

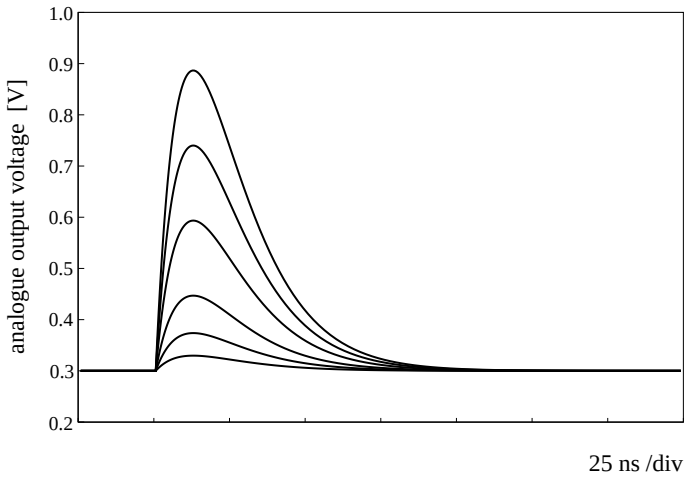


Figure 2.28: Simulated pulse waveforms for an ideal CR-RC shaper with 12.5 ns peaking time and different input charges.

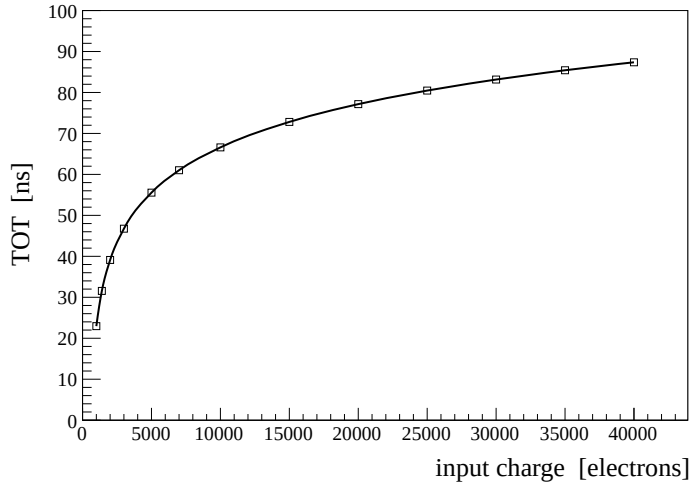


Figure 2.29: TOT as a function of the input charge for a CR-RC shaper.

As an example, let consider the well-know impulse response of a Front-End chain that implements a CR-RC pulse shaping coupled to an ideal charge integrator [Leo 1994, Spieler 2005, Rossi 2006]. For a δ-like sensor input current $i_{in}(t) = Q_{in}\delta(t)$ and assuming in the filter design equal integration and differentiation time constants τ, the output voltage as a function of time is given by

$$V_{out}(t) \;=\; \frac{Q_{in}}{C_F}\left(\frac{t}{\tau}\right)e^{-t/\tau}$$

Transient simulations for a CR-RC shaper with 12.5 ns peaking time are presented in Figure 2.28. As one can see, simulated TOT numerical values for a fixed threshold voltage exhibit a strong non linear relationship with the input charge. Since a theoretic ideal Gaussian filtering would give a logarithmic dependence between charge and TOT, this behaviour is sometimes referred to as logarithmic compression [Delagnes 2000, Kasinski 2012]. Actually, TOT non linearity does not represent a limiting factor for charge measurement. For instance, compression of high-amplitude signals has been exploited in the past [Kipnis 1997] to expand the dynamic range and increase TOT resolution for channels with small signals. Nevertheless, a linear trend is usually preferred in order to avoid elaborate offline calibration algorithms.

Triangular shaping is a natural choice to achieve a linear relationship between the width of the analogue pulse and the collected input charge. This is accomplished by means of constant current feedback solutions [Fisher 2001, Peric 2006, Llopart 2007, Karagounis 2011, Kugathasan 2011].
As sketched in Figure 2.30, the CSA feedback capacitor C_F is discharged with a constant current source I_F provided by a dedicated feedback circuitry that activates under proper bias conditions. Depending on the practical transistor level implementation, the feedback current can be generated either with basic current mirror techniques or by an auxiliary transconductance amplifier. Assuming an infinite open-loop gain and neglecting the finite rise time of the signal due to the limited BW of the core amplifier, the CSA output voltage immediately reaches its nominal maximum value Q_{in}/C_F. Then I_F turns on and begins to discharge the feedback capacitor. The impulse response in the time domain is therefore a straight line,

$$V_{out}(t) \;=\; \frac{Q_{in}}{C_F} - \left(\frac{I_F}{C_F}\right)\, t$$

The analogue voltage decreases with a constant slope and goes below the threshold at time $t = \mathrm{TOT}$. Since the threshold value must be set as close as possible to the DC baseline we can approximate $V_{out}(t = \mathrm{TOT}) \approx 0$, which in turn leads to

$$\mathrm{TOT} \;=\; \frac{Q_{in}}{I_F}$$

The discharge time has therefore a linear dependence with the input deposited charge. This can be appreciated from simulation results reported in Figure 2.32. For a given charge, the TOT does not depend on the feedback capacitance C_F and neither on the sensor capacitance for a core amplifier with infinite open-loop gain. Since the discharging current is constant, the above relationship remains valid even when the CSA output voltage exceeds the linear dynamic range, saturating the core amplifier. This represents a further advantage of the TOT technique with respect to direct pulse height sampling and digitization, allowing for charge measurements over a wide range of input charges despite reduced voltage swings available in deep submicron technologies.

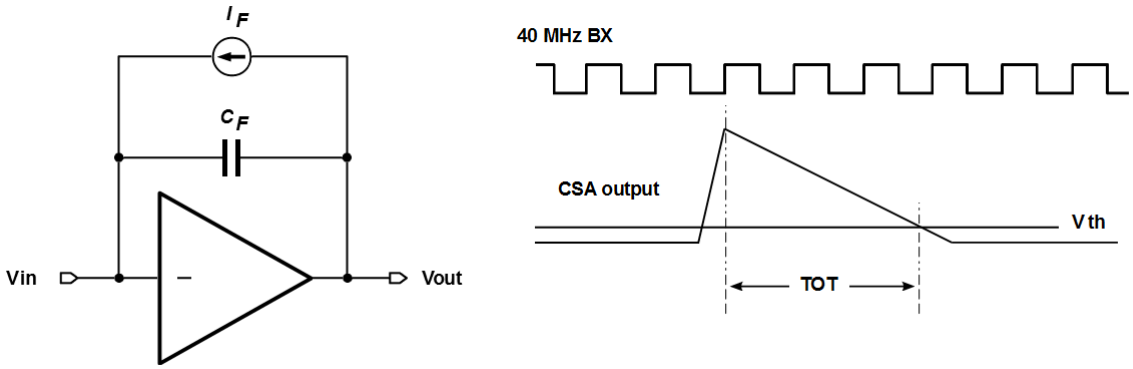


Figure 2.30: Principle of operation of a constant-current feedback (left) and definition of TOT in perspective of synchronous discriminator operations (right).

As a matter of fact, with minimum circuit complexity the TOT method offers the best compromise among power consumption, charge digitization resolution, speed and area. Such a technique is therefore well suited to be employed in multi-channel readout systems, and a TOT encoding scheme based on constant current feedback has been adopted as a baseline solution for the pixel Front-End. Depending on the foreseen event rate for the target application, the main limitation of the TOT technique can arise from the maximum conversion time, which in turn can lead to unacceptable data losses due to electronics dead times. This is determined by the frequency of the system clock used for the TOT counting logic. As discussed in great details later in the chapter, different speed and resolutions can be achieved according to the chosen counting strategy. In the proposed synchronous Front-End architecture a discrete-time voltage comparator samples the CSA analogue pulse at its nominal peaking time and TOT digitization is performed on the constant current discharging ramp, similarly to what is usually performed in Wilkinson-type A/D converters. A novel solution has been implemented to achieve high-speed and high-resolution charge encoding at the pixel level by turning the synchronous comparator into a local voltage-controlled oscillator, addressing speed requirements and providing flexible digitizations in terms of number of bits.
Note that previous considerations on triangular shaping have been derived under the assumption of ideal circuit components for both the charge integration and the constant current feedback generation. However, due to the finite open-loop gain of the core amplifier, the input capacitance reduces the maximum amplitude value Q_{in}/C_F provided by an ideal integrator, thus TOT values also depend on the sensor capacitance. Furthermore, a real constant current feedback generator implemented with MOS transistors always exhibits finite output resistance and non-zero turn on/turn off transients. As a result, a constant current discharge of the feedback capacitance is effectively achieved over a reduced amount of time. Interesting, thanks to synchronous operations a non-zero rise time for the integrated signal due to the finite BW of the CSA does not affect the effective TOT value, which in the adopted strategy is actually retrieved as if a true ideal triangular shaping would be implemented.

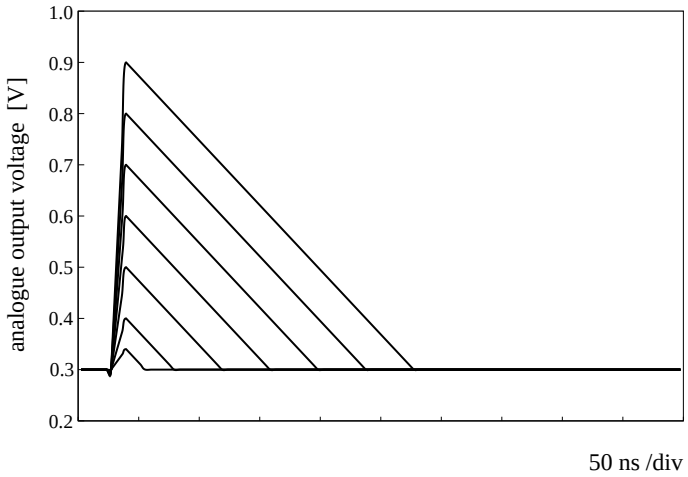


Figure 2.31: Simulated pulse waveforms with triangular shaping assuming a 12.5 ns peaking time and different input charges

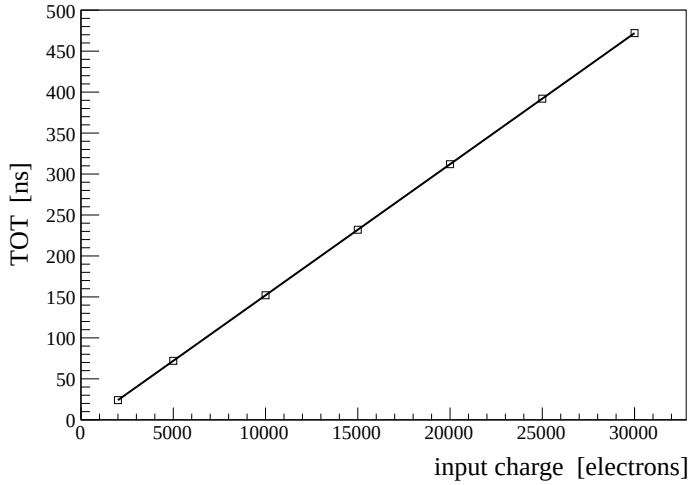


Figure 2.32: TOT as a function of the input charge for triangular pulse shaping.

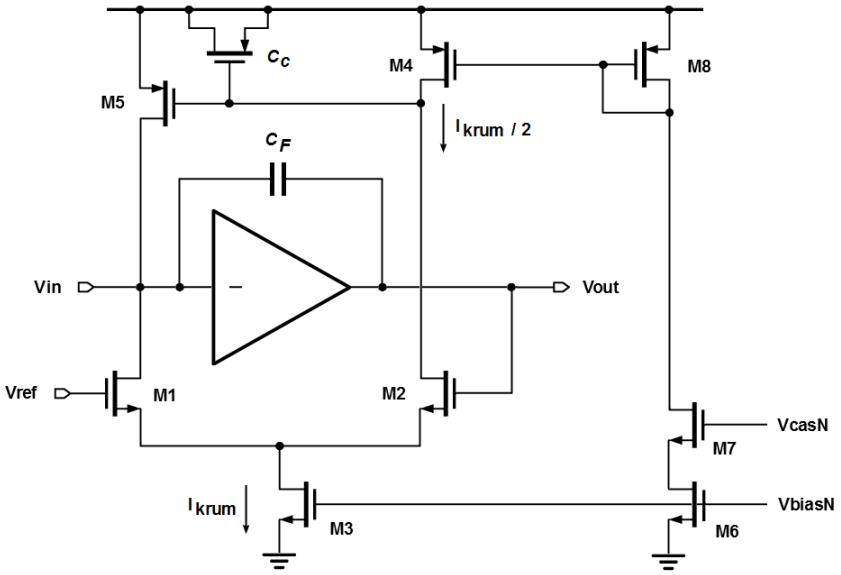


Figure 2.33: Transistor level implementation of the Krummenacher feedback network adopted for the Front-End amplifier.

Among the reach variety of active feedback networks described in literature, without doubts the Krummenacher scheme is a natural choice to accomplish triangular pulse shaping and perform at the same time a reliable leakage current compensation in a compact way. Originally proposed in [Krummenacher 1991], over the years this feedback circuit architecture has gained increased popularity in analogue Front-End design for pixel ASIC applications [Snoeys 2001, Llopart 2002, Ballabriga 2006, Llopart 2007, Valerio 2012, Ceresa 2014, De Gaspari 2014].
Figure 2.33 shows the transistor level implementation of the feedback solution adopted for the pixel Front-End. Fast recovery for large input charges has not been included instead, remanding to a second iteration the design of a dedicated clipping circuit.

Despite simplicity of this architecture, detailed inspection analysis is required to appreciate circuit operations [Szczygiel 2010]. The Krummenacher feedback is a continuous-time active reset based on an auxiliary transconductance stage. A differential pair M1-M2 compares the CSA output voltage with an external DC reference value. As reviewed thereafter, the usage of NMOS devices in the pair is optimized to compensate leakage currents coming from n-type sensors. The feedback loop is closed by connecting the drain of the transistor M1 driven by such a reference voltage to the input node of the core amplifier, providing the necessary DC operating point for the input transistor of the core amplifier. The differential pair is biased with a total tail current I_{Krum} provided by the NMOS current source M3. A second current source $I_{Krum}/2$ is implemented with PMOS transistor M4, whereas M5 and the PMOS bypass capacitor C_C play a fundamental role both in the leakage compensation process and in the signal shaping. The circuit works as follows.

At equilibrium, no charge-induced signal is integrated on the feedback capacitance C_F. Without sensor leakage current contributions, at the quiescent point the system is perfectly balanced. A bias current $I_{Krum}/2$ flows in each branch of the differential pair and M1-M2 devices have the same gate-source voltage, $V_{GS1} = V_{GS2}$. On the one hand, neglecting offset contributions due to random mismatches in the differential pair, the DC output voltage of the core amplifier is determined by the reference voltage as $V_{out,DC} = V_{REF}$. On the other hand, the input one settles above the common drain-source voltage across the tail current source M3 according to the value of V_{DS1}, $V_{in,DC} = V_{DS1} + V_{DS3}$. A first clear advantage of this architecture resides therefore in the possibility of performing different DC level optimizations for CSA input and output nodes.
Let now introduce a leakage current I_{leak} that shunts from the input node to ground, which is the case of n-type sensors. Such a DC current is integrated on the feedback capacitance C_F, rising the voltage at the CSA output node. Since the gate voltage of transistor M2 increases as well, a current larger than $I_{Krum}/2$ would flow in the right branch through M2. However, M4 is a current source that accepts only $I_{Krum}/2$, hence any extra DC current due to leakage is integrated on capacitor C_C. The gate voltage of transistor M5 is pulled down and the total current driven by M5 is regulated such that it equals $I_{Krum}/2 + I_{leak}$, thus compensating for I_{leak} and restoring equilibrium. Leakage current compensation is therefore accomplished by the low-frequency response of the feedback loop through M2, C_C and M5.
Interesting, the compensation path exhibits an inductive behaviour. In terms of small-signal parameters and complex quantities in fact, very slow voltage variations at the CSA output node $V_{out}(s)$ are converted into a small-signal current $g_{m2}\, V_{out}(s)/2$ through half the transconductance of M2 (due to single-ended operations). Such a current is then integrated on the bypass capacitor, resulting into a voltage drop in the Laplace domain

$$V_L(s) = \frac{g_{m2} V_{out}(s)}{2\, sC_C}$$

Finally, this voltage on the gate of transistor M5 is converted back into an input current $I_{in}(s)$ through the transconductance g_{m5}, hence

$$I_{in}(s) = g_{m5} \left[\frac{g_{m2} V_{out}(s)}{2\, sC_C} \right]$$

As a result, the overall transfer function $V_{out}(s)/I_{in}(s)$ is equivalent to an inductor with a complex impedance

$$Z_L(s) = \frac{2\, sC_C}{g_{m2}\ g_{m5}}$$

connected in parallel with the feedback capacitance C_F. Potential detector leakage currents (DC component) are therefore absorbed by the compensating device M5 and not by transistor M1 in the differential pair, minimizing the circuit sensitivity to DC input currents.
Note that the feedback loop is optimized to compensate DC currents discharging the input. In case of p-type sensors instead, the complementary circuit with a PMOS differential pair and NMOS loads must be adopted to properly compensate leakage contributions entering the input node.

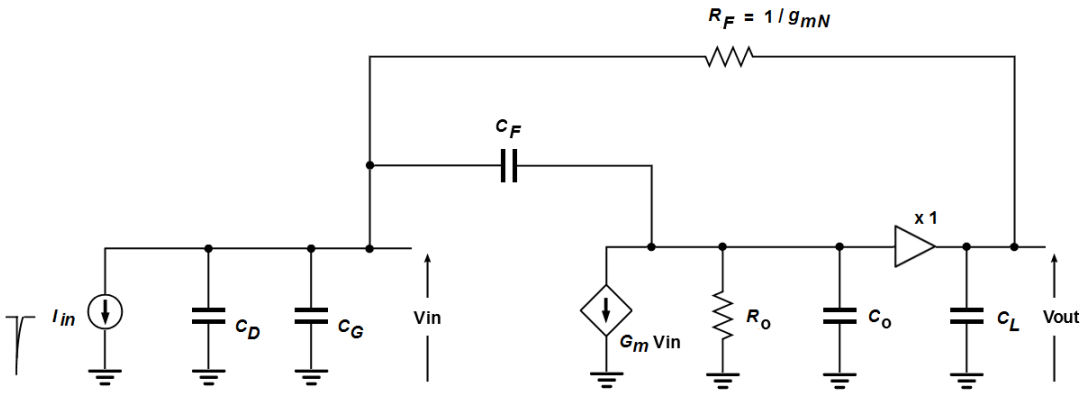


Figure 2.34: Krummenacher scheme small-signal model. For small input charges the feedback path provides an exponential discharge through a finite feedback resistance $R_F = 1/g_{mN}$. For large enough input charges the feedback capacitance is discharged instead by a constant current $I_{Krum}/2$. The output buffer isolates the system from any additional extra capacitive load C_L due to other elements connected to the Front-End amplifier.

Signal shaping is provided by the high-frequency response of the circuit instead. Neglecting the effect of the input capacitance, if a negative charge is collected in the sensor the CSA output voltage rapidly increases to about Q_{in}/C_F. Since the gate of transistor M2 is pulled up with a fast transition, the low-frequency feedback loop is not triggered. The increase of the gate-source voltage of M2 unbalances the differential pair and part of the tail current I_{Krum} is steered through M2, whereas the current driven by M1 reduces. Depending on the magnitude of the input charge, the output response has different behaviours.
For small input charges, the circuit can be linearized around the bias point and the small-signal model depicted in Figure 2.34 applies. In this case, the feedback capacitance C_F is discharged through a small signal-resistance $R_F = 1/g_{mN}$, being $g_{mN} = g_{m2}/2$ the effective transconductance available due to single-ended operations. It is therefore expected that for a δ-like input current $i_{in}(t) = Q_{in}\,\delta(t)$ the output voltage decreases with the typical time response of a transimpedance amplifier. Analytical expressions can be derived by solving nodal equations in the Laplace domain for the circuit [Peric 2004, Rossi 2006, Karagounis 2011],

$$\begin{cases} I_{in}(s) + s(C_D + C_G)V_{in}(s) + (sC_F + 1/R_F)\,[V_{in}(s) - V_{out}(s)] = 0 \\ \\ G_m V_{in}(s) + V_{out}(s)/R_o + sC_o V_{out}(s) + (sC_F + 1/R_F)\,[V_{out}(s) - V_{in}(s)] = 0 \end{cases}$$

As one can see, the feedback impedance $1/sC_F$ of a pure integrator is now replaced by the term $(1/sC_F + R_F)$ which includes the finite value of R_F.

Under the usual assumption that the core amplifier has a large open-loop gain $A_o = G_m R_o \gg 1$ the resulting transfer function $H(s) = V_{out}(s)/I_{in}(s)$ can be approximated as

$$H(s) \;\approx\; \frac{R_F\,(1 - sC_F/G_m)}{1 + as + bs^2}$$

where

$$a = R_F C_F + \left(\frac{C_D + C_G + C_o}{G_m}\right) \qquad \text{and} \qquad b = \frac{R_F}{G_m}\left[(C_D + C_G + C_o)C_F + (C_D + C_G)C_o\right]$$

Similarly to the CSA closed-loop transfer function, the system maintains the zero G_m/C_F at high frequencies due to the direct coupling through the feedback capacitance between input and output nodes. The denominator exhibits instead more complex dependences in terms of various circuit parameters, but still represents a second-order system with two poles. Depending on the actual parameter values in the small-signal model, radices of $1 + as + bs^2 = 0$ can be either real-poles or complex-conjugate poles. This can introduce instability (undershoots and ringing effects) in the output waveform. Neglecting the zero and assuming to have a stable loop, the transfer function can be put in time constant form as

$$H(s) \;\approx\; \frac{R_F}{1 + as + bs^2} \;=\; \frac{R_F}{(1 + s\tau_r)(1 + s\tau_f)}$$

where τ_r and τ_f are respectively the *rise* and the *feedback* time constants associated to two real poles. In the time domain, the response to a δ-like current pulse $Q_{in}\delta(t)$ is therefore given by the inverse Laplace transform, obtaining the transient behaviour

$$V_{out}(t) = -\frac{Q_{in} R_F}{\tau_f - \tau_r}\left[e^{-t/\tau_f} - e^{-t/\tau_r}\right] u(t)$$

valid for a stable CSA with limited bandwidth (non-zero rise time) and finite feedback resistance. Analytical equations for τ_r and τ_f as a function of all circuit parameters are obtained by solving the equation $1 + as + bs^2 = 0$. This is performed with the usage of Taylor expansions, resulting into quite elaborated formulas [Peric 2004, Karagounis 2011]. Indeed, under the assumption of well separated real poles, time constant expressions can be simplified as

$$\tau_r \;\approx\; \frac{1}{G_m}\left[(C_D + C_G + C_o) + \frac{(C_D + C_G)\;C_o}{C_F}\right] \qquad \text{and} \qquad \tau_f \;\approx\; R_F C_F$$

which in turn hold if the discharge of the feedback capacitance is longer than the CSA signal rise time, as usually demanded in most of Front-End applications. As one can see, τ_r is essentially the rise time already introduced for a charge integrator with limited bandwidth. $R_F C_F$ represents instead the discharge time constant associated to an ideal CSA with infinite bandwidth (zero rise time) and a finite resistance connected in parallel to C_F, which is characterized by the exponential decay $V_{out}(t) = Q_{in}/C_F\, e^{-t/(R_F C_F)}$.

Assuming $\tau_f = R_F C_F$, the impulse response of the two-poles small-signal model becomes

$$V_{out}(t) = -\left(\frac{Q_{in}}{C_F}\right)\frac{\tau_f}{\tau_f - \tau_r}\left[e^{-t/\tau_f} - e^{-t/\tau_r}\right]u(t)$$

The peaking time can be therefore derived as

$$t_p = \frac{\tau_r}{1 - \tau_r/\tau_f}\ln\left(\frac{\tau_f}{\tau_r}\right)$$

by requiring $dV_{out}(t)/dt = 0$, whereas the pulse amplitude is

$$V_{out}(t = t_p) = -\frac{Q_{in}}{C_F}\left(\frac{\tau_r}{\tau_f}\right)^{\tau_r/(\tau_f - \tau_r)}$$

On the one hand, the peaking time does not depend on the input charge Q_{in}. On the other hand, the pulse amplitude depends on both time constants and exhibits an amplitude loss with respect to the step amplitude Q_{in}/C_F of an ideal charge integrator. This is the well-known *ballistic deficit*, which affects any Front-End systems whenever the charge integration (signal formation) and the discharge of the feedback capacitance (continuous reset) act simultaneously with comparable time scales, representing de facto two competitive processes [Rossi 2006]. In order to minimize such an amplitude loss, the discharge time of the feedback capacitance must be much longer than the signal rise-time of the CSA, requiring $\tau_f \gg \tau_r$ as mentioned. The choice of large values for R_F would be therefore preferable, resulting also into a reduced thermal noise contribution. This is achieved by choosing a small feedback current I_{Krum}, thus reducing the transconductance g_{m2}. Certainly, for a given 50 kHz/pixel event rate specification, the usage of a too small feedback current can not be tolerated in order to avoid pulse overlaps for larger signals, for which the Krummenacher scheme has a different behaviour as discussed shortly thereafter.
The transient simulation presented in Figure 2.35 shows the output pulse obtained with the final optimized Krummenacher feedback network for 1 ke^- input charge, 4 fF feedback capacitance and 20 nA total feedback current, assuming 100 fF detector capacitance. Simulated data have been compared with the theoretical predictions of the above discussed two-poles approximation. As one can see, the simplified model is in reasonable agreement with the actual transistor-level circuit response.

Transient simulations in Figure 2.36 compare the impulse response for two different input charges under same bias conditions and circuit parameters. If the charge is small, the linear model holds and the Krummenacher scheme mimics a finite resistor R_F of value proportional to $1/g_{m2}$. Indeed, for practical TOT measurements the large-signal response of the circuit must be considered. When the initial voltage swing at the CSA output node is large enough in fact, the gate-source voltage of transistor M2 increases such that the tail current I_{Krum} is completely steered through M2, whereas M1 turns off. Hence the differential pair becomes fully unbalanced. Nevertheless, thanks to the presence of the capacitor C_C, the gate-source voltage of transistor M5 is kept constant. The feedback capacitance is therefore discharged at the CSA input node with a constant current $I_{Krum}/2$ supplied by transistor M5. As a result, the output node restores to the baseline with triangular shaping, suitable for linear TOT measurements. Both small-signal and large-signal considerations apply also for positive input charges, thus the circuit naturally offers bipolar shaping capabilities regardless the usage of a NMOS or PMOS differential pair. Two different circuit optimizations are instead required for the leakage compensation, as discussed.

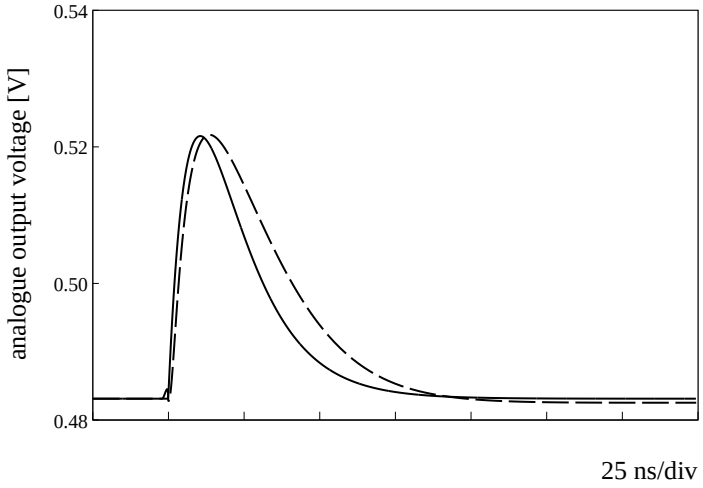


Figure 2.35: Simulated Krummenacher feedback impulse response (dashed curve) assuming 1 ke^- input charge, 100 fF detector capacitance, 4 fF feedback capacitance and 20 nA tail current. Comparison with the impulse response of a simplified two-poles linear model (solid curve). From DC operating points, a 20 nA total bias current leads to $g_{m2} \approx 280$ nS, resulting into a small-signal resistance $R_F \approx 1.8$ MΩ. The two-poles approximation is in reasonable agreement with the actual transistor-level circuit response.

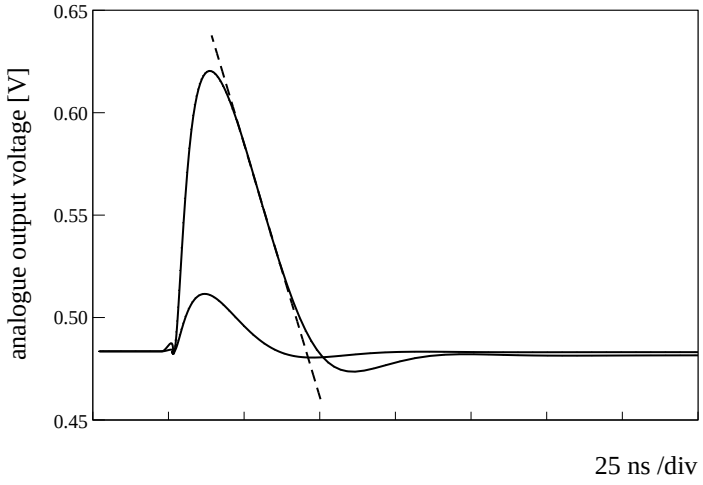


Figure 2.36: Simulated Krummenacher feedback impulse response for 1 ke^- and 4 ke^- input charges assuming same Front-End parameters. If the initial voltage swing is large enough, the differential pair M1-M2 becomes fully unbalanced and the feedback capacitance is discharged with a constant current $I_{Krum}/2$ supplied by M5, providing triangular shaping for linear TOT charge encoding.

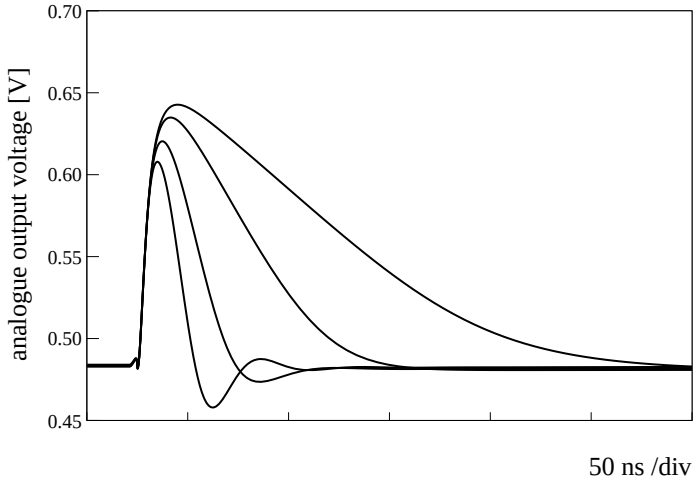


Figure 2.37: Simulated Krummenacher feedback impulse response for different values of feedback current I_{Krum} = 10 nA, 20 nA, 50 nA and 100 nA, assuming 4 ke^- input charge (MIP), 100 fF detector capacitance and 4 fF feedback capacitance. Depending on the chosen current, the actual circuit response is affected by amplitude losses (ballistic deficit), peaking time variations and loop instability (undershoots or ringing).

A constant-current linear discharge is available only over a limited amount of time. The actual transistor-level circuit response has in fact non-linear transients. In particular, the output signal exhibits an undershoot with respect to the baseline. Circuit analysis revealed the existence of two feedback loops. The leakage current compensation is assured by a low-frequency loop with an inductive behaviour. Signal shaping is provided instead by the high-frequency response. Indeed, the effective circuit response is determined by both paths, which in turn act simultaneously and are two competitive processes. The actual closed-loop transfer function depends on a large number of zeroes and poles when the complete frequency response of the feedback network coupled the core amplifier is considered. Simple analytical calculations are no more available when the overall transistor-level implementation is considered. As an example, the core amplifier has been always considered a single-pole system, which is only an approximation. The output stage has been in fact supposed to be a unity-gain ideal buffer. However, due to the low bias current adopted for the output source follower in order to cope with the severe power budget defined for the Front-End, the non-zero output impedance of the stage introduces an additional pole in the system. Furthermore, the low-frequency loop actually contributes to the signal shaping, differentiating the output pulse due to its inductive behaviour. This can introduce significant undershoots or ringing effects in the output waveform if a very limited bandwidth is not guaranteed for the leakage compensation feedback path. As shown in transient simulations presented in Figure 2.37, for a given input charge the actual impulse response is affected by amplitude losses, peaking time variations and loop instability depending on the value of the total feedback current. A small feedback current I_{Krum} leads to a longer discharge time, thus maximizing the charge-to-voltage gain. Furthermore, the return of the signal to the baseline is not affected by undershoots. However, a too long discharge time can not be tolerated in order to cope with the foreseen event rate. A larger feedback current provides a faster discharge of the feedback capacitance, but also increases the ballistic deficit and significantly degrades the loop stability. As a matter of fact, the presence of an excessive overshoot above the baseline is undesirable.

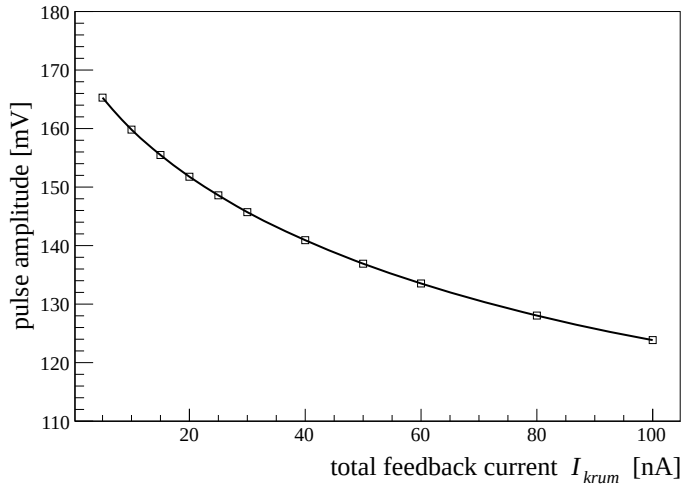


Figure 2.38: Pulse amplitude as a function of the total feedback current I_{Krum} assuming 4 fF feedback capacitance, 100 fF detector capacitance and 4 ke^- input charge. Due to ballistic deficit, the Krummenacher scheme is affected by amplitude losses for larger feedback currents.

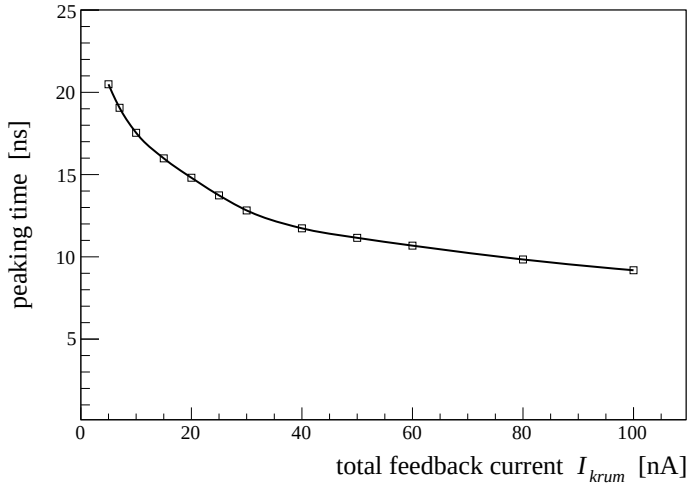


Figure 2.39: Peaking time variations as a function of the total feedback current I_{Krum} assuming nominal Front-End parameters and 4 ke^- input charge.

Pre-layout characteristics for the ballistic deficit and the peaking time variation as a function of the feedback current are shown in Figure 2.38 and Figure 2.39 respectively. Note that a nominal peaking time of 12.5 ns is obtained with a total feedback current of about 40 nA. As discussed later in the chapter, peaking time variations around this nominal value do not represent a key limiting factor in perspective of synchronous operations for the hit discrimination, provided that a signal is found above the threshold when the CSA analogue waveform is sampled by the comparator.

Due to the complexity of the feedback loop, practical transistor sizing for the Krummenacher scheme becomes more challenging. A final optimized solution can be only validated upon CAD simulation results. Nonetheless, fundamental design guidelines can be derived from loop stability requirements. Referring to schematic in Figure 2.40, it turns out that four major poles can be identified inspecting most important capacitance contributions in the circuit under consideration. A first pole ω_{p1} is due to the compensation capacitor C_C and can be expressed as

$$\omega_{p1} = \frac{g_{m5}}{C_C}$$

At low frequencies, the leakage compensation path must be slow enough such that the loop is sensitive to DC only or very slow variations. The bypass capacitor C_C must be therefore made large in order to limit the bandwidth. Hence it is expected that ω_{p1} is the first dominant pole at low frequencies. A second pole ω_{p2} is associated to the feedback capacitance C_F. As already discussed, when the small-signal behaviour of the circuit is considered this leads to a time constant $\tau_f \approx R_F C_F$, being R_F a small-signal resistance proportional to the inverse of the transconductance g_{m2}. Hence we can write

$$\omega_{p2} \approx \frac{g_{m2}}{C_F}$$

A third pole ω_{p3} located at intermediate frequencies arises from the total parasitic capacitance C_b seen at the source of transistors M1-M2 in the differential pair,

$$\omega_{p3} = \frac{g_{m2}}{C_b}$$

Finally, an additional pole ω_{p4} is due to the finite bandwidth of the core amplifier, which in turn determines a non-zero rise time constant τ_r for the signal integrated over C_F. This poles includes contributions from different circuit parameters, such as the effective transconductance G_m of the cascode input stage, the total input capacitance $C_{in} = C_D + C_G$ seen at the input node, the feedback capacitance C_F and the the capacitance C_o that determines the limited bandwidth ω_o of the CSA, as already discussed. Assuming that the core amplifier simply behaves as a single-pole system with transfer function $A_o/(1 + s/\omega_o)$, this pole can be essentially expressed as

$$\omega_{p4} \approx \frac{C_F}{C_{in}} A_o \omega_o = \frac{G_m C_F}{C_{in} C_o}$$

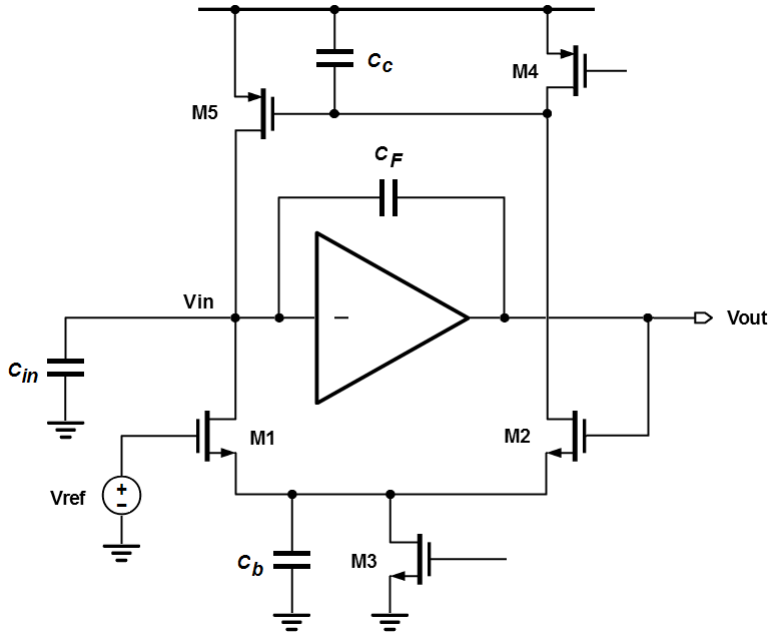


Figure 2.40: Most important capacitance contributions in determining the loop stability of the implemented Krummenacher feedback network.

Due to the large DC gain of the core amplifier $A_o \gg 1$, ω_{p4} is the pole at highest frequency. The loop stability is guaranteed if the poles are well separated, thus requiring

$$\omega_{p1} \ll \omega_{p2} \ll \omega_{p3} \ll \omega_{p4}$$

This introduces some fundamental design considerations and trade-offs. At low frequencies, the condition $\omega_{p1} \ll \omega_{p2}$ implies that

$$\frac{g_{m5}}{C_C} \ll \frac{g_{m2}}{C_F}$$

Such a request does not represent a key issue. As mentioned, the compensation capacitor C_C is made large, usually of the order of hundreds fF. A precise capacitance value is actually not required, thus a simple MOS capacitor with sufficiently large channel length and width can be adopted. This leads to a significant area contribution. Certainly C_C is much larger than the feedback capacitance C_F, which has a nominal value of 4 fF. On the other hand, the transconductance g_{m5} must be minimized, whereas the choice for g_{m2} depends on further considerations. Note that transistors M5 and M6 work as current sources, hence device operations must be pushed in strong inversion and their thermal noise must be minimized. As already discussed, this can be accomplished with small channel widths and increasing the channel length up to 1-3 μm, benefiting also of a larger output impedance.

The DC baseline is determined by the auxiliary reference voltage V_{ref}, which has a nominal value of about 500 mV. Since the bias currents involved in the feedback network are very small, from a few nA up to 100 nA, transistors M1-M2 are biased in weak inversion. Under this operating condition the transconductance-to-current ratio g_m/I_D is constant and can expressed as

$$\frac{g_m}{I_D} = \frac{1}{nk_BT/q}$$

being k_BT/q the well known thermal voltage (26 mV at $T = 25$ °C) and n a numerical coefficient (weak-inversion slope factor) that only depends on the CMOS process [Tsividis 1999, Enz 2006, Sansen 2006, Jespers 2010]. At equilibrium, transconductance values g_{m12} are therefore uniquely determined by the bias current $I_{Krum}/2$ that flows in each branch of the differential pair, without dependences on devices aspect ratios $(W/L)_1$ and $(W/L)_2$. Indeed, available transconductance values span about two orders of magnitudes, from a few nS up to μS, as already presented in the characteristic of Figure 2.8.

In practice, most important design trade-offs for the Krummenacher architecture derives from the intermediate poles ω_{p2} and ω_{p3}. The condition $\omega_{p2} \ll \omega_{p2}$, which in turn requires

$$C_F \gg C_b$$

is not a priori guaranteed. The lumped capacitance C_b is determined by parasitic contributions from transistors M1-M2 as well as from the tail current source M3. In order to ensure loop stability, C_b should be minimized, thus requiring transistors with small aspect ratios. However, transistors M1-M2 should be made large in order to mitigate current mismatches in the differential pair. Furthermore, the channel length of transistor M3 must be increased to guarantee strong inversion operations and low thermal noise. Parasitic contributions of a few fF easily become comparable with the small value of the feedback capacitance C_F adopted for the CSA. Hence the parasitic capacitance C_b heavily contributes in degrading the stability of the feedback response, introducing undershoot or ringing effects in the shape of the output waveform. The usage of a larger feedback capacitance is therefore preferable, increasing also the magnitude of the fourth pole ω_{p4} at high frequency and improving the phase margin of the core amplifier. Nonetheless, a larger value of C_F decreases the charge-to-voltage gain of the CSA, which can not be tolerated in order to cope with the 1 ke^- minimum detectable charge constraint defined for the pixel Front-End. As a result, the circuit exhibits a trade-off between channel-to-channel uniformity, loop stability and charge-to-voltage gain. As mentioned, a final optimized transistor sizing can be assessed only after extensive stability and transient simulations, verifying the transistor-level response across PVT corners, process and mismatch variations and worst-case leakage currents. Furthermore, additional contributions from layout parasitics can be significantly modify the circuit response, requiring therefore a careful layout floorplan and post-layout simulations.

Detailed final transistor sizing adopted for the feedback network is presented in Table 2.4. Most important pre-layout circuit performance are summarized instead in Table 2.5.

Device	W/L [μm/μm]
M1-M2	1.2/1
M3	0.2/4
M4-M5	0.5/3
M6	0.2/4
M7	0.2/3
M8	0.5/3
PMOS cap	8/7

Table 2.4: Feedback network final optimized transistor sizing.

Parameter	Value/specification
total feedback current I_{krum}	selectable, 1 nA - 100 nA
differential pair DC reference voltage	490 mV nominal
compensation capacitor	$\approx$ 100 fF, PMOS capacitor
static power dissipation	0.12 μW at 100 nA
peaking time	$\approx$ 12.5 ns at 40 nA
ENC at 100 fF input capacitance	90 e^- RMS at 40 nA
charge-to-voltage gain (4 fF feedback cap.)	36 mV/ke^- at 40 nA
charge-to-voltage linear range	8-10 ke^-
TOT linearity	up to 40 ke^-
PSRR	$<$ 0 dB at all frequencies

Table 2.5: Krummenacher feedback performance summary table (pre-layout).

Most significant simulation performance results for the final optimized circuit are presented in the following. If not otherwise specified, nominal Front-End parameters in terms of feedback capacitance and detector capacitance are assumed. Furthermore, the TOT has been evaluated assuming a nominal threshold value of 12 mV above the baseline.
Transient simulations in Figure 2.41 shows the analogue waveform by increasing the input charge. For the same test bench, the pulse amplitude as a function of the input charge is presented in Figure 2.42. As one can see, charge-to-voltage linearity is maintained only for a limited range of charges, up to about 8 ke^-. Indeed, a saturation of the Front-End amplifier does not affect the linearity of the TOT, as shown in Figure 2.43. For a fixed 4 ke^- input charge, a plot of the TOT as a function of the feedback current I_{Krum} is presented in Figure 2.44. The expected reverse proportionality between TOT and feedback current is well confirmed by the linearity of the TOT reported as a function of $1/I_{Krum}$, as shown in Figure 2.45.
Transient simulations in Figure 2.46 have been obtained instead by introducing leakage currents up to 50 nA at the input of the CSA. Stability of the feedback loop up to the maximum worst-case leakage current foreseen after irradiation is guaranteed. As shown in Figure 2.47, leakage-induced DC baseline variations are negligible and efficiently compensated by the Krummenacher scheme.
ENC performance have been investigated with both transient noise simulations and AC noise analyses. A small sample of 10 transient noise simulations for a minimum charge signal of 1 ke^- is presented in Figure 2.48, whereas a distribution of 100 sampled noise-induced baseline variations is shown in Figure 4.18. Assuming 4 fF feedback capacitance, 100 fF detector capacitance and 40 nA feedback current the RMS noise at the CSA output node is about 3.5 mV, which in turn lead to an ENC of about 90 e^- RMS. ENC characteristics as a function of the detector capacitance for different feedback currents are presented in Figure 2.50 instead. As one can see, the final optimized design guarantees low-noise requirements demanded from design specifications. A simulation for the power-supply rejection ratio (PSRR) is presented in Figure 2.51. Finally, sample transient MC simulations for the output waveform are presented in Figure 2.52. A distribution of TOT values across MC iterations is shown instead in Figure 2.53.

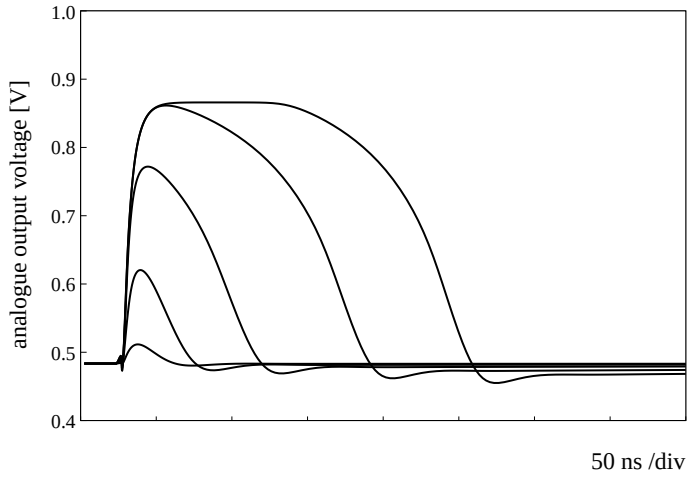


Figure 2.41: Simulated Krummenacher feedback impulse response for different input charges Q_{in} = 1 ke^-, 4 ke^-, 10 ke^-, 20 ke^- and 30 ke^- assuming nominal Front-End parameters.

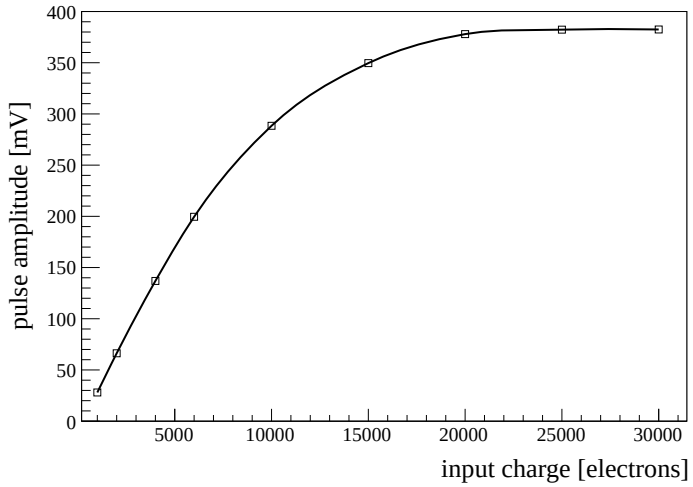


Figure 2.42: Pulse amplitude as a function of the input charge assuming 40 nA feedback current. Charge-to-voltage linearity is maintained up to about 8 ke^-. A saturation of the amplifier does not affect the linearity of TOT measurements.

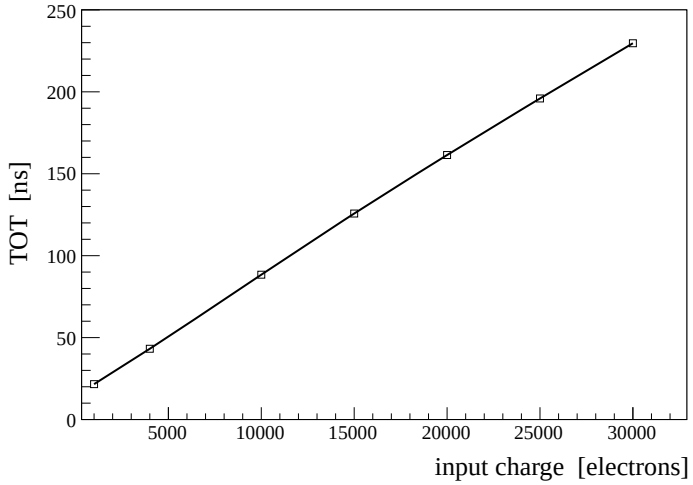


Figure 2.43: TOT as a function of the input charge.

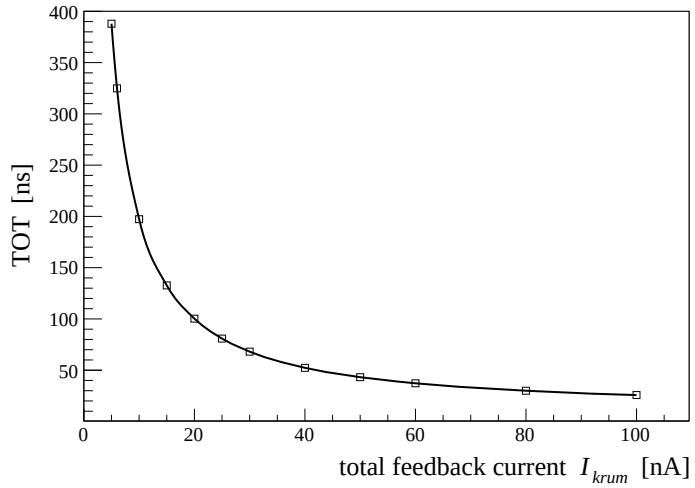


Figure 2.44: TOT as a function of the total feedback current I_{Krum} for a fixed 4 ke^- input charge.

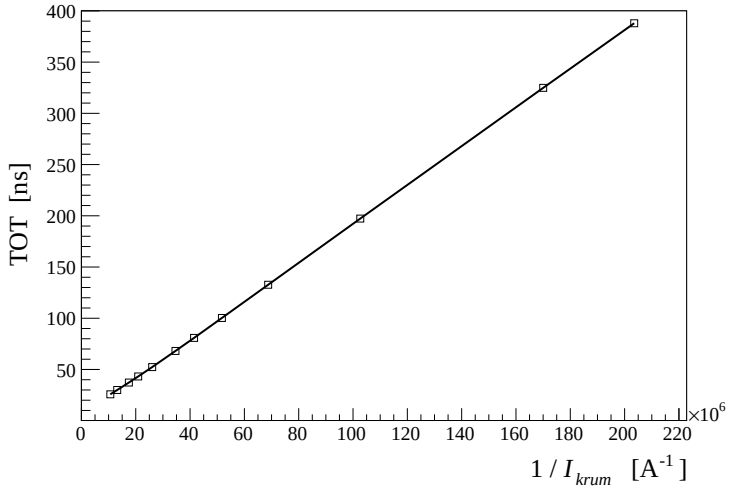


Figure 2.45: TOT reported versus $1/I_{Krum}$. The linearity confirms the reverse proportionality between TOT and the feedback current.

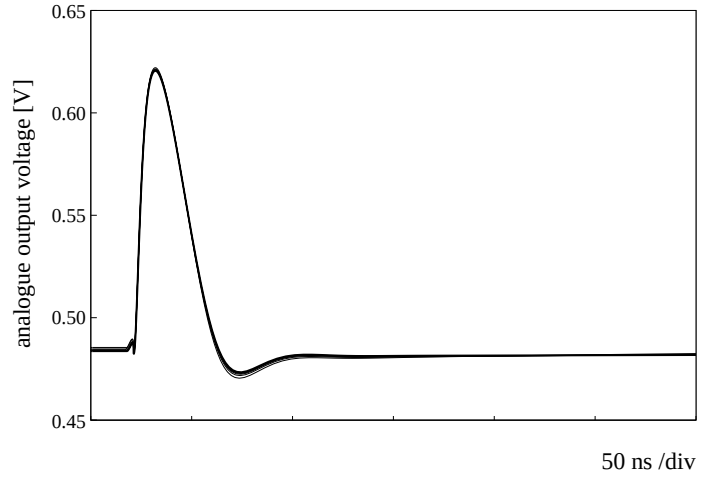


Figure 2.46: Analogue waveform for different values of sensor leakage current, up to 50 nA. Stability of the feedback loop up to the maximum worst-case leakage current foreseen after irradiation is guaranteed.

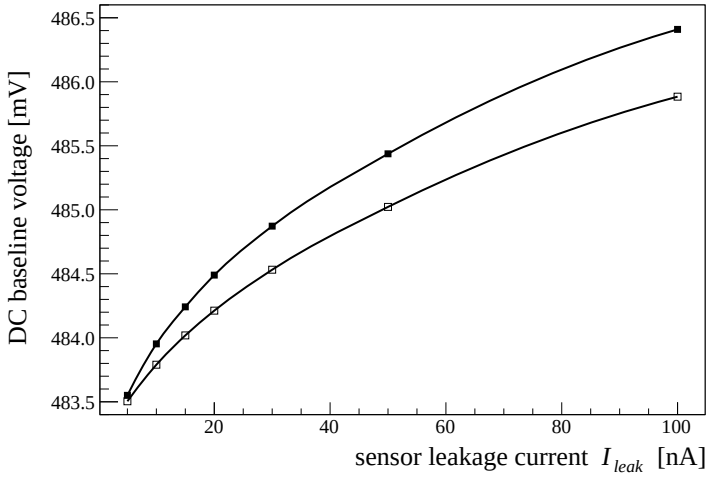


Figure 2.47: DC baseline shift as a function of the leakage current assuming 10 nA (filled points) or 20 nA (empty points) feedback current. Leakage-induced baseline variations are negligible and efficiently compensated by the Krummenacher scheme.

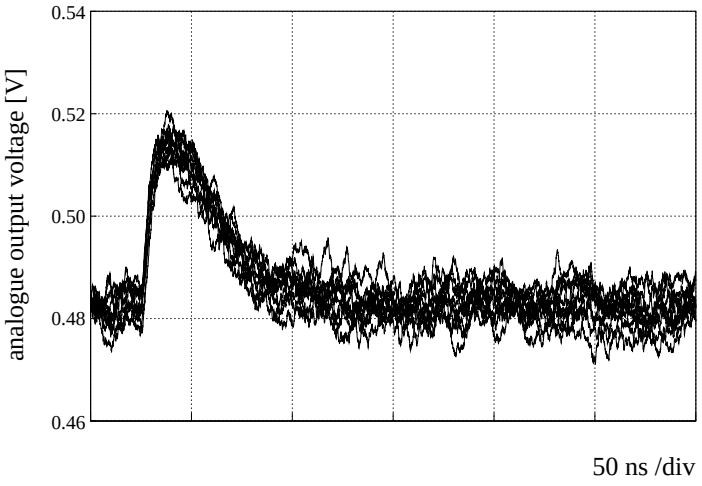


Figure 2.48: Sample of 10 transient noise analyses for a minimum input charge of 1 ke^- and nominal Front-End parameters. With a nominal pulse amplitude of about 30 mV and a noise floor of 3.5 mV RMS, the signal-to-noise ratio for the minimum signal of interest is SNR $\approx$ 8.

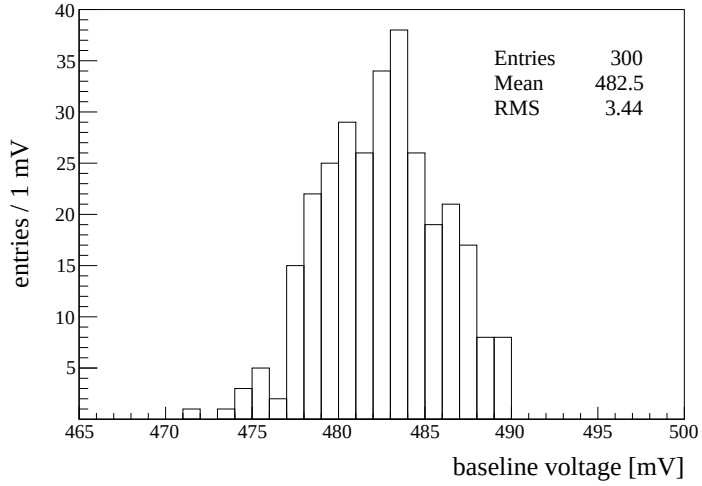


Figure 2.49: Distribution of noise-induced random baseline variations. For nominal Front-End parameters, the noise floor is about 3.5 mV RMS, corresponding to an ENC of 90 e^- RMS.

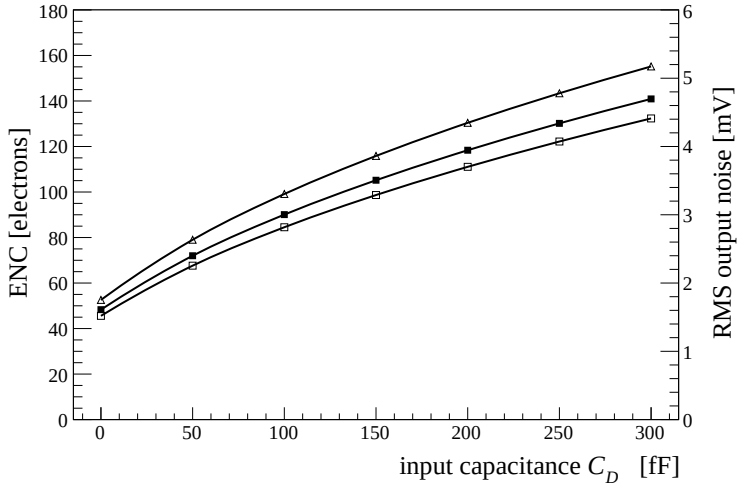


Figure 2.50: Simulated ENC as a function of the detector capacitance for 10 nA, 20 nA or 40 nA feedback current. For a 100 fF input capacitance, the design requirement ENC $< 150\ e^-$ RMS is always satisfied. Note that due to ballistic deficit, the ENC increases by increasing the feedback current.

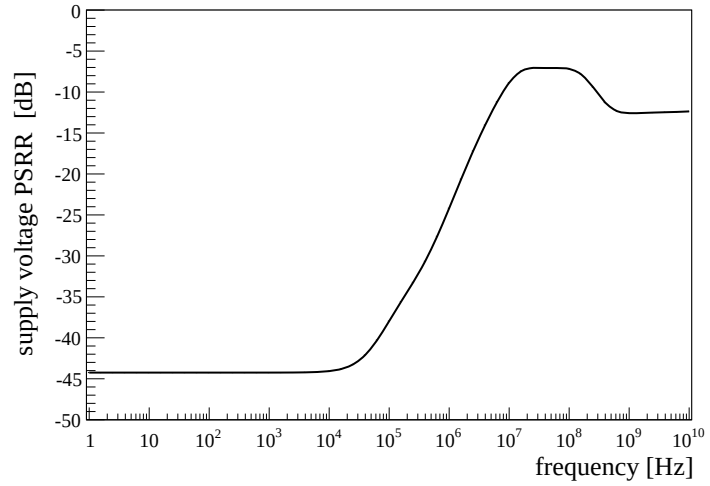


Figure 2.51: Simulated power-supply rejection ratio (PSRR) as a function of frequency. As one can see, PSRR < 0 dB at all frequencies, as requested to ensure adequate noise-immunity against disturbances on the analogue supply rail *VDDA*.

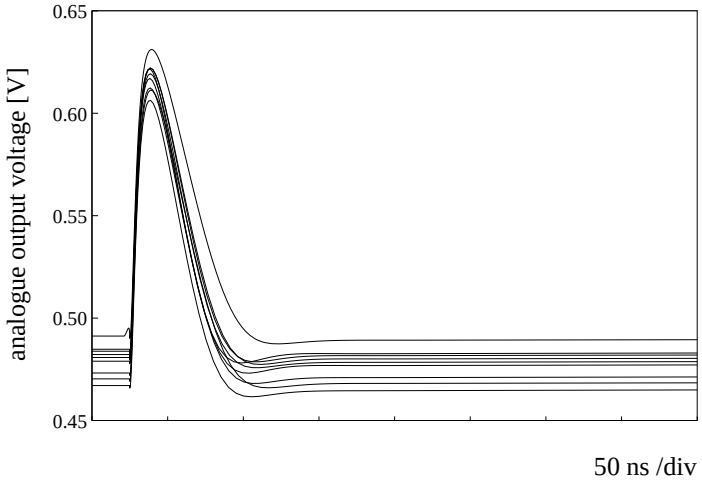


Figure 2.52: Transient MC simulations assuming nominal Front-End parameters and 1 ke^- input charge.

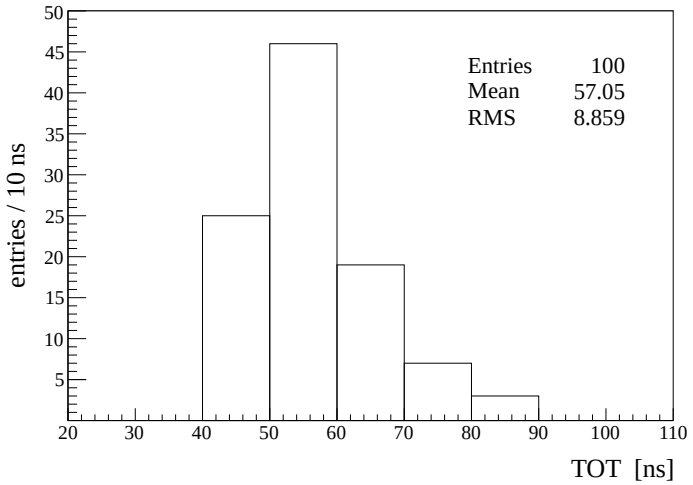


Figure 2.53: Distribution of the TOT value across 100 transient MC iterations for a minimum input charge of 1 ke^-.

Test charge injection circuit and sensor emulation

The design of a reliable on-pixel calibration circuit plays a fundamental role in order to verify proper analogue Front-End functionality and characterize the overall pixel ASIC when no detector is connected. This remains an essential task also when a pixel sensor is bump-bonded to the chip. Furthermore, systematic calibration procedures are extensively required for both commissioning and monitoring purposes during in-situ operations. Thereby each cell must include the possibility of inject precise test charges at the CSA input node. As usually performed, this is accomplished by applying an analogue voltage step with programmable amplitude over one or more well-defined test capacitors connected at the input of the Front-End chain [Spieler 2005, Rossi 2006].
A voltage step must be therefore available in each pixel cell. In order to electrically stimulate only selected pixels, a configuration bit is required to enable/disable the test feature in each pixel disconnecting the injection capacitor from the test pulse. The choice of a practical pulsing strategy is an important point. Certainly, a global voltage step generated in the chip periphery or provided off-chip with suitable instrumentation can be distributed to all pixels. This is a simple and popular solution adopted for small chip prototypes. Indeed, such an approach has a fundamental drawback. In fact, the direct propagation of an analogue pulse is susceptible to injection timing degradation due to delays introduced by metal interconnections. The usage of a low-resistance bus is therefore mandatory in order to minimize RC-effects. Nevertheless, interconnection delays become an issue in the design of full-size chips, where line resistances are not negligible if distances of the order of cm are considered. The usage of a dedicated pulse generator in the chip periphery also significantly increases the overall chip power dissipation. From design specifications, a rise time below 5 ns must be guaranteed for calibration step signals. In practice, the requirement of a well-defined injection timing becomes more important in a small pixel array in perspective of synchronous Front-End operations.

As a matter of fact, a more realistic and power-efficient solution is represented instead by a local generation of the analogue test pulse starting from a static DC voltage distributed to all pixels. A digital signal is then used to properly drive switches that commute the voltage on the top plate of the injection capacitor from ground to the supplied DC level and vice versa. A precisely timed voltage step is therefore achievable thanks to the usage of a digital control. Following such an approach, the calibration circuit presented in schematic of Figure 2.54 has been adopted.
A test capacitor *Ccal* driven by the *Vcal* voltage is connected in series with the CSA input node. The external DC voltage *Vcal_level* is distributed to all pixels along with a CMOS digital control signal *TestP*. A *TestP_EN* configuration bit is then used to enable the charge injection on the pixel. A couple of CMOS switches controlled by logic values $SW1$ and $SW2$ is used to toggle the voltage *Vcal* between the analogue ground rail and *Vcal_level*. For precise timing, additional CMOS switches always turned on are used to compensate inverter delays. Finally, three selectable shunt capacitors of values 25 fF, 50 fF of 100 fF have been added in order to emulate different values of sensor capacitance by means of SEL_CIN25F, SEL_CIN50F and $SEL\_CIN100F$ configuration bits. This is of primary importance for an experimental characterization of the Front-End noise as a function of the input capacitance. When *TestP_EN* is low, $SW1$ is forced low and $SW2$ to the supply voltage. Hence *TestP* is not propagated and *Vcal* is always kept to the ground rail. With *TestP_EN* set to high instead, *TestP* drives CMOS switches in complementary ways. If *TestP* is initially low and goes high, so does $SW1$, whereas $SW2$ goes low. Thus *Vcal* commutes from ground to *Vcal_level* and a positive charge is injected. On the contrary, if *TestP* is initially high and switches to low a negative charge is injected. As a result, current signals of both polarities can be presented at the input of the Front-End chain, as shown in the transient simulation reported in Figure 2.55. From design specifications, only negative input charges are of interest, but the Krummenacher feedback can naturally handle signals of both polarities, as discussed.

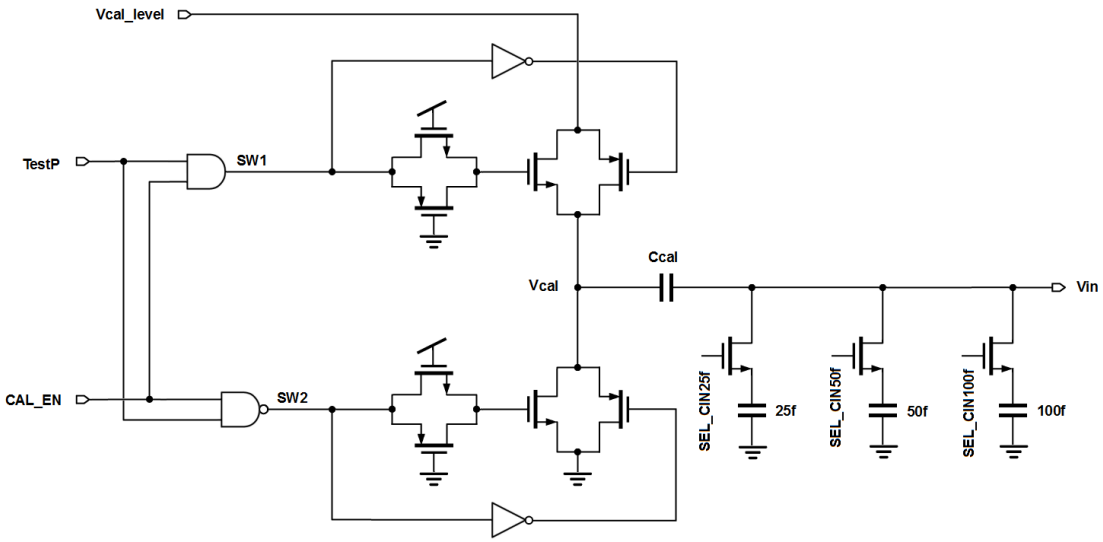


Figure 2.54: Test charge injection circuit. Selectable shunt capacitors of values 25 fF, 50 fF and 100 fF have been added in order to mimic different sensor capacitances at the CSA input node.

The absolute value of the injected charge is $Qcal = Ccal\ Vcal\_level$. Without doubts, a practical choice of the value of the injection capacitance is the most critical point. On the one hand, *Ccal* must allow to inject test charges covering the full dynamic range demanded for the pixel application. On the other hand, pixel-to-pixel differences due to process and mismatch variations in the capacitance value must be taken into account. According to analogue Front-End requirements, the calibration circuit has to provide test charges from a minimum value of 0.1 fC up to 10 fC. The maximum voltage level that *Vcal_level* can assume is the analogue supply voltage, 1.2 V. Thus a capacitor in the 8-10 fF range is required. As a final choice, a capacitance value of 8 fF has been selected. Since the overall Front-End calibration relies on the precision of the injection capacitor, the usage of a characterized library element in terms of simulation models for mismatches and process variations is recommended. A Metal-Oxide-Metal (MOM) capacitor offered by the 65 nm CMOS process has been therefore adopted, as already performed for feedback capacitances in the charge sensitive amplifier. A calibration curve for the amount of injected charge as a function of the DC calibration level is presented in Figure 2.56. The charge value has been retrieved as a numerical integral of the injected current pulse at the CSA input node.
Transient simulations in Figure 2.57 show the analogue waveform obtained for different values of shunt capacitors connected at the input of the CSA, assuming 100 mV DC level (about 5 ke^- input charge), 4 fF feedback capacitance and 40 nA total feedback current. Figure 2.58 shows instead the pulse amplitude as a function of the DC calibration voltage. As already discussed, linearity is maintained only over a limited range, up to about 8 ke^-. However, saturation of the amplifier does not compromise the linearity of TOT measurements.
As described in the next chapter, in the final assembled small pixel arrays the reference voltage *Vcal_level* must be provided off-chip using suitable instrumentation. The usage of high-linearity and programmable 7-8 bit resolution DC levels generated by a dedicated D/A converter placed at the chip periphery will be addressed in a second iteration. Furthermore, the possibility of distributing *TestP* as a digital CMOS differential signal can be considered, as already performed in a previous design of the INFN Torino VLSI Design Laboratory from which the above circuit has been derived [Kugathasan 2011].

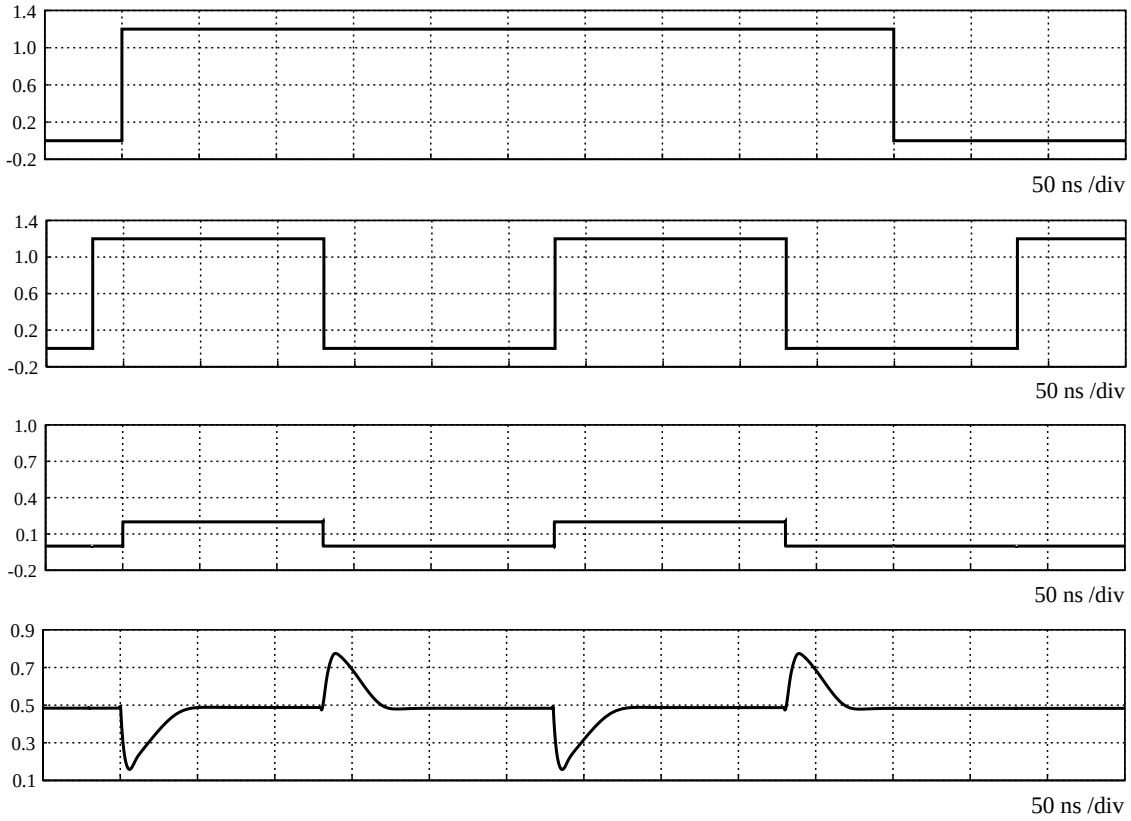


Figure 2.55: Transient simulation for the test charge injection circuit. From top to bottom: *TestP_EN*, *TestP*, *Vcal* and the resulting analogue waveform at the CSA output node.

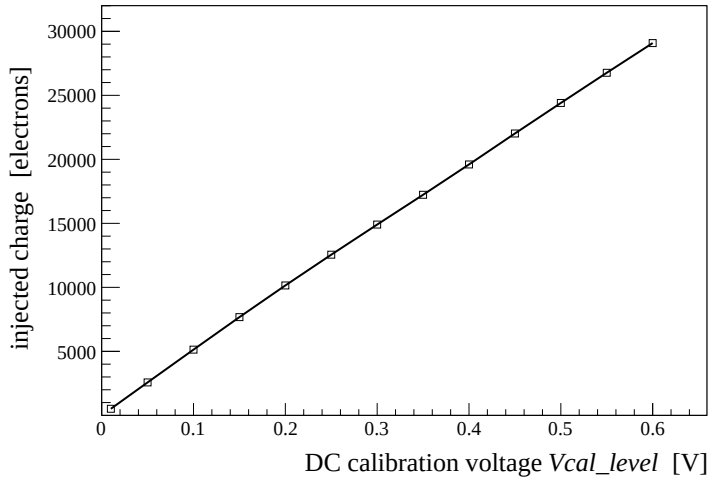


Figure 2.56: Injected charge as a function of the DC calibration level. The charge value has been retrieved as a numerical integral of the injected current pulse at the CSA input node.

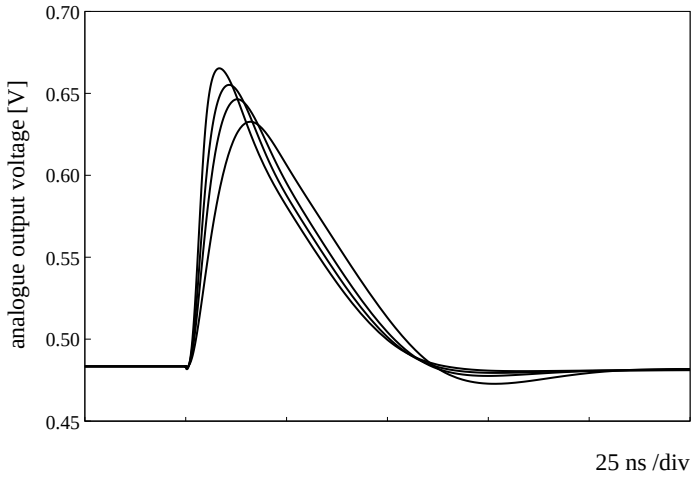


Figure 2.57: Simulated analogue waveform for different values of shunt capacitors connected at the input of the CSA (0 fF, 25 fF, 50 fF and 100 fF) assuming 100 mV DC calibration voltage (about 5 ke^-input charge) and nominal Front-End paramenters (4 fF feedback capacitance, 40 nA total feedback current).

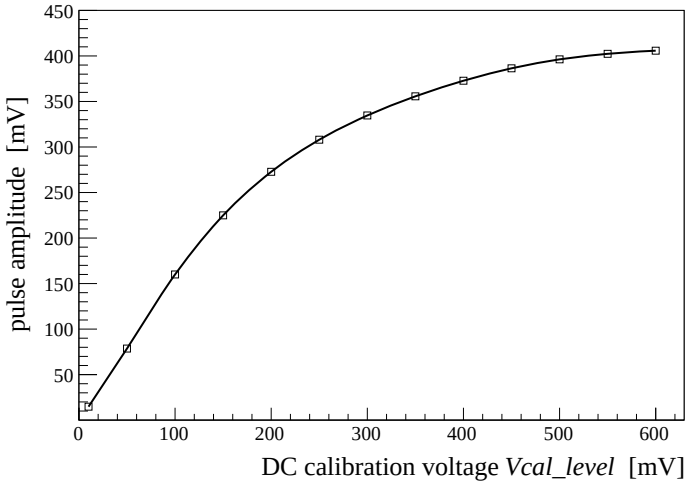


Figure 2.58: Pulse amplitude as a function of the DC calibration level. Saturation of the amplifier above 8 ke^- (about 150 mV DC level) does not affect the linearity of TOT measurements.

2.5 Discriminator design

The last stage of the pixel Front-End chain is a hit discriminator that receives as inputs the CSA analogue output and a reference threshold voltage[4]. The circuit is used to generate a digital pulse when the analogue waveform provided by the Front-End amplifier is found above the given threshold. Hit information is then fed to a dedicated on-pixel control logic circuitry for necessary registering and signal processing into the digital domain.
Despite large design efforts were dedicated to the overall optimization of the Front-End amplifier, the most interesting aspects in the implemented pixel Front-End architecture certainly are offered by the proposed synchronous solution for the hit discrimination. Design and optimization of the Front-End comparator are therefore discussed in great details in the following sections.

Architecture choice

Comparators are the most widely used building blocks in mixed-signal integrated circuit design. They are at the core of any A/D data conversion systems. In its simplest form a comparator can be considered as a 1-bit A/D converter as well. Comparators are also key components for other applications, such as data transmission and switching power regulators. Due to the importance and widespread use of these circuits, extensive theoretical analyses and a large variety of CMOS circuit topologies are proposed and discussed in reference textbooks [Razavi 1995, Johns 1996, Gregorian 1999, Allen 2002, Maloberti 2003, Baker 2010, Goll 2015].
Indeed, a few important considerations and practical design aspects must be pointed out when the usage of comparators for the hit discrimination is referred to pixel ASIC applications at the CERN LHC [Campbell 2001, Horisberger 2001, Rossi 2006].
It is a matter of fact that for a given SNR (or equivalently ENC) specification introduced for the Front-End amplifier coupled to its sensor, ultimate performance in terms of hit detection efficiency and low noise occupancy rely on the quality of the discriminator. On the one hand, the threshold value must be set low enough to maximize the detection efficiency, but well above the noise floor in order to keep fake hit events at negligible levels. Comparator resolution is therefore of primary importance. On the other hand, the time response of the discriminator in combination with the rise time of the signal generated by the Front-End amplifier is a crucial aspects for pixel systems at the LHC. Only triggered data (zero-suppressed) can be readout and transmitted to counting rooms for the on-line reconstruction, thus requiring time-stamping and pipelined buffering for the whole trigger latency. Hence the hit generation must occur in less then 25 ns in order to associate an event with the correct beam crossing.
Speed and precision are therefore key input specifications in the comparator design. However, due to the severe power budget constraint defined for the pixel upgrade, the speed-precision trade off is exacerbated. In practice, with 2-2.5 μA bias currents already allocated in the Front-End amplifier, no more than 1-1.5 μA can be used for the comparator design in order to keep the total analogue power dissipation at a maximum level of 5-6 μW/pixel. Nevertheless, the discriminator must be able to unambiguously produce a hit pulse within 25 ns when a nominal minimum detectable charge of 1000 e^- is collected by the sensor. As a result of low-threshold and in-time threshold requirements, design and optimization of the Front-End discriminator become more challenging and the choice of a baseline circuit topology plays a fundamental role.

[4] The comparison can be performed either in current mode or voltage mode. According to the proposed solution threshold will always refer to a voltage quantity.

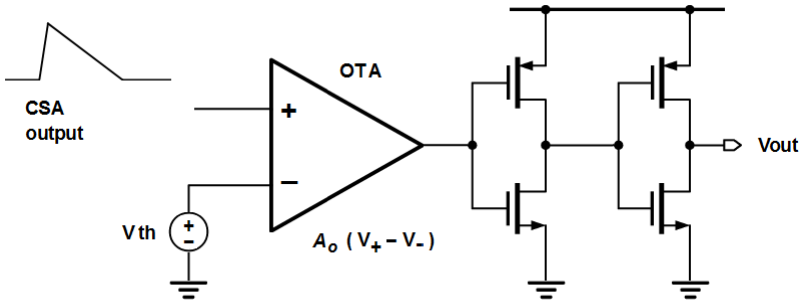


Figure 2.59: Continuous-time voltage comparators are usually implemented by means of a high open-loop gain differential amplifier coupled to CMOS inverters. The actual number of inverters depends on the required drive strength and fanout. This is a common choice adopted for the design of the hit discriminator in a Front-End chain.

Continuous-time comparator architectures represent a standard choice for applications involving radiation detectors. The usage of asynchronous solutions is primary motivated by the fact that the arrival time of the events is usually unknown. Most of Front-End discriminators for particle physics, spectroscopy and imaging are therefore implemented as high-gain Operational Transconductance Amplifiers (OTAs) in open-loop configuration. As depicted in Figure 2.59, the OTA provides a small-signal gain A_o and amplifies the voltage difference $(V_+ - V_-)$ at comparator input nodes. Because of its high-gain, the amplifier output voltage saturates to some V_{OH} close to the power supply or V_{OL} close to the ground rail according to the sign of the input difference. CMOS output inverters are then used to obtain rail-to-rail digital levels and ensure adequate drive strength and fanout for subsequent logic stages. Since the OTA works in open-loop, stability is not an issue and internal frequency compensation is not required. Popular circuit topologies adopted in the pixel ASIC community are telescopic or folded cascode OTAs, two-stage OTAs or even a single MOS differential pair coupled to inverters. More complex and specific continuous-time solutions have been proposed for precise timing applications instead, where zero-crossing discriminators (ZCDs) and constant-fraction discriminators (CFDs) are employed.
The overall performance offered by OTA-based comparators in terms of speed, resolution and power consumption are actually quite poor with respect to circuit architectures extensively adopted for commercial applications. On the one hand, both small signal bandwidth and slew-rate (SR) contribute in determining the propagation delay, hence the maximum speed. In particular, if the differential input presented to the comparator is quite large, the amplifier becomes fully unbalanced and the total load capacitance seen at the OTA output node is charged and discharged with the maximum available current in the circuit. Therefore speed is limited by the circuit SR and the OTA output voltage can only change with a constant slope. SR limitations can potentially degrade speed performance at unacceptable levels. In that cases, dedicated internal clipping structures such as diode-connected clamp devices must be used to limit charge and discharge processes at sensitive nodes. On the other hand, comparator resolution is only determined by the open-loop gain of the amplifier, because the minimum input that can be resolved (neglecting noise and offset contributions) is ideally equal to $(V_{OH} - V_{OL})/A_o$. Thus high resolutions necessary require high gains, at the cost of larger power dissipations and slower time response. Low-power techniques such as cascodes and positive feedback can be certainly used to increase gain and speed, as performed for example in class AB hysteresis comparators.

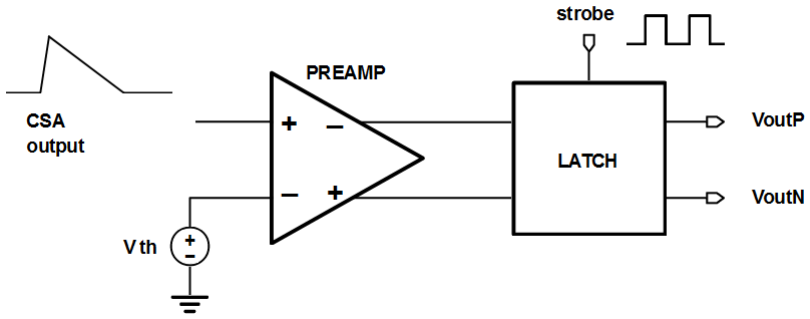


Figure 2.60: Simplified architecture of the proposed synchronous pixel Front-End discriminator. For fast and high-resolution performance a low-gain differential amplifier (preamplifier) is coupled to a positive feedback latch stage. Pixel-to-pixel threshold variations are then minimized by using an autozeroed solution (not shown in figure).

During preliminary Front-End architecture explorations, a few continuous-time solutions have been scaled in 65 nm, derived from past designs of the INFN Torino VLSI Design Laboratory already implemented in higher technology nodes. Nevertheless, hit detection efficiency performance have not been found satisfactory. Moreover, power dissipation values were unrealistic.
At the end, in order to take advantage of speed offered by a 65 nm CMOS process and explore innovative solutions for the hit discrimination and the charge encoding, the final choice has been to employ an track-and-latch autozeroed comparator. With such a solution, synchronous operations have been introduced for the analogue Front-End system.
It is well known in fact that high-speed, high-resolution comparators can be implemented by cascading a certain number of low-gain and wide-bandwidth differential amplifiers followed by a regenerative latch stage. During a first *track phase*, voltage differences at comparator input nodes are amplified and presented to the latch stage. In a second *latch phase* instead, voltage differences at latch input nodes are further rapidly amplified up to opposite rail-to-rail digital levels due to the usage of positive feedback operations. Hence discrete-time comparators overcome the speed-precision trade off by separating in two different moments a preliminary signal amplification and the actual final comparison. Positive feedback in the latch stage must be enabled only at the time of the comparison and reset after the comparator decision has been registered. Therefore, depending on the actual circuit topology one or more appropriate digital clock waveforms are necessary to trigger comparator operations and periodically enable and reset the latch after regenerative transients have completed. As a result, the comparator becomes a true clocked mixed-signal system.
With respect to continuous-time solutions, synchronous architectures offer better performance in terms of speed, resolution and power consumption. Interesting, discrete-time comparators are systematically employed in semiconductor industry for the implementation of high-speed and high-precision A/D data converters. As a matter of fact, they represent the standard choice adopted for any high-performance design and an impressive amount of literature is focused on synchronous architectures.

A simplified block diagram of the proposed track-and-latch Front-End discriminator is shown in Figure 2.60. The circuit uses a single stage low-gain ($A_o \approx 6$) differential amplifier (referred to as preamplifier in the following) coupled to a positive feedback latch stage in order to obtain fast response and high-resolution performance for the hit discrimination. A CMOS *strobe* digital control signal is used to enable/reset positive feedback in the latch. As will be shown, depending on the latch decision the differential outputs *VoutP* and *VoutN* settle to rail-to-rail complementary logic levels at each *strobe* cycle.

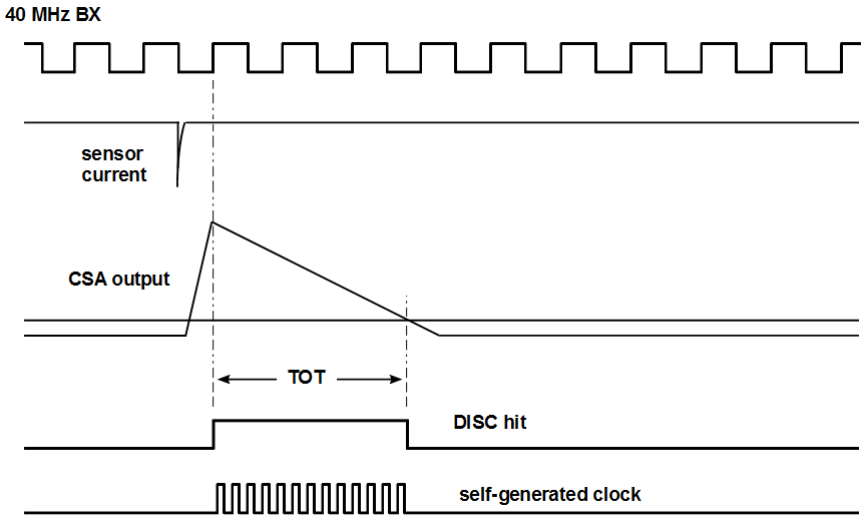


Figure 2.61: Timing diagram for analogue Front-End operations synchronized with a bunched machine activity.

The final choice of a synchronous solution for the pixel Front-End is supported by some important considerations. From a pure design point of view, the usage of a latched comparator makes easier to meet low-power constraints along with low-threshold and speed requirements defined for the pixel upgrade. In fact, large voltage amplifications are achieved by means of positive feedback in the latch stage with negligible static power dissipation (after positive feedback transients, only leakage currents contribute to power consumption). Thus bias currents in the 1-1.5 μA range can be effectively employed for the design of the first low-gain stage without compromising on time response and sensitivity of the overall system. Provided that low-noise performance are guaranteed in the design of the Front-End amplifier, speed and high resolution are then obtained at minimum power dissipation using positive feedback, such that small voltage differences at comparator input nodes can be resolved. As a result, very low charge-induced signals of 8-10 mV above the nominal threshold are pre-amplified and then presented at latch input nodes, which in turn thanks to regenerative operations can unambiguously produce a hit information for the event in less than about 1 ns. From an experimental point of view, synchronous operations are certainly suitable in those applications in which particles arrive at known times. This is the case of colliding experiments and in perspective of the CMS pixel upgrade at HL-LHC a synchronous Front-End approach offers some important and fairly appealing features that naturally meets the requirements introduced by a bunched environment. In fact, particles at the LHC are produced every 25 ns at each beams intersection. Due to the the position of pixel detectors closest to the beam spot, the arrival time of particles traversing the pixelated layers (a few hundreds ps after the bunch interaction) de facto corresponds to the bunch crossing. It is therefore possible to distribute a 40 MHz clock to all pixels and synchronize the hit generation in the analogue Front-End with machine operations.

Assuming for instance that a charge signal is collected in a pixel cell shortly after a falling edge of the 40 MHz machine clock, positive feedback in the latch can be enabled in correspondence of the next rising edge in the same clock cycle, as depicted in Figure 2.61. The discriminator samples the CSA analogue output around its nominal peaking time of 12.5 ns and as derived from CAD simulations, the latch decision is promptly ready in less than about 1 ns after positive feedback has been enabled. Hence the comparator delay becomes a well defined and constant quantity roughly equal to the peaking time of the Front-End amplifier and no time-walk issues can affect the time-stamp assignment for the event. Moreover, noise-induced leading edge uncertainties (jitter) that usually exhibit time invariant comparators disappear with such an approach.
Certainly due to the large number of channels foreseen in the final readout chip, the simultaneous triggering of all pixel cells both in the analogue and digital parts can become an issue in terms of noise and power dissipation. Nevertheless, adequate substrate isolation can be achieved by proper layout choices and usage of Deep N-Well (DNW) structures, whereas current peaks can be avoided by distributing the master clock along pixel columns with non-zero skew. This is already performed in pixel ASICs in which a clock signal is propagated to all digital pixel cells [Llopart 2007, Hemperek 2009, Valerio 2014, Gaioni 2015].
Within the pixel ASIC community, a clocked dynamic comparator has been already prototyped in 65 nm CMOS technology, as reported in [Havranek 2014]. Nevertheless, design specifications are not compliant with performance requirements defined for the CMS pixel upgrade at HL-LHC. Indeed, the proposed track-and-latch architecture has been equipped with two additional features. On the one hand, discrete-time comparators can naturally include autozeroing techniques for the offset compensation, as extensively adopted in industry. Pixel-to-pixel threshold variations have been therefore minimized by means of capacitors using an autozeroed scheme, without the need of a on-pixel D/A converter for digital trimming. The absence of a local D/A converter introduces fundamental advantages in perspective of pixel operations in a harsh radiation environment.
On the other hand, without doubts the most interesting feature offered by a latched comparator resides in the possibility of turning it into a local oscillator by means of asynchronous control logic. As will be discussed in great details in fact, the implemented latch stage can become a compact voltage-controlled oscillator (VCO) and in case a particle hit has been detected above the threshold voltage, fast time-over-threshold (TOT) digitizations with 5-8 bits resolutions can be accomplished using on-pixel self-generated clock signals. As derived from CAD simulations, up to GHz frequencies can be achieved thanks to speed performance offered by a 65 nm CMOS process.

In the following, practical implementation of the Front-End discriminator and simulation results are discussed. Both preamplifier and latch stages are described at the transistor level and most important circuit topology and transistor sizing choices are justified. A short review of positive feedback in CMOS circuits is given as well. After preamplifier and latch stages design, the threshold adjustment by means of an auotzeroing technique is described. Eventually, the feasibility of performing fast TOT encoding by turning the latched comparator into a local oscillator is discussed. A description of the necessary on-pixel digital control logic designed to support discriminator operations is given in Section 7.7 instead.

Low-gain preamplifier

The first stage of the proposed Front-End discriminator is a low-gain fully-differential amplifier. During normal operations, it receives as inputs the analogue signal from the CSA and a global threshold voltage. Differences at input nodes are amplified and then presented to inputs of the subsequent regenerative latch stage. The choice of fully-differential operations is mandatory in order to guarantee necessary immunity against common-mode noise and voltage supply variations introduced by the foreseen intense switching activity in the latch itself and in the pixel digital part. A dedicated calibration period is required for the autozeroing instead. As discussed later in the chapter, in the offset compensation phase both preamplifier inputs are disconnected and shorted to a common-mode reference voltage using switches.

For preliminary design studies, a basic NMOS differential pair with PMOS diode-connected loads has been employed in order to derive first constraints on transistor sizes and quantify mismatches and process variation effects. A schematic of the circuit is presented in Figure 2.62.
The total small-signal resistance seen at output nodes is given by the parallel combination between the output resistance $1/g_{ds12}$ of input transistors and the small-signal resistance $1/(g_{m34}+g_{ds34})$ that exhibit diode-connected loads. Hence the differential gain of the circuit is equal to

$$|A_o| = \frac{g_{m12}}{g_{ds12}+g_{ds34}+g_{m34}} \approx \frac{g_{m12}}{g_{m34}}$$

Adequate performance can be achieved with $|A_o| \approx$ 5-10 and proper transistor sizing is required to satisfy $g_{m12} > g_{m34}$. In order to meet the power budget constraint, 1 μA tail current I_{SS} has been used to bias the stage, resulting into 1.2 μW static power contribution at 1.2 V supply voltage[5]. The channel length of transistor M5 has been increased to 2 μm in order to push the device in strong inversion, ensuring better matching in the current mirror used to bias the stage. With a 150-200 mV drain-source voltage across M5 and the input common-mode set to 500-600 mV, input devices M1 and M2 are biased in deep moderate inversion, with underdrive voltages in the 200-250 mV range. Hence the transconductance of input devices is essentially determined by the 500 nA bias current flowing in each branch of the pair. As already reviewed in fact, in weak inversion the g_m/I_D ratio is roughly constant, whereas in moderate inversion g_m has only weak dependences on transistor sizing. Thus, wider input transistors do not increase significantly g_{m12} values but only contribute to parasitic input capacitances. As a baseline choice, $(W/L)_{12} = 10$ has been adopted, resulting into a $g_{m12}/I_{D12} \approx 26$ V^{-1}. Certainly larger aspect ratios improve matching. Transconductance values g_{m34} must be therefore minimized. As already discussed for the Front-End amplifier design, this is accomplished by increasing the channel length L of PMOS devices above 1 μm such that their aspect ratio (W/L) is minimized, at the cost of a larger output capacitance.

Preliminary simulation results demonstrated that differential gains larger than 3 would not have been feasible under the above described bias conditions. Furthermore, due to the large threshold values that exhibit MOS transistors in the chosen 65 nm process, reduced input common-mode voltages can be only obtained by means of low-Vt devices in DNW configuration, such that the threshold voltage can be reduced avoiding the body effect. The total output-referred DC offset determined from MC simulations was about 8 mV RMS.

[5] As already mentioned for the design of the Front-End amplifier, current mirror branches are external to the pixel cell, hence they not contribute to the compute of the total power dissipation at the pixel level.

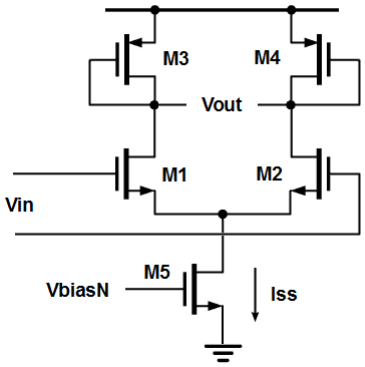


Figure 2.62: Low-gain differential amplifier implemented with a NMOS input differential pair and diode-connected PMOS loads.

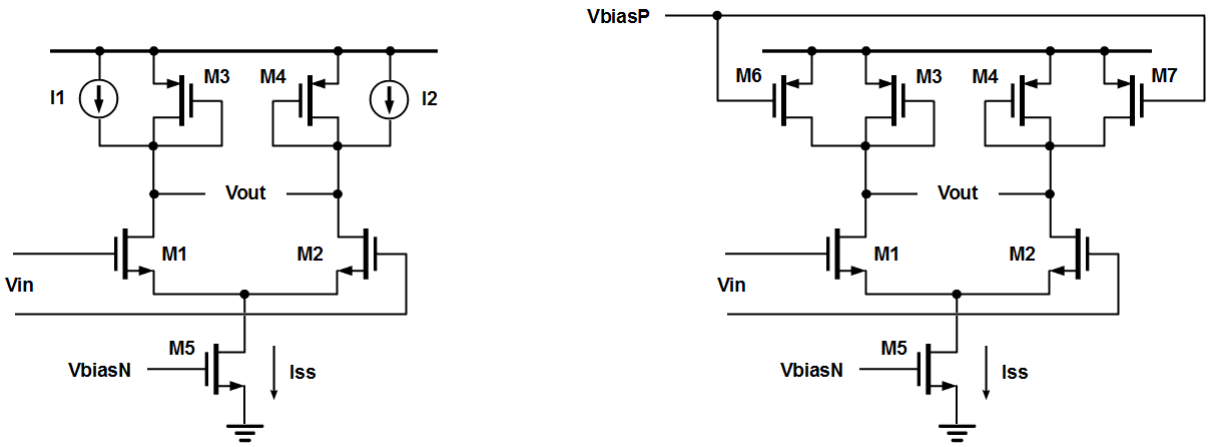


Figure 2.63: Gain enhancement using lateral shunt current sources.

In order to increase the gain, in the final implementation shunt current sources have been added at drain nodes of NMOS input devices. This is presented in Figure 2.63. Other solutions such as positive feedback and cascodes could be adopted as well. As already performed in the telescopic cascode of the charge integrator, thanks to current splitting PMOS transconductances g_{m34} can be reduced by decreasing the current flowing in M3 and M4 without modifying the 500 nA bias current in input transistors M1 and M2. Moreover, auxiliary currents allow to optimize the DC output node voltage if required. Assuming square-law devices due to strong inversion operations for M3-M4 we can write

$$g_{m34} \;=\; \sqrt{2\mu_p C'_{ox} \left(\frac{W}{L}\right)_{34} I_{D34}}$$

With a total tail current of 1 μA the choice

$$I_1 \;=\; I_2 \;=\; 0.8 \left(\frac{Iss}{2}\right) \;=\; 400 \text{ nA}$$

reduces the bias current in M3-M4 from 500 nA to 100 nA, without modifying transcondutance values g_{m12} of NMOS input transistors. Assuming same device sizing, it is expected that the differential gain should increase by a factor

$$\sqrt{\frac{500 \text{ nA}}{100 \text{ nA}}} \;\approx\; 2.2.$$

Basic square-law analytical calculations are in agreement with the simulated Bode plot presented in Figure 2.64. The resulting DC open-loop gain roughly doubles to about 6, with a BW of 45 MHz. Figure 2.65 shows instead the distribution of the DC voltage difference at the preamplifier output nodes across 100 MC iterations. The output-referred offset slightly increases to about 10 mV RMS due to additional mismatches in lateral shunt current sources. As discussed later in the chapter, such an offset contribution is completely cancelled in the implemented autozeroed solution.
A sample transient simulation for differential output voltages is presented instead in Figure 2.66. The final optimized transistor sizing is summarized in Table 2.6.

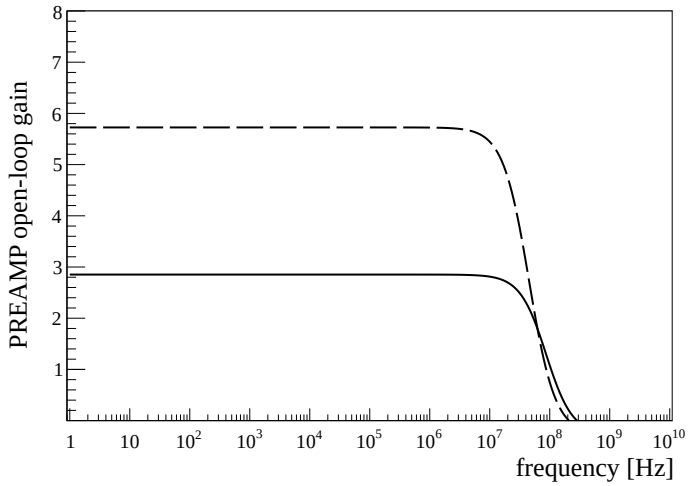


Figure 2.64: Simulated AC gain for simple diode-connected PMOS loads (solid) and with gain enhancement by means of lateral current sources (dashed).

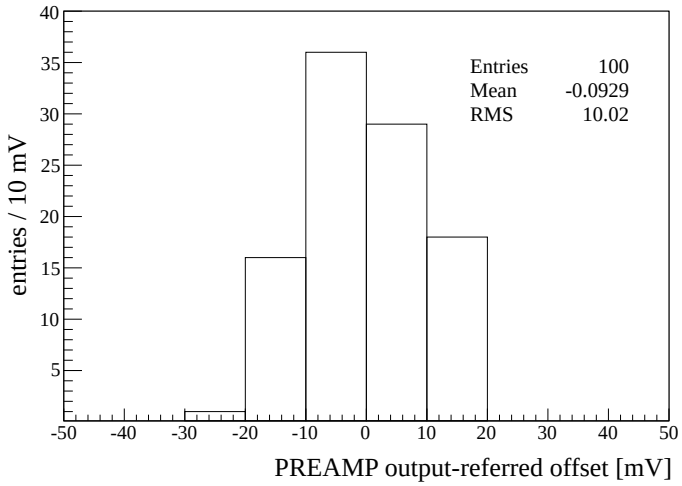


Figure 2.65: Distribution of random DC voltage differences at preamplifier output nodes across 100 MC iterations assuming a common-mode input voltage of 500 mV. The resulting output-referred offset is 10 mV RMS.

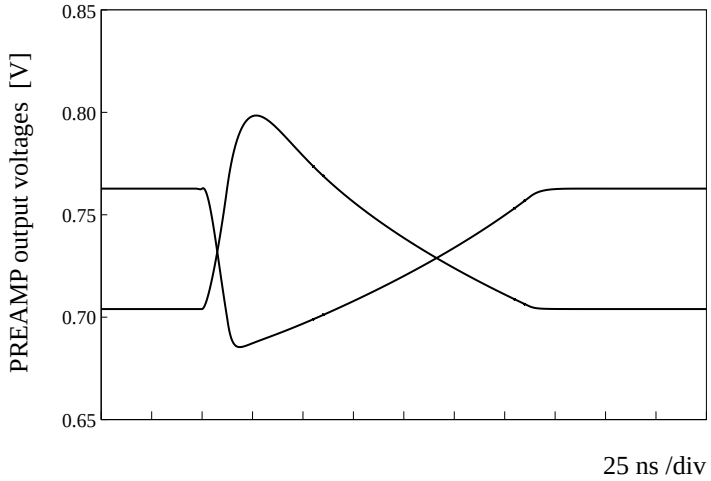


Figure 2.66: Simulated preamplifier differential output voltages assuming 1 ke^- input charge.

Device	W/L [μm/μm]
M1-M2	10/1
M3-M4	0.3/1.5
M5	1.42/2
M6-M7	0.6/1.5

Table 2.6: Final optimized transistor sizing for the synchronous comparator low-gain preamplifier.

Positive feedback latch

At the core of the Front-End discriminator resides a dynamic latch stage. It tracks preamplifier differential outputs and generates full swing CMOS logic levels *VoutP* and *VoutN* according to the sign of the voltage difference presented at comparator input nodes. An external digital control signal *strobe* is required to periodically enable and reset positive feedback in the circuit, hence synchronous operations are introduced in the analogue Front-End system.
Before a description of the actual latch transistor level implementation, regenerative behaviour of CMOS cross-coupled inverting amplifiers are shortly reviewed hereafter [Razavi 1995, Johns 1999, Allen 2002, Kang 2003, Figueiredo 2009].

Positive feedback[6] is one of the most popular techniques adopted to increase speed and gain performance in low-power and low-voltage analogue integrated circuit design. Same concepts apply to CMOS digital circuits too, where regenerative operation represents the fundamental expedient to achieve memory in sequential circuits such as S/R and D-type latches, FlipFlops, registers and Static-RAM (SRAM) cells. As shown in Figure 2.67, positive feedback is obtained by cross connecting two identical inverting amplifiers. According to the proposed latch implementation we refer to a couple of CMOS inverters used as voltage amplifiers. In digital electronics this configuration is the well known basic bistable element. Single NMOS or PMOS cross-coupled pairs are also widely employed in latch circuit topologies.
When used as an analogue amplifier, the input bias voltage of a CMOS inverter is chosen such that the DC output voltage sits at an acceptable level between power and ground rails. In a perfectly symmetric design this value is half the supply voltage and appropriate transistor sizing [Rabaey 2003] ensures that $V_{in} = V_{out} = V_{DD}/2$. This can be appreciated in the simulated DC voltage transfer characteristic (VTC) for minimum channel length devices and different W_P/W_N ratios, as reported in Figure 2.68. Small input voltage variations superimposed to the quiescent point are therefore amplified and inverted to the output. Thanks to both NMOS and PMOS contributions to the effective total transconductance $g_m = g_{mN} + g_{mP}$ of the system, for a given total load capacitance C_L at the output node the CMOS inverter provides the largest small-signal gain and GBW product $g_m/2\pi C_L$ achievable with a common-source amplifier [Sansen 2006].
In the basic CMOS cross-coupled pair the output voltage of the first inverter determines the current flowing into the second one. Conversely, the output of the second inverter controls the current in the first one. The regenerative behaviour of this structure can be therefore inspected as follows. Let suppose that both X and Y nodes are initially tied to the same DC level at about half the supply voltage. Since $V_X = V_Y$ the system is balanced and the same current flows in each side of the circuit. Nevertheless, this is an unstable bias condition. Due to the large small-signal gain of both inverters any small perturbation at internal nodes can fully unbalance the structure. As an example, suppose that at time $t = 0$ the circuit exhibits a small voltage difference of a few mV such that $V_X > V_Y$. Since the voltage at node X is higher, a larger current must flow in the right branch. Hence V_Y further decreases, the current in the left branch diminishes as well and at the end V_X is pulled up. As a result, the initial voltage difference $V_X - V_Y$ has been amplified. As shown in transient simulations of Figure 2.69, output voltages rapidly diverge and the process continues until V_X and V_Y saturate to power and ground rails respectively. Neglecting leakage effects, no currents flow after rail values have been reached, thus the circuit dissipates only during the regenerative transients. Provided that no external large perturbations are subsequently introduced at internal nodes and that power is supplied to the circuit, after V_X and V_Y settle to complementary rail values the system holds. In fact, the basic bistable element is at the core of any static memory circuits in digital circuits. Certainly, in practical CMOS latch implementations appropriate switches must be introduced in order to periodically enable and reset positive feedback operations.

[6] Also referred to as *negative resistance* in analogue integrated circuit design.

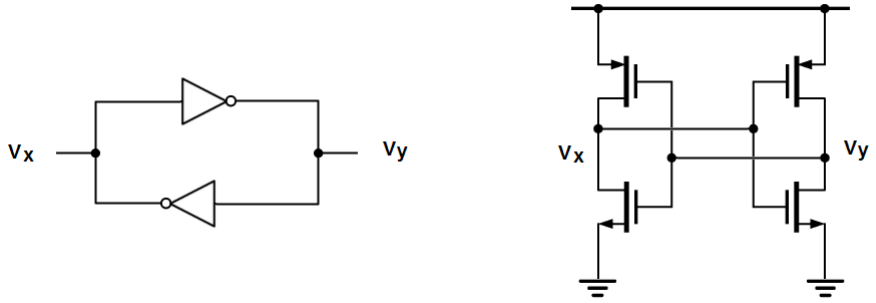


Figure 2.67: Positive feedback achieved by means of cross-coupled inverters (left) and practical CMOS implementation (right). It is supposed that at $t = 0$ output nodes exhibit a small non-zero voltage difference V_{XY0} around a quiescent point $\approx V_{DD}/2$.

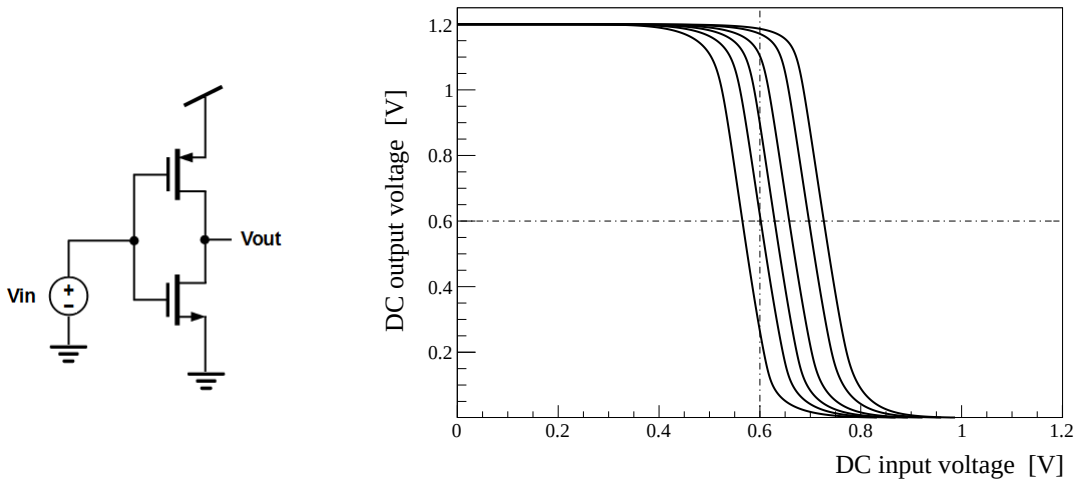


Figure 2.68: DC characteristics for a 65 nm CMOS inverter assuming minimum channel length, $W_N = W_{\min}$ and increasing the channel width W_P of the PMOS transistor.

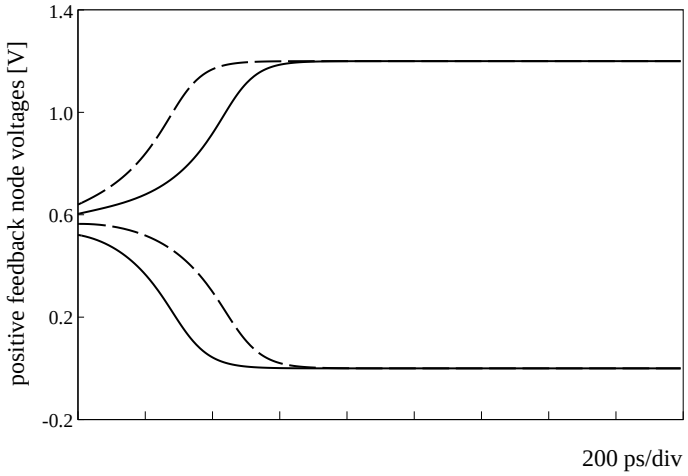


Figure 2.69: Transient simulations of positive feedback node voltages V_X (solid) and V_Y (dashed) for +10mV and -30 mV initial voltage differences V_{XY0} around a 600 mV quiescent point.

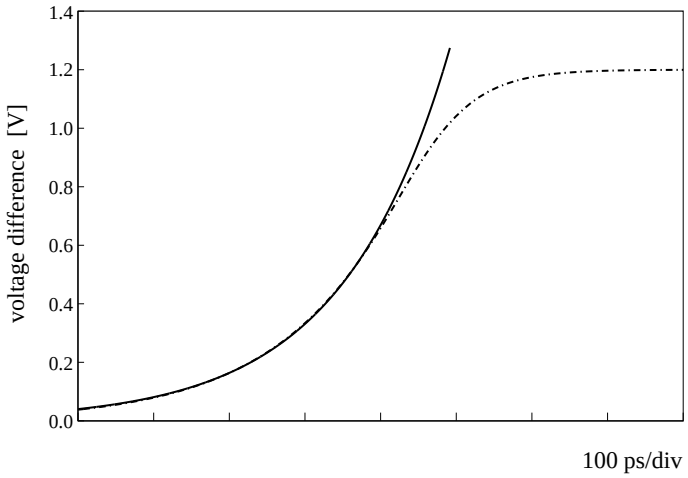


Figure 2.70: Simulated voltage difference $V_{XY} = V_X - V_Y$ as a function of time (dashed) with +40 mV initial condition and comparison with an exponential fit $A\,e^{B\,t}$ (solid). Before saturation the foreseen exponential behaviour is verified. The latch time constant τ_L can be extracted from fit results as $\tau_L = 1/B$.

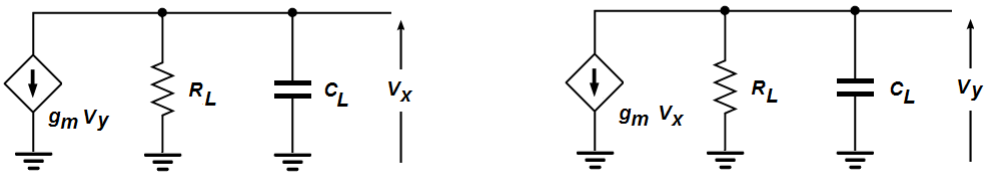


Figure 2.71: Small-signal model for a system of cross-coupled inverting voltage amplifiers.

Analytical calculations for the evolution of the voltage difference $V_{XY}(t) = V_X(t) - V_Y(t)$ as a function of time can be derived by referring to the cross-coupled inverters small-signal equivalent circuit presented in Figure 2.71. CMOS inverters are supposed to be identical and each stage is a single-pole inverting voltage amplifier with a transconductance $g_m = g_{mN} + g_{mP}$, output resistance $R_L = 1/(g_{dsN} + g_{dsP})$ and total load capacitance C_L. Thus

$$A_o = -g_m \; R_L \qquad \text{and} \qquad A(s) = \frac{A_o}{1 + s\tau_o}$$

being $\tau_o = R_L C_L$ the time constant associated to the low-pass frequency behaviour of the system. By applying the Kirchoff current law in the time domain at output nodes X and Y we can write

$$g_m V_Y + \frac{V_X}{R_L} + C_L \frac{dV_X}{dt} = 0$$

$$g_m V_X + \frac{V_Y}{R_L} + C_L \frac{dV_Y}{dt} = 0$$

that rearranged lead to a pair of coupled differential equations

$$\tau_o \frac{dV_X}{dt} + V_X = -A_o V_Y$$

$$\tau_o \frac{dV_Y}{dt} + V_Y = -A_o V_X$$

Subtracting the second equation from the first one we finally obtain

$$\tau_o \frac{dV_{XY}}{dt} = (A_o - 1) \; V_{XY}$$

As a result, for a given initial condition $V_{XY0} = [V_X - V_Y]_{t=0}$ the time evolution is

$$V_{XY}(t) = V_{XY0}\ e^{\,t/\tau_L}$$

where

$$\tau_L = \frac{\tau_o}{A_o - 1}$$

represents the *latch time constant* of the system. Thus, the voltage difference at positive feedback nodes increases as a function of time with an exponential behaviour until amplifiers saturate and voltage gains drop to zero. This can be appreciated in the simulation presented in Figure 2.70. As one can see, before saturation the predicted exponential behaviour is well confirmed and the time constant can be easily extracted from an exponential fit performed on simulated data.

The most important aspect in the latch design is the time required to produce full logic levels after regeneration begins in the circuit. Both the time constant τ_L and the initial condition V_{XY0} contribute in determining the effective amount of time needed to reach full-swing voltages. Let suppose that a certain voltage difference V_{XY1} is necessary before it can be interpreted as a logic 1. If $T1$ is the time required to reach such a value,

$$V_{XY1} \ = \ V_{XY0}\ e^{\,T1/\tau_L}$$

we can immediately derive a relationship between $T1$ and V_{XY0} as

$$T_1 \ = \ \tau_L \ \ln\left(\frac{V_{XY1}}{V_{XY0}}\right) \ = \ \frac{\tau_0}{A_o - 1} \ \ln\left(\frac{V_{XY1}}{V_{XY0}}\right)$$

Certainly, the time required by V_X and V_Y to reach power and ground rails decreases by applying a larger initial voltage difference V_{XY0}. This represent the primary motivation for the usage of a preamplifier stage in a track-and-latch comparator design. In fact, in order to achieve high-speed and high-precision voltage comparisons, it is preferable to present to the positive feedback stage a sufficiently large initial difference V_{XY0} by means of a low-gain differential amplifier.
A set of transient simulations for the voltage difference as a function of time assuming minimum size CMOS inverters and different initial conditions V_{XY0} is presented in Figure 2.72. For the same testbench, Figure 2.73 shows the time required by V_{XY} to reach 1 V as a function of the initial condition. Simulated data have been then compared with the logarithmic small-signal prediction, extracting the voltage gain A_o from the CMOS inverter DC characteristic[7] and the latch time constant from an exponential fit over simulated data. As one can see, simulation results are in good agreement with the analytical model.

[7] Strictly speaking, 1 V is already below the lower limit of the noise margin (NM) defined by the simulated DC characteristic for a CMOS inverter in 65 nm at 1.2 V supply voltage. This can be appreciated in Figure 2.68.

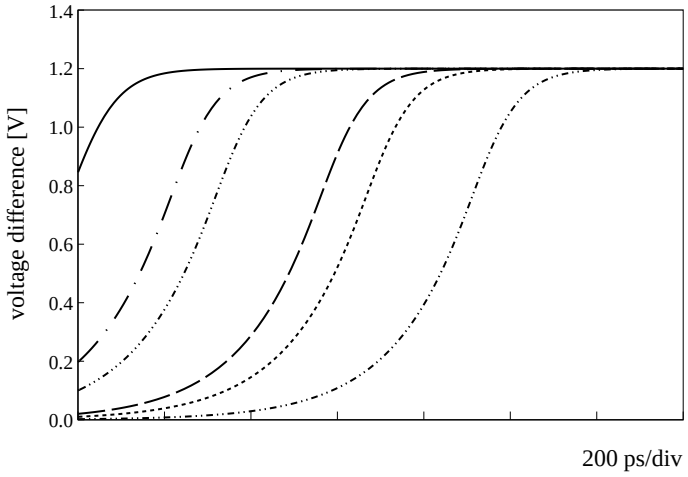


Figure 2.72: Simulated voltage difference $V_{XY} = V_X - V_Y$ as a function of time for different initial conditions V_{XY0} and minimum size CMOS inverters. From right to left: 1 mV, 5 mV, 10 mV, 50 mV, 100 mV and 500 mV.

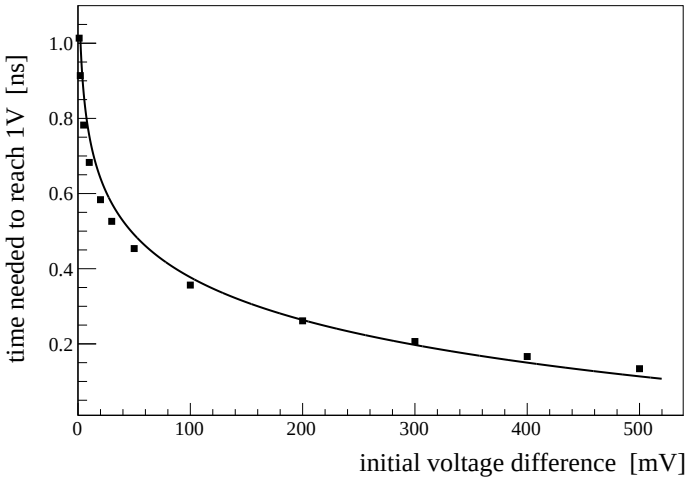


Figure 2.73: Simulated time required to reach 1 V as a function of the initial voltage difference (points) and comparison with the small-signal model analytical prediction (solid curve).

Beside a large enough initial voltage difference, for a fast response the latch time constant τ_L must be minimized. Since typical values for the DC gain of a CMOS inverter are $A_o \geq 10$ we can assume

$$\tau_L \approx \frac{\tau_o}{A_o} = \frac{C_L}{g_m}$$

in order to derive practical transistor sizing considerations. The latch time constant is therefore roughly equal to the inverse of the unity-gain frequency of each CMOS inverter/amplifier. On the one hand, if no extra loading is presented to output nodes then C_L is mainly due to gate-source capacitances, which are proportional to the gate area $W\,L$. On the other hand, with a quiescent point at about half the supply voltage $V_{DD}/2$, a CMOS inverter used as an amplifier works in strong inversion, hence the effective transconductance g_m can be derived assuming well-known square-law expressions,

$$I_D = \frac{1}{2}\,\mu\,C'_{ox}\,\frac{W}{L}\,(V_{GS} - V_{TH})^2 \qquad \text{and} \qquad g_m = \frac{\partial I_D}{\partial V_{GS}} = \sqrt{2\,\mu\,C'_{ox}\left(\frac{W}{L}\right) I_D}$$

Upon these assumptions we can further derive

$$\tau_L \propto \frac{WL}{\sqrt{(W/L)\;I_D}} = \sqrt{\frac{WL^3}{I_D}}$$

which in turn lead to $\tau_L \propto L^2$ due to the linear dependence of the drain current with the transistor aspect ratio (W/L). As expected, the choice of minimum channel length devices maximizes the speed. Interesting, for a given equilibrium bias condition $V_X = V_Y \approx V_{DD}/2$ and assuming that $L = L_{\min}$ is adopted, the latch time constant primarily depends on process technology parameters such as carrier mobilities and gate capacitance for unit area included in the implied proportional constant, but not on device sizing. Certainly, an extra capacitive load connected to positive feedback nodes decreases the speed, with a foreseen linear increase of τ_L.
In perspective of latch design, transient simulations have been performed in order to investigate the effective dependence of the regeneration time constant as a function of extra load capacitance and MOS channel length. Figure 2.74 shows a set of transient simulations for the voltage difference for different values of load capacitance C_L, assuming minimum size inverters and a +10 mV initial condition. Time constants have been then extracted from exponential fits and reported as a function of the load capacitance, as shown in Figure 2.75. The linear dependence is confirmed. In practice, a capacitive load larger than 150-200 fF increases the time constant above 1 ns.
The same study has been performed for the same initial condition and increasing the channel length of CMOS inverters in order to inspect the foreseen quadratic relationship. Simulation results are shown in Figure 2.76. Furthermore, time constant values have been reported as a function of the channel length squared. This is presented in Figure 2.76. The expected dependence $\tau_L \sim L^2$ is well confirmed by a linear characteristic $\tau_L \sim 1/L^2$ if a channel length larger than 1.5-2 μm is adopted. Short channel effects and the much higher complexity in the description of the MOS transistor for submicron channel lengths can justify instead the non-linearity that exhibits the characteristic for $L < 1$ μm.

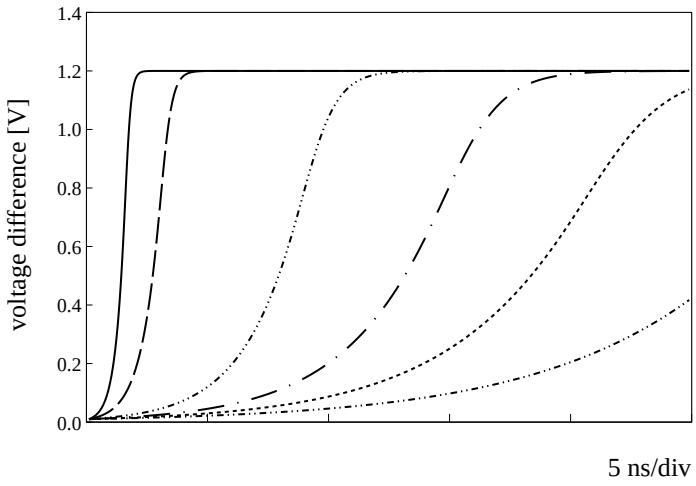


Figure 2.74: Voltage differences as a function of time by increasing an extra load capacitance at the positive feedback nodes, assuming minimum size CMOS inverters and +10 mV initial condition.

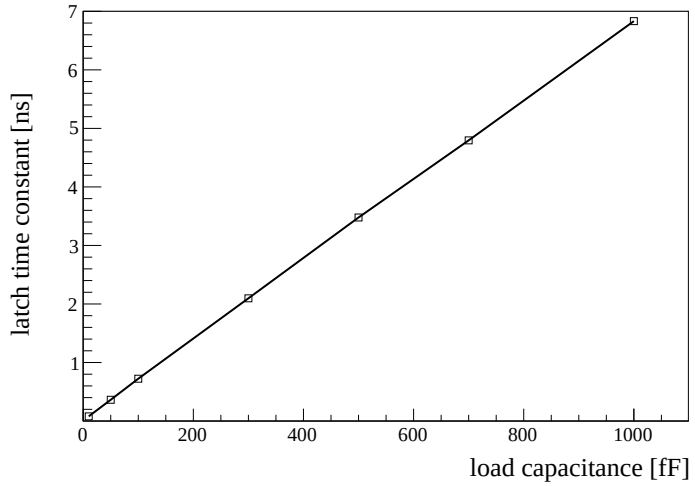


Figure 2.75: Latch time constant τ_L as a function of the total load capacitance C_L.

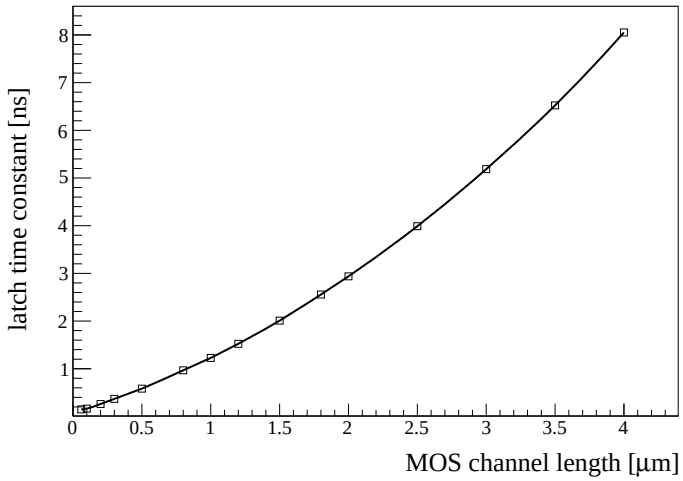


Figure 2.76: Latch time constant τ_L as a function of CMOS inverters channel length L.

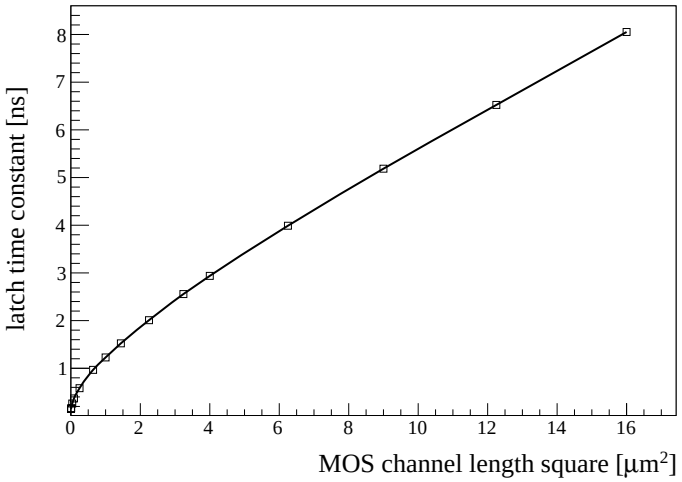


Figure 2.77: Time constant values reported as a function of L^2. The foreseen quadratic relationship $\tau_L \sim L^2$ is well verified for $L >$ 1.5-2 μm. Short channel effects and the much higher complexity in the description of the MOS transistor for submicron channel lengths can justify instead the initial non-linearity in the characteristic.

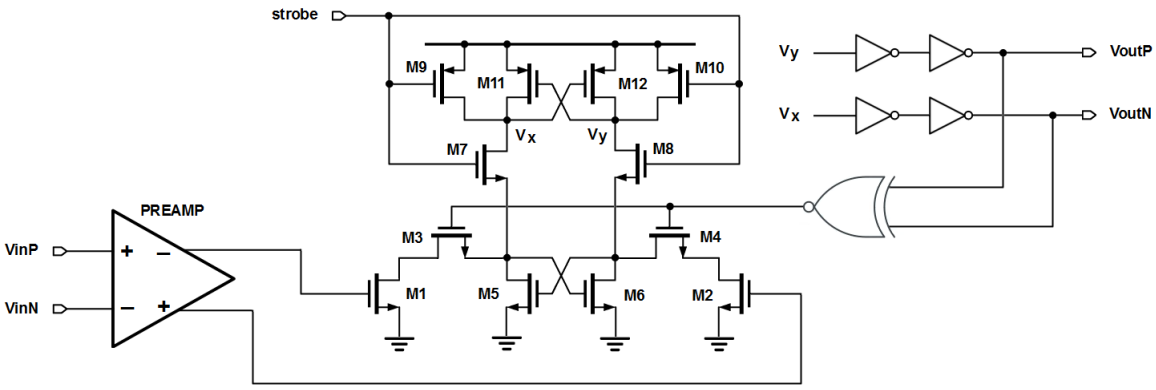


Figure 2.78: CMOS dynamic latch architecture adopted in the proposed synchronous Front-End discriminator.

General considerations and simulation results discussed for the basic CMOS bistable element are now applicable to practical latch transistor-level implementation in the proposed synchronous Front-End discriminator. Within the large variety of circuit topologies described in literature in fact, latches based on cross-coupled inverters offer high performance in terms of speed and resolution. Indeed, power dissipation really depends on the actual latch architecture.
In order to meet the power budget constraint defined for the analogue Front-End the baseline choice has been to use a dynamic latch, according to key advantages of a dynamic approach with respect to less power-efficient class A (static) and class AB solutions [Figueiredo 2006, Goll 2015]. Different circuit topologies have been explored during a preliminary design phase, and the final choice for the overall comparator architecture was essentially driven by power and autozeroing performance derived from CAD simulations. Eventually, the dynamic CMOS latch presented in schematic of Figure 2.78 has been adopted [Huang 2013].

Differential outputs from the low-gain preamplifier stage are presented to latch input transistors M1-M2. The actual coupling to the preamplifier is achieved by means of storage capacitors used to implement an autozeroing technique for the offset cancellation. This is described later in the chapter. MOS switches M7-M8 and M9-10 controlled by an external clock signal *strobe* are used to periodically enable and reset positive feedback in the circuit. Regenerative operations are provided by cross-coupled inverters formed by M5-M6 and M11-M12. CMOS inverters are used to generate full swing output logic levels *VoutP* and *VoutN*. Moreover, they provide necessary drive strength and fanout for subsequent gates in the digital part. Finally, a XNOR gate drives M3-M4 switches according to *VoutP* and *VoutN* logic values. On the one hand, it breaks DC current paths after positive feedback transients have completed, limiting therefore the static power dissipation to leakage contributions only. On the other hand, it reduces charge injection effects on latch inputs through M1-M2 drain-gate capacitances (kickback noise) due to large voltage swings during regenerative transients [Figueiredo 2004, 2006].

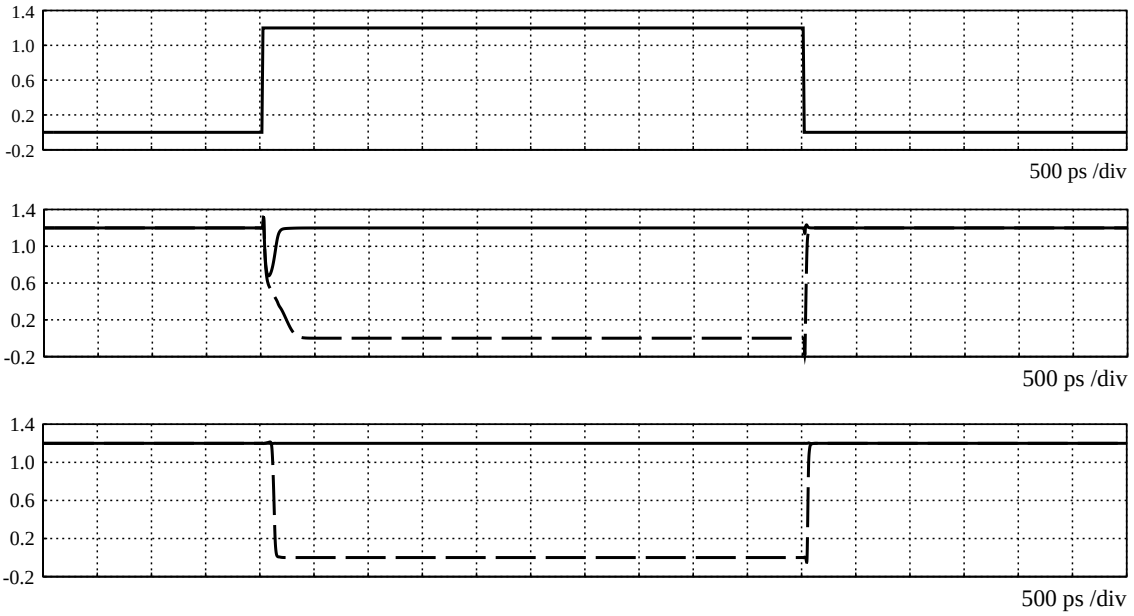


Figure 2.79: Simulated latch transient behaviour. From top to bottom: *strobe*, internal nodes positive feedback voltages and digital outputs *VoutP*/*VoutN*. The simulation was performed with a + 10 mV voltage difference at preamplifier input terminals. Since digital outputs are active-low signals, the dashed curve represents *VoutP*. Thanks to regeneration, the voltage comparison is promptly ready in less than 500 ps.

Latch operating modes are determined by logic levels of the *strobe* signal. A sample transient simulation obtained with final optimized transistor sizing and a +10 mV voltage difference at preamplifier input terminals is presented in Figure 2.79.
In the latch reset phase *strobe* is low. Hence NMOS switches M7-M8 are off whereas PMOS counterparts M9-M10 are on. Positive feedback is disabled, no currents flow in the system and internal nodes V_X and V_Y are pulled up to the the supply voltage through the on-resistance of transistors M9-M10. As a results, buffered outputs *VoutP* and *VoutN* are tied high as well. Since $VoutP = VoutN$ the output of the XNOR gate is high, thus NMOS switches M3-M4 are on.
When *strobe* toggles from low to high instead, NMOS switches M7-M8 turn on and PMOS switches M9-M10 turn off. Internal nodes are therefore disconnected from the supply rail, the latch enters in its regenerative phase and a voltage comparison is accomplished. Indeed, as shown in the simulation a certain amount of time (a few hundreds ps) is required before positive feedback triggers. In fact, when *strobe* goes high and switches M7-M8 turn on, currents start flowing into input transistors M1-M2, discharging X and Y nodes previously forced to the supply voltage. Since cross-coupled PMOS transistors M11-M12 are initially off, both V_X and V_Y decrease together. When the voltage at internal nodes has decreased enough to turn on M11-M12 devices, then regenerative operations begin and V_X and V_Y rapidly withdraw due to the large gain of cross-coupled inverters. After the positive feedback transient has completed, outputs settle to complementary logic value according to the sign of the voltage difference at preamplifier inputs. As one can see, thanks to the usage of the positive feedback mechanism the latch decision is promptly ready within about 1 ns. Neglecting transients and abstracting logic levels from analogue values, if $VinP > VinN$ then *VoutP* goes low, whereas *VoutN* remains high. Latch outputs are therefore active-low signals. Certainly, in order to perform a new comparison, the *strobe* signal must switch again from high to low, such that equilibrium in the fully unbalanced structure is restored by forcing both internal nodes to the supply rail and disconnecting cross-coupled loads from each other.

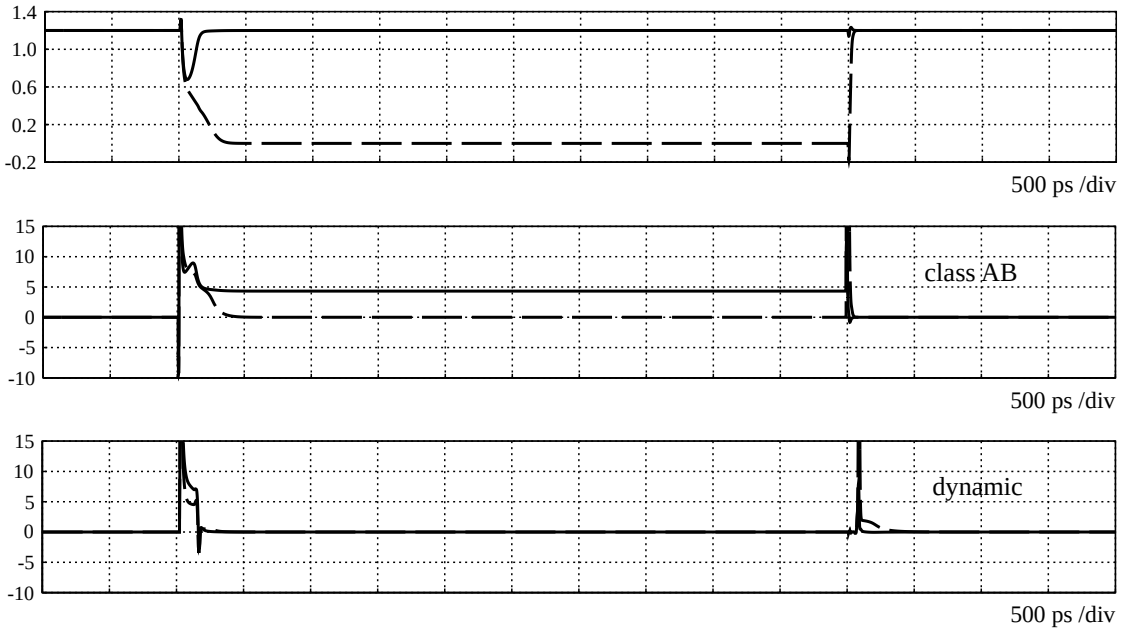


Figure 2.80: Shunt transient currents in class AB or fully dynamic latch operations. From top to bottom: positive feedback voltages and supply shunt currents through M1-M2 without or with NMOS switches M3-M4 driven by a XNOR gate. Without the switches addition, after positive feedback transients have completed a DC current of about 5 μA always flows in the branch in which regeneration has pulled up the internal node to the supply rail (solid).

Note that when latch outputs *VoutP* and *VoutN* have reached complementary values, the output of the XNOR gate toggles from high to low and NMOS switches M3-M4 turn off. Therefore input transistors M1-M2 are disconnected from NMOS cross-coupled devices and no DC shunt current can flow in the input device of the branch in which the regenerative node has been pulled up to the supply voltage. This can be appreciated referring to transient currents reported in Figure 2.80. Interesting, without the usage of M3-M4 switches [Yukawa 1985, Johns 1996] a static current of about 5 μA always shunts from the supply voltage to ground through M1 (if $V_X = V_{DD}$) or M2 (if $V_Y = V_{DD}$) after the positive feedback transient has completed, resulting into class AB latch operations and rising the static power dissipation well above the maximum power budget defined for the analogue Front-End. The usage of a XNOR gate is therefore mandatory to break current paths and achieve a true dynamic behaviour. Relevant power dissipation contributions are limited to regenerative transients only, whereas residual static currents are imputable to negligible leakage effects only. Furthermore, the actual turn-off process in switches M3-M4 is gradual, hence input transistors are effectively disconnected from latch internal nodes before positive feedback nodes exhibit the maximum voltage swing. As a result, the kickback noise towards the inputs is reduced.

The chosen latch architecture offers remarkable advantages in terms of high input impedance, large gain for fast regeneration, rail-to-rail outputs, minimum power consumption and reduced kickback noise. Circuit complexity only slightly increases with respect to more traditional and widely employed fully dynamic latched comparators based on a differential pair input stage and cross-coupled inverters load [Kobayashi 1993].

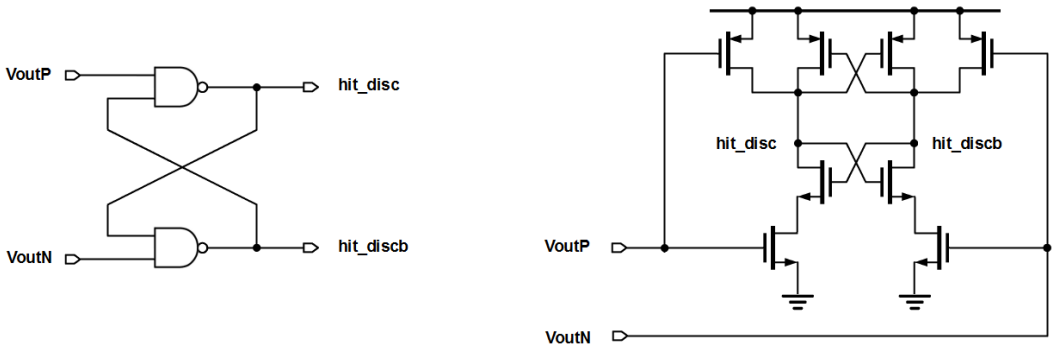


Figure 2.81: NAND-based S/R latch (left) and practical CMOS implementation (right). The digital circuit is used to combine latch differential outputs *VoutP*/*VoutN* and generate a single-ended pulse *hit_disc* fed to the subsequent control logic.

During normal comparator operations[8] the analogue output of the Front-End amplifier is fed to the non-inverting input *VinP* of the low-gain preamplifier, whereas the nominal threshold voltage is applied on the inverting input *VinN*. Provided that a switching signal is propagated to the latch, *VoutP* and *VoutN* are tied high when *strobe* is low (reset phase), and settle to complementary logic values according to the sign of the voltage difference at preamplifier input nodes when *strobe* goes high (regenerative phase). Since latch outputs *VoutP* and *VoutN* are active-low signals, a single-ended discriminator output pulse *hit_disc* is finally obtained by means of a NAND-based S/R latch placed in the on-pixel digital part, as shown in Figure 2.81. As discussed later in the chapter, due to an initial unavailability of a design kit in 65 nm technology with full schematic an layout views of digital components, all required CMOS logic gates have been re-designed from scratch as full-custom cells.

A simulation of the comparator that shows synchronous operations is presented in Figure 2.82. The *strobe* signal is a 40 MHz clock and positive feedback in the latch is enabled every 12.5 ns, corresponding to the nominal peaking time of the Front-End amplifier. Neglecting regenerative transients, if no signal is found above the threshold *VoutN* goes low and *VoutP* remains high. Consequently, *hit_disc* is reset to low. On the contrary, if a particle hit has been detected above the threshold voltage, then *VoutP* goes low whereas *VoutN* remains high. Thus *hit_disc* is set to high. During latch reset phases instead, *strobe* is low, both outputs are forced to the supply voltage and the S/R latch keeps memory of *hit_disc* until a new comparison is enabled. Note that both leading-edge and trailing-edge occurrences of *hit_disc* are always synchronized with a *strobe* positive edge, hence the overall duration of the hit pulse can be only an integer number of clock cycles. Such a solution is therefore suitable either to implement a simple binary readout or to perform slow and low-resolution (3-4 bit) TOT digitizations using the 40 MHz master clock. Speed and resolution can be improved by increasing the frequency of the *strobe* signal. Certainly, in a pixel array it is unrealistic to distribute clock signals with frequencies higher than a few hundreds MHz to all pixel cell. Nevertheless, as already anticipated the latch can be turned into a compact oscillator by means of asynchronous logic. With this technique, locally-generated clock signals up to the GHz level are available in 65 nm CMOS technology and can be used to perform fast TOT encoding up to 8-bit resolution without increasing the power consumption at unacceptable values.

[8] The offset compensation performed by means of an autozeroed solution requires a dedicated calibration cycle, slightly complicating the timing scheme of synchronous operations.

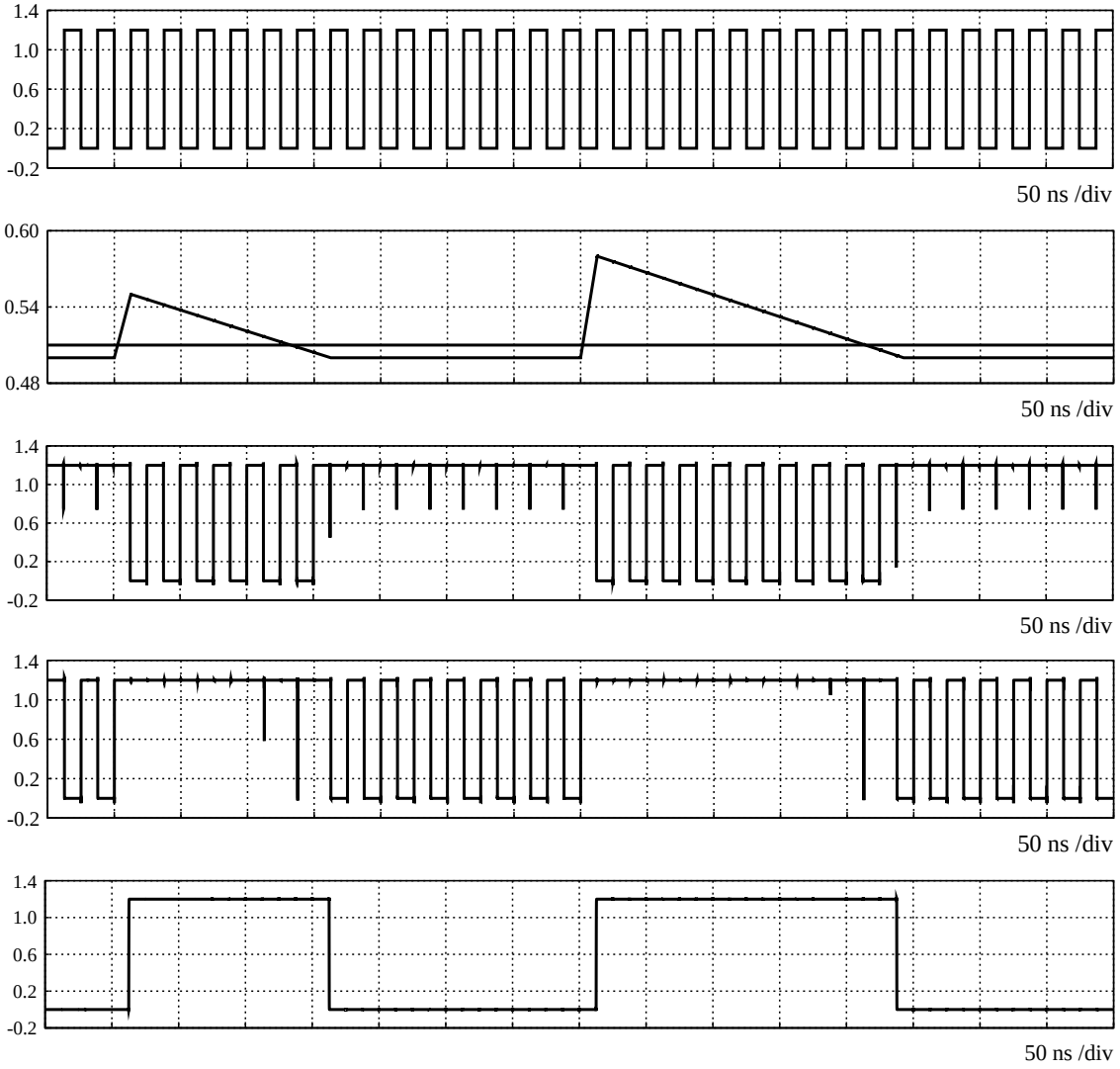


Figure 2.82: Transient simulations for synchronous comparator operations. From top to bottom: 40 MHz *strobe* signal, analogue waveforms and threshold, latch digital outputs *VoutP*/*VoutN* and hit pulse *hit_disc*. For a better visualization, the output of the Front-End amplifier has been simply modelled with a pulse voltage source that mimics a triangular shaping with 12.5 ns peaking time. The latch is enabled at each positive edge of *strobe*. The duration of the hit pulse is therefore an integer number of 40 MHz clock cycles, hence a slow TOT encoding is naturally included in a synchronous solution.

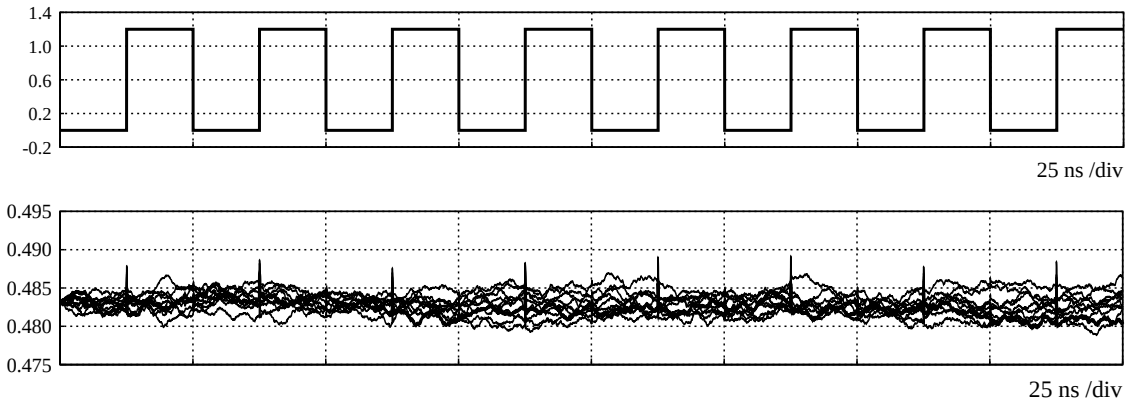


Figure 2.83: Transient noise simulations for the Front-End amplifier coupled to the synchronous comparator. From top to bottom: 40 MHz *strobe* and transient noise superimposed to the nominal baseline when no charge signal is collected at the CSA input node. The noise induced by the digital switching activity in the latch is comparable with the noise at the Front-End output.

Pre-layout transient simulations for the full analogue Front-End chain coupled to the synchronous comparator are presented instead in Figure 2.83 and 2.84. As one can see, the digital switching activity in the latch introduce spikes in correspondence of each *strobe* transition.
Indeed, the magnitude of such a digital noise is comparable with the simulated transient noise superimposed to the nominal baseline. Note that the largest spikes are only in correspondence of a rising edge of the clock signal, when positive feedback is enabled in the latch and kickback noise due to large voltage swings at latch internal nodes in injected towards the preamplifier through parasitic capacitances. This can be further appreciated in transient noise simulations for the preamplifier output voltages reported in Figure 2.85. Certainly, since only voltage differences determine the actual latch comparison, the common-mode noise that affect the waveforms is cancelled thanks to fully-differential operations, as usually required for any reliable mixed-signal design.

Optimum transistor sizing for latch devices has been derived from W/L parametric simulations with the block coupled to its necessary digital control logic, optimizing capacitive loads and drive strength. In particular, inappropriate transistor sizing can introduce asymmetries in the shape of the high-frequency self-generated clock used to perform fast time-over-threshold encoding by turning the latch into a local oscillator. Practical transistor sizing considerations are therefore discussed later in the chapter.
A final remarkable aspect related to the usage of a latch stage resides instead in the *dynamic offset* that affects these circuit topologies. The usage of positive feedback increases both speed and resolution. However, due to mismatches and process variations the effective minimum initial voltage difference that can be unambiguously discriminated increases. Hence the actual resolution of the Front-End discriminator is degraded. An important task in a synchronous Front-End design is therefore to quantify the circuit dynamic offset V_{osL} and its contribution in determining the effective discriminator threshold dispersion.

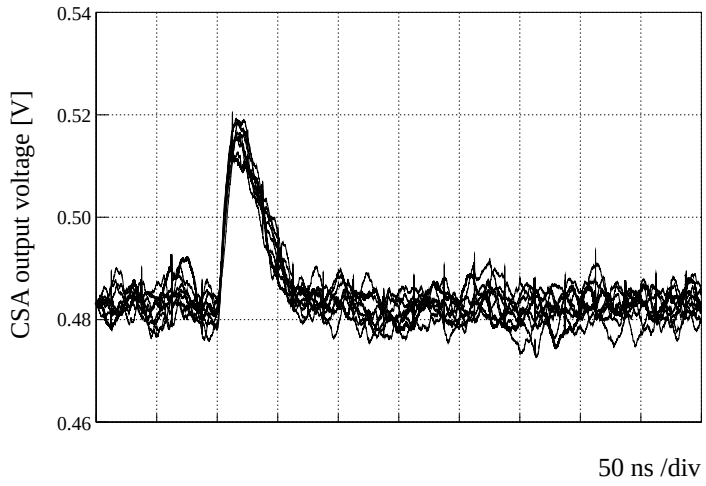


Figure 2.84: Transient noise simulations for the Front-End amplifier coupled to the synchronous comparator and assuming a minimum input charge of 1 ke^-. The injected digital noise does not degrade the quality of the analogue pulse.

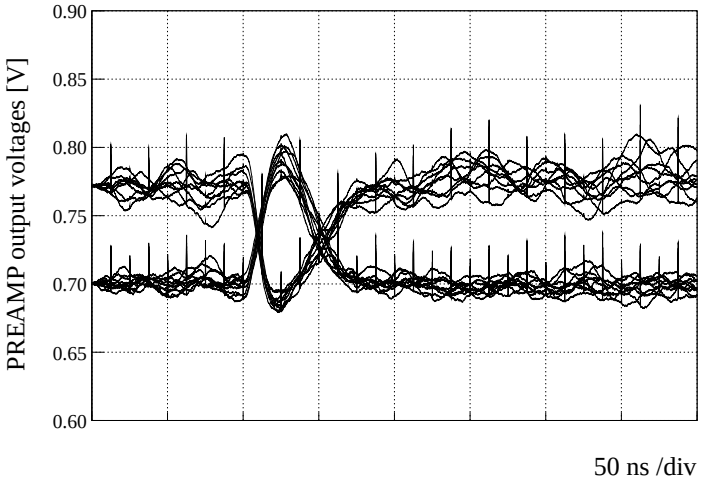


Figure 2.85: Transient noise simulations for the output voltages of the low-gain preamplifier coupled to the latch. Only voltage differences determine the latch comparison, hence the common-mode digital noise is cancelled due to fully-differential operations.

For continuous-time comparators, a DC (static) offset can be easily determined from a sample of bias point analyses across a sufficiently large number of MC runs. Indeed, in a discrete-time architecture mismatches and process variation effects can be appreciated only from the transient (dynamic) response of the system. Simple DC simulations are in fact no longer sufficiently since the operating point depends on the circuit transient behaviour. As an example, Figure 2.86 shows a small sample of 5 transient MC simulations for the time evolution of positive feedback internal nodes. The latch was disconnected from the low-gain preamplifier and a fixed initial voltage difference of +5 mV has been applied at latch input terminals, assuming a quiescent point at half the supply voltage. As one can see, once *strobe* enables the positive feedback, the latch decision varies across MC iterations due to random variations introduced for MOS device parameters.
The more general offset-simulation methodology for comparators based on S-curves is therefore required [Graupner 2006]. The latch dynamic offset can be in fact simulated by varying the initial voltage difference at small steps around the quiescent point. A sample of N transient MC runs is performed for each initial condition and the number of iterations n_i in which the latch output *VoutP* is low[9] (or equivalently, *VoutN* high) is registered. Assuming Gaussian-distributed device variations, the relative frequency

$$z_i = \frac{n_i}{N}$$

as a function of the latch input voltage difference follows the well known Normal cumulative distribution function

$$\begin{aligned} \Phi(z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} dt\ e^{-t^2/2} \\ &= \frac{1}{2}\left[1 + erf\left(\frac{z}{\sqrt{2}}\right)\right] \end{aligned}$$

Simulated data can be therefore fitted with an error function. The actual fit prototype is obtained by replacing z with $(z-\mu)/\sigma$ and the latch dynamic offset V_{osL} is finally retrieved as the standard deviation σ provided by fit parameters. Not surprisingly, this is the same technique extensively adopted in order to determine the effective value of the threshold voltage and its channel-to-channel dispersion in a binary-only Front-End system [Spieler 2005].
Simulation results obtained with 100 transient MC simulations for the final optimized latch are presented in Figure 2.87. As expected, when a 0 mV input voltage difference is presented to latch inputs (i.e. both transistors M1-M2 are simply connected to half the supply voltage) digital outputs *VoutP* and *VoutN* randomly settle to high or low with equal probability. Certainly, by increasing the magnitude of the voltage difference to be discriminated the probability that a wrong comparison is performed decreases. In practice, for input differences larger than 60-70 mV, latch outputs always settle to right values. The standard deviation σ retrieved from fit results indicates that the latch is affected by a dynamic offset of about 22 mV. A description of the implemented offset compensation technique and how the contribution of such a dynamic offset affects the overall Front-End discriminator resolution is addressed in the following section.

[9] As already discussed, in the chosen latch architecture digital outputs are active-low signals.

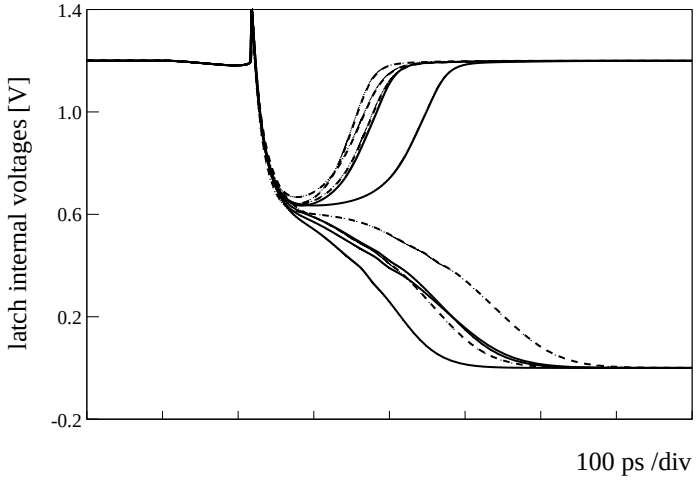


Figure 2.86: Set of 5 transient MC simulations for positive feedback voltages V_X and V_Y assuming a fixed +5 mV voltage difference at latch input nodes. Due to random variations introduced for MOS device parameters, the result of the comparison varies across different MC iterations.

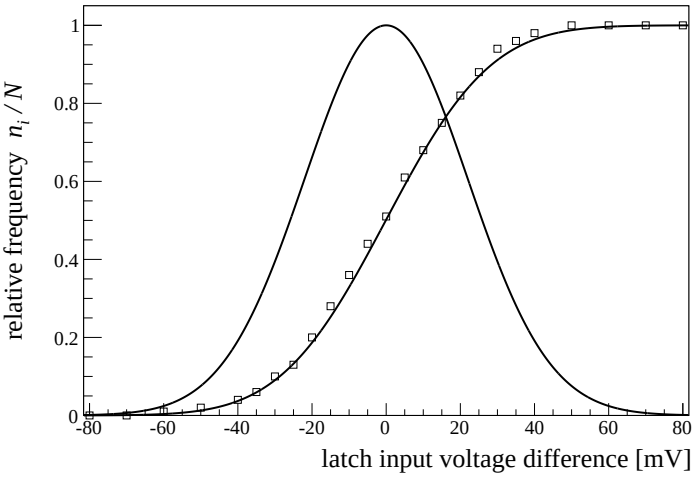


Figure 2.87: Characterization of the latch dynamic offset by means of transient MC simulations for different input voltage differences. From fit results, the standard deviation of the S-curve indicates that the latch is affected by a dynamic offset of about 22 mV. Only for a better visualization, the corresponding Gaussian distribution has been normalized to 1 instead of $1/\sqrt{2\pi}\sigma$.

Local threshold adjustment

Ultimate performance of the analogue Front-End are limited by uncertainties and imperfections that are introduced at each step of the manufacturing flow. Because of mismatches and process variations that affect CMOS fabrication technologies in fact, the effective value of the minimum detectable charge differers from the nominal one. Pixels with a lower effective threshold are more susceptible to fake hit events due to noise, whereas pixels with higher effective threshold values are less efficient. Thus hit detection performance degrade and channel-to-channel random fluctuations must be taken into account. In order to guarantee high detection efficiencies and minimize the noise-induced hit rate a careful design of all circuit components determining the threshold value is therefore required and additional dedicated on-pixel circuitry must be introduced to perform local threshold corrections [Rossi 2005]. Since transistor matching and process variations get worse with CMOS scaling, MOS devices in 65 nm exhibit larger mismatches with respect to higher technology nodes. This makes more challenging to meet low-threshold requirements in the Front-End system. Accurate characterization and modelling of mismatch effects are of primary importance and the extensive MC simulations performed to quantify the foreseen pixel-to-pixel threshold variations rely on statistical models provided by the foundry.
According to performance requirements already discussed for the pixel upgrade, the analogue Front-End must operate with a nominal threshold of about 1000 e^- without compromising hit efficiency and noise occupancy. In the final chip, a global threshold will be provided to all pixels by a D/A converter placed in the chip periphery. A local fine adjustment of the threshold is then required in each pixel cell. Before correction the overall threshold dispersion must be < 400 e^- RMS and both contributions from discriminator mismatches and Front-End amplifier gain variations must be considered. A residual total dispersion < 40 e^- RMS is demanded after threshold tuning.

In the pixel ASIC community, digital trimming is usually employed to minimize channel-to-channel threshold variations [Snoeys 2001, Llopart 2002, Erdmann 2005, Peric 2006, Karagounis 2011, Kugathasan 2011, Valerio 2012]. Depending on the specific resolution required by the target application, 3-7 bit binary words are stored in every pixel and fed to a dedicated D/A converter that generates a calibration voltage or current for the Front-End discriminator. In principle, this solution can be adopted without any restrictions also for the proposed latched comparator. As an example, current sourcing/sinking provided by a on-pixel current-steering DAC can be introduced at preamplifier output nodes. Nevertheless, an alternative method has been implemented. In fact, one of the most relevant characteristic of discrete-time comparators is the possibility of achieve high-precision offset corrections by means of autozeroing techniques [Razavi 1995, Johns 1996, Gregorian 1999, Allen 2002, Maloberti 2003, Figueiredo 2009].
In CMOS autozeroed architectures, sampling switches are introduced and voltage differences due to mismatches and process variations are periodically sensed, stored on capacitors and finally added to comparator sensitive nodes such that offset cancels itself. A dedicated period is needed to carry out the calibration, thus voltage comparison and offset correction become complementary phases. Appropriate digital waveforms must be generated to control MOS switches activity, complicating the timing scheme of synchronous operations. Certainly a physical limit is set by the discharge of compensation capacitors due to device leakage currents, requiring a periodic calibration cycle after a certain amount of time. Since leakage currents increase with technology scaling, detailed characterizations of their contributions are mandatory for any autozeroed systems. However, after validation of an autozeroed approach, efficient calibration schedules can be defined according to on-line machine operations.

Autozeroing techniques are extensively used in IC industry for the design of high-resolution A/D converters for commercial applications and electronic instrumentation. Some Front-End solutions for the readout of segmented silicon sensors already include discrete-time autozeroed comparators [Rivetti 2001, Degerl 2003]. The fundamental advantage in adopting an autozeroing procedure for the local threshold tuning is that the usage of a on-pixel D/A converter is no more required. The circuit complexity of the discriminator increases, but the absence of a local D/A converter simplifies the design of the overall analogue Front-End. Depending on the actual circuit topology and layout this can lead to a more compact design. Moreover, in perspective of pixel operations in a hostile radiation environment the lack of a on-pixel converter also simplifies the design of the digital part, avoiding the necessity of dedicated Single Event Upset (SEU) tolerant registers to store the configuration bits for digital trimming. As a result, the available area for local temporary data storage and signal processing can significantly increase.

A reach variety of CMOS circuit topologies is proposed in literature for practical implementation of autozeroed comparators and operational amplifiers. A few different autozeroing schemes have been explored in the preliminary design phase. As a final solution, the Output Offset Storage (OOS) technique was adopted, according to key advantages of this architecture with respect to the Input Offset Storage (IOS) counterpart [Razavi 1992]. More complex autozeroed comparators proposed in literature for high-speed and high-precision applications can employ both IOS and OOS techniques [Brianti 1997]. Nevertheless, the natural choice for a synchronous comparator that uses a single stage preamplifier coupled to a latch is OOS. The circuit is illustrated in the schematic of Figure 2.88. Two autozeroing capacitors C_{az} are inserted in series between the low-gain preamplifier and the latch stage. CMOS sampling switches controlled by ϕ_{1A}, ϕ_{1B} and ϕ_2 digital signals are then used to trigger discriminator operations with opportune timing. A sample timing diagram is presented in Figure 2.89. The circuit operates as follows.

During offset compensation ϕ_{1A} and ϕ_{1B} switches close, while ϕ_2 switches open. Both latch and preamplifier input nodes are therefore shorted to a common-mode reference voltage V_{BL}. In practice, the value has been chosen equal to the nominal baseline voltage of the Front-End amplifier. Such an external reference voltage ensures proper bias conditions for preamplifier input transistors, hence the total input-referred offset V_{os} seen at comparator input nodes is amplified by the voltage gain A_v of the preamplifier. The resulting output-referred offset voltage $A_v V_{os}$ is stored on series capacitors and the circuit consisting of preamplifier and autozeroing capacitors ideally exhibits zero DC offset voltage. Positive feedback in the latch can be disabled if required, but strobe activity does not interfere with the offset cancellation procedure since in the compensation phase latch inputs are tied together. During normal operations instead, preamplifier and latch stages are disconnected from the reference voltage by opening ϕ_{1A} and ϕ_{1B} switches, whereas ϕ_2 switches close such that the comparator can track voltage differences presented to its inputs.
The choice and practical implementation of autozeroing using the OOS technique introduce a few important considerations. Without doubts a main advantage of such a solution is offered by the minimum increase of circuit complexity with respect to the basic preamplifier-latch architecture. Only CMOS sampling switches and series capacitors are added in the system, without the need of additional stages. Since autozeroing capacitors are placed in series there is no separation between the signal path and the offset compensation path. The preamplifier always works in open-loop configuration, hence stability is not a problem. However, when offset cancellation is achieved the preamplifier may saturate if the product $A_v V_{os}$ of the gain and the input-referred offset exceeds the maximum voltage swing allowed at output nodes. Thereby OOS necessary requires the usage of a preamplifier stage with a limited gain, usually less then 10.

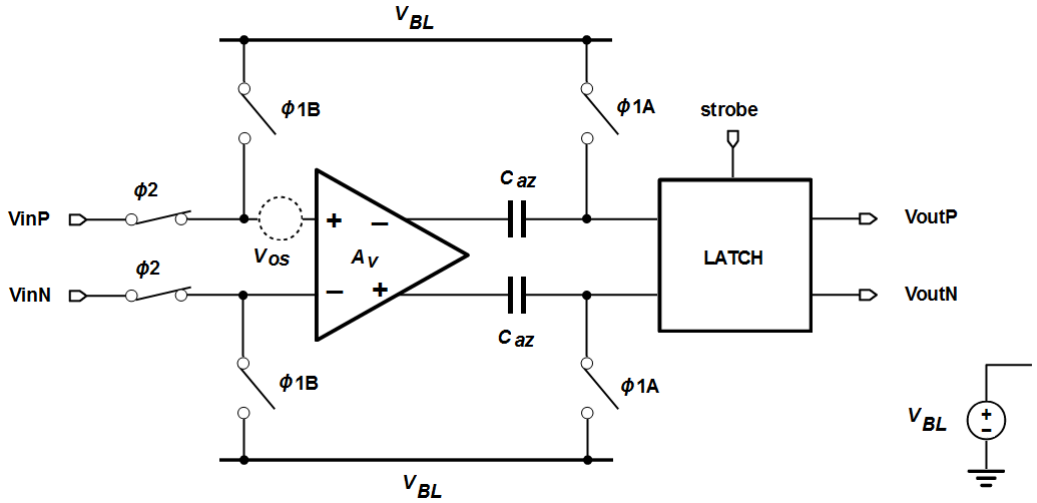


Figure 2.88: Autozeroing performed by means of Output Offset Storage (OOS). The input-referred offset is represented as a voltage source V_{os} of random magnitude and sign connected at the input of the low-gain preamplifier. All switches are CMOS and proper timing is required for ϕ-signals to reduce charge injection contributions to the residual offset of the system.

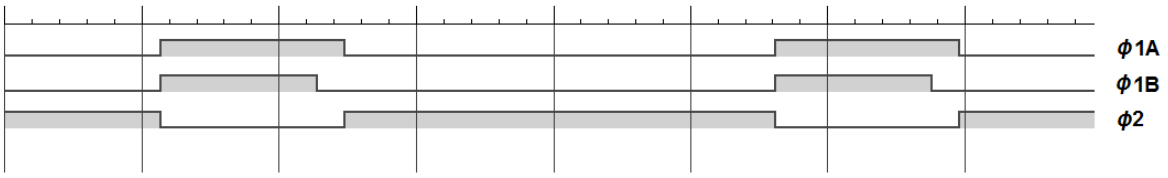


Figure 2.89: Proper timing for ϕ-signals (arbitrary timescale).

Indeed, the most notable advantage of OOS resides in the fact that the output-referred offset of the preamplifier $A_v V_{os}$ is completely cancelled, a key feature of the OOS technique in contrast with IOS. The residual offset seen at latch input nodes after compensation is determined by two contributions. The first term is the already discussed dynamic offset of the latch V_{osL}, which is not cancelled. A second contribution to the residual offset is introduced instead by channel charge injection mismatches that affect CMOS switches. The total residual input-referred offset V'_{os} after compensation is therefore obtained by dividing these two contributions for the open-loop gain A_v of the preamplifier stage,

$$V'_{os} = \frac{\Delta q}{A_v C_{az}} + \frac{V_{osL}}{A_v}$$

where Δq represents the difference in the amount of charge injected at latch input nodes when switches ϕ_{1A} and ϕ_{1B} opens. As one can see, the effect of charge injection mismatches $\Delta q/C_{az}$ is scaled on the preamplifier gain, which is a further advantage of OOS with respect to IOS.
Actually, depending on the exact turn-on and turn-off timing of ϕ_{1A} and ϕ_{1B} signals, significant contributes to the residual offset due to charge injection mismatches can be introduced at latch input nodes. In fact, if after compensation ϕ_{1B} switches turn-off later than ϕ_{1A} counterparts, an additional uncancelled offset due to charge injection mismatches is caused at the inputs of the latch stage when ϕ_{1A} switches open. Indeed, ϕ_{1B} devices must turn-off before ϕ_{1A} counterparts, such that charge injection contributions when ϕ_{1A} open can no more affect latch input nodes. This justifies the timing diagram depicted in Figure 2.89. Similar considerations apply also in multi-stage autozeroed comparators that implement OOS at each stage, where sequential clocking is adopted to minimize charge injection mismatch effects. Proper digital waveforms for ϕ_{1A} and ϕ_{1B} signals can be generated by means of asynchronous logic using delays and coincidences. The ϕ_2 control signal can be simply obtained by inverting the longer ϕ_{1A}. Such a solution has been implemented in the final design. A clock generator implemented as a NOR-based S/R latch with asynchronous delays inserted on the feedback path can be adopted as well, as reported in [Johns 1996]. If a synchronous logic design approach is adopted instead, level-to-pulse converters, counters or shift registers can be used, hence all pulse durations for ϕ-signals are derived from the master clock used to synchronize digital operations. Additionally, more sophisticated techniques must be adopted if precise non-overlapping control signals are required, as usually demanded for high-speed applications.
Storage capacitors C_{az} have been implemented using MOM capacitors provided by the 65 nm library. The final choice for the capacitance value has been 80 fF, representing a good compromise between layout area and offset compensation performance over long periods of time.

Autozeroing performance have been investigated by means of iterated transient MC analyses over 1 ms simulation time. Figure 2.90 shows the distribution of the voltage difference at preamplifier output nodes sampled for each MC iteration at 100 μs after a calibration cycle. Not surprisingly, before storage capacitors the RMS value of the distribution is about 9 mV, in agreement with the expected output-referred offset of the preamplifier stage already determined from simple DC analyses across different MC runs. The distribution of sampled voltage differences at latch input nodes is presented instead in Figure 2.91. As one can see, the RMS value of the distribution indicates that after compensation the residual offset at latch inputs is reduced to about 760 μV RMS. As mentioned above, such a contribution is due to the charge injection mismatches introduced when ϕ_{1A} switches turn-off. The latch dynamic offset V_{osL} has been previously determined as 22 mV RMS. With an open-loop gain $A_v \approx 6$, after compensation the total input-referred residual offset is reduced to about 3 mV RMS, a value comparable with the simulated transient noise superimposed to the nominal baseline of the Front-End amplifier. Assuming for the sake of simplicity a single-electron to voltage gain q_e/C_F and 4 fF feedback capacitance, it turns out that after autozeroing the residual offset if reduced to rough 75 e^- RMS.

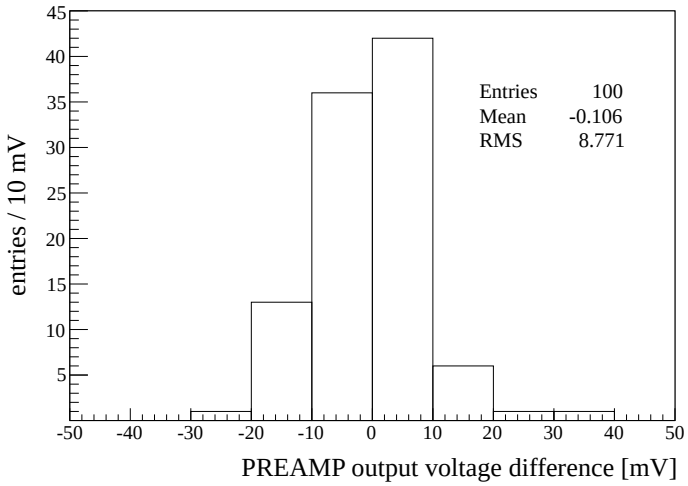


Figure 2.90: Distribution of voltage differences at preamplifier output nodes from 100 transient MC simulations. For each MC iteration, the voltage difference has been registered 100 μs after the calibration cycle. As expected, the RMS of the distribution is in agreement with the DC output-referred offset of the low-gain preamplifier.

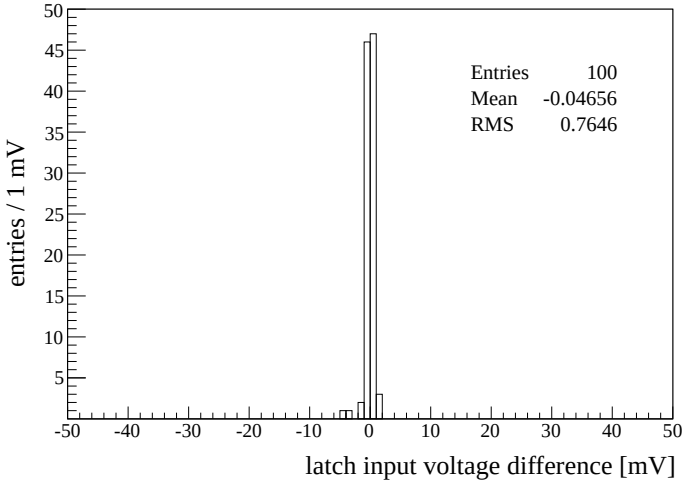


Figure 2.91: Distribution of voltage differences at latch input terminals for the same testbench. The residual offset at latch inputs is reduced to about 760 μV RMS.

A small sample of transient MC simulations over a 1 ms period is reported in Figure 2.92, whereas Figure 2.93 shows the RMS value of the residual offset at latch input terminals as a function of time. The offset is efficiently stored on autozeroed capacitors up to 100 μs. Transient MC simulations with a charge-induced signal after offset compensation are presented in Figures 2.94 instead.

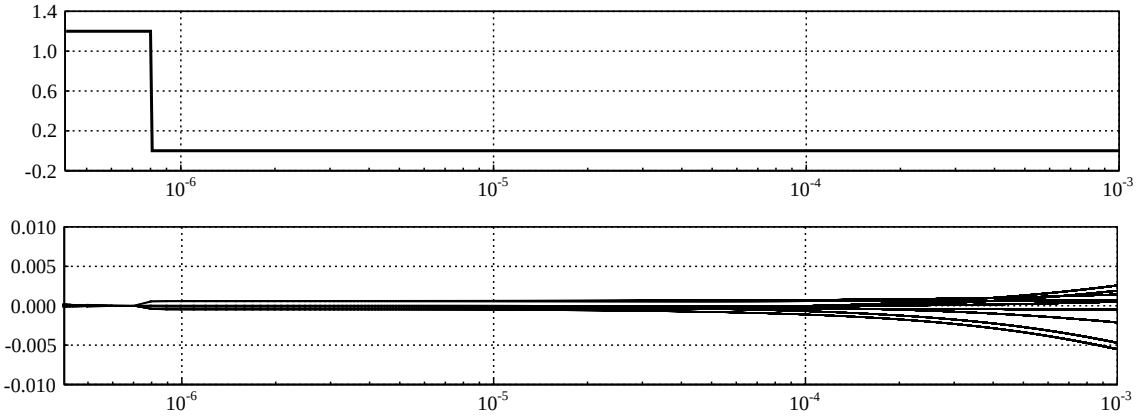


Figure 2.92: Sample of transient MC carlo simulations for the voltage difference at latch input terminals. The calibration cycle completes when ϕ_{1A} (top signal) goes low. Note the logarithmic timescale up to 1 ms.

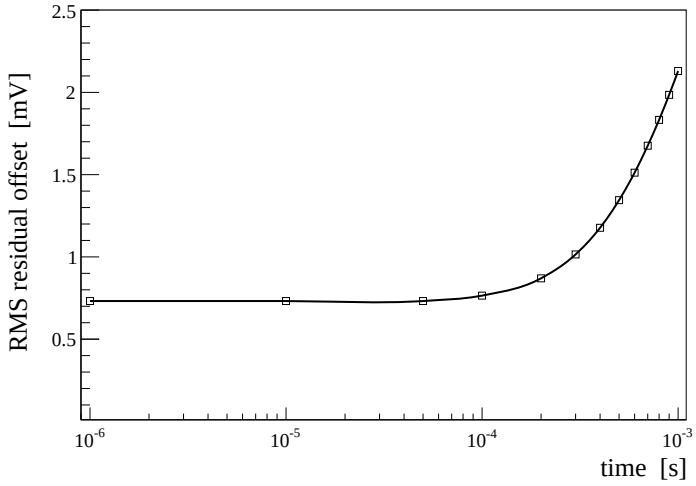


Figure 2.93: RMS residual offset at latch input nodes as a function of time.

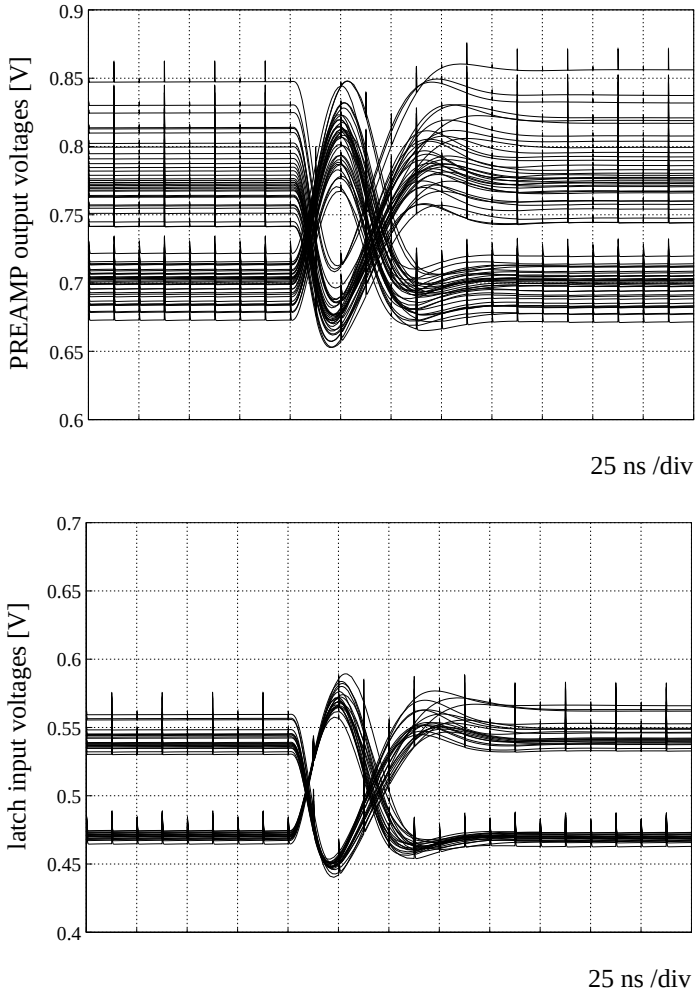


Figure 2.94: Complete transient MC simulations including the latch *strobe* activity. A minimum charge signal of 1 ke^- has been injected at the CSA input node after an offset calibration cycle. From top to bottom: differential voltages at the low-gain preamplifier outputs (before storage capacitors) and at latch input terminals (after storage capacitors).

AC coupling with the Front-End amplifier

The analogue waveform generated by the Front-End amplifier is fed to the synchronous comparator for the hit discrimination. The coupling between the two stages can be either DC or AC. If a DC solution is adopted, the CSA directly drives the non-inverting input of the comparator, without the need of additional circuitry. In case an AC coupling is performed instead, a capacitor C_{AC} is inserted in between the two stages, thus suppressing the DC component of the signal. On the one hand, this eliminates pixel-to-pixel baseline fluctuations due to mismatches and process variations. On the other hand, proper DC bias point must be guaranteed at the input of the comparator. Both AC and DC coupling solutions have been explored for the proposed synchronous Front-End architecture. In the final implementation, AC coupling was adopted for the first prototype. As sketched in schematic of Figure 2.95, the DC operating point at the input node of the low-gain preamplifier is established by means of a bias resistor R_b connected to the common-mode voltage V_{BL} employed for the offset compensation. A sample of transient MC simulations for the analogue waveform at the Front-End amplifier output node is presented in Figure 2.96. Certainly, a Krummenacher feedback network is more sensitive to mismatches with respect to other constant-current feedback solutions. As discussed, trade-offs exists between transistor sizing and loop stability, which in turn lead to a trade-off between mismatches and loop stability. This introduces quite large worst-case variations in the DC value of the baseline voltage at the Front-End amplifier output node, about 10% with respect to its nominal voltage, thus increasing the effective offset contribution. This can be appreciated in the MC distribution of Figure 2.97. As shown in Figure 2.98, baseline variations are cancelled instead after the decoupling capacitor. The final choice for an AC-coupled solution was essentially motivated by the request of avoid excessive baseline fluctuations at the discriminator input node.

The actual transistor level implementation is presented in schematic of Figure 2.99. A 100 fF Metal-Oxide-Metal (MOM) capacitor from the 65 nm library has been adopted. The bias resistor is implemented instead with a couple of NMOS transistors, resulting into an on-resistance of about 200 MΩ. The extra filtering and thermal noise introduced by the RC network has been carefully investigated, optimizing the transistor sizing for bias devices and CMOS switches.

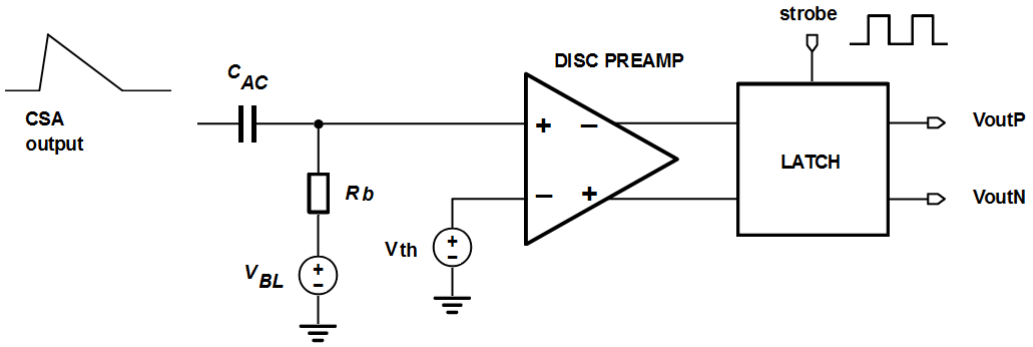


Figure 2.95: In the final architecture, AC coupling is performed between the Front-End amplifier and the synchronous discriminator. Autozeroing capacitors and sampling switches are not shown in figure.

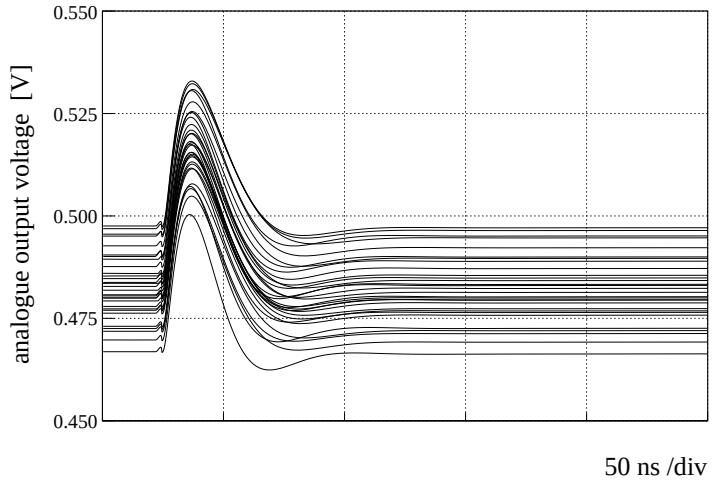


Figure 2.96: Transient MC simulations for the analogue waveform at the Front-End amplifier output node, assuming 1 ke^- input charge, 100 fF input capacitance, 4 fF feedback capacitance and 40 nA total feedback current. Baseline variations affect the signal due to mismatches and process variations.

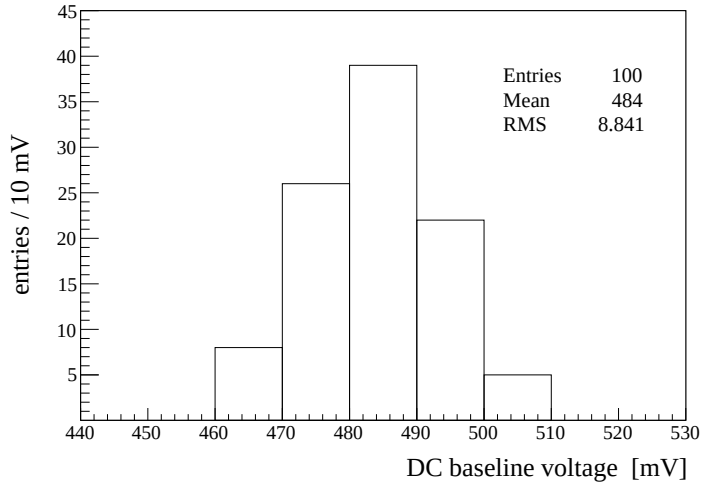


Figure 2.97: Simulated MC distribution for DC baseline values at the Front-End amplifier output node. Due to the usage of a Krummenacher feedback architecture, baseline variations are quite large, about 50 mV (10%).

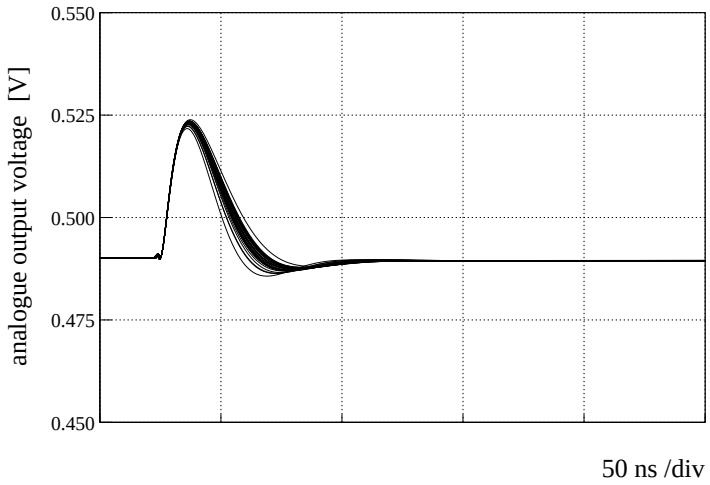


Figure 2.98: Transient MC simulations for the analogue waveform fed to the low-gain preamplifier. Thanks to AC-coupling, baseline fluctuations are cancelled.

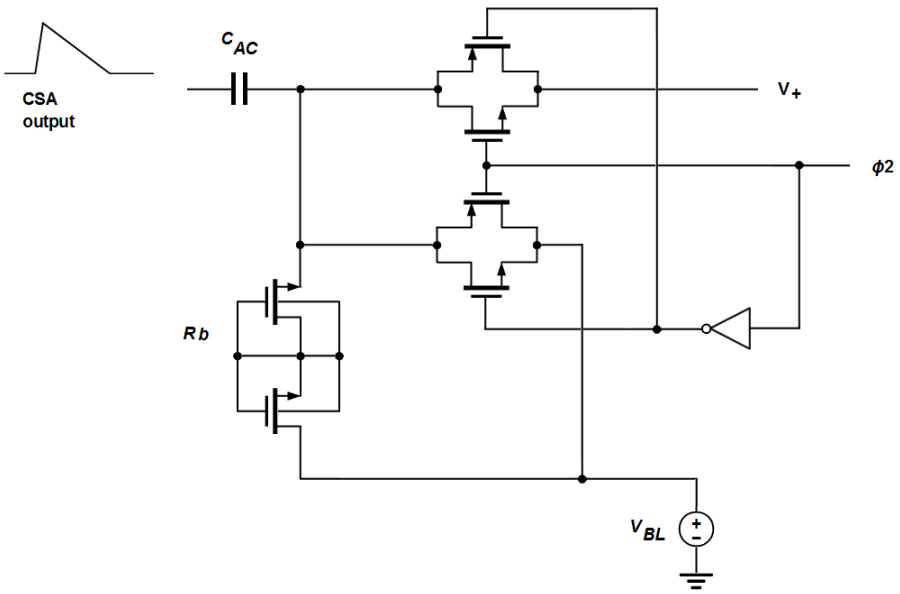


Figure 2.99: Practical AC coupling transistor level implementation. A 100 fF decoupling capacitor is implemented with a MOM capacitor from the 65 nm process library. A 200 MΩ bias resistance is provided by a couple of wide NMOS devices. Note that CMOS switches controlled by ϕ_2 work in complementary phases. This is required in order to establish a well-defined DC level when the discriminator is disconnected from the decoupling capacitor during autozeroing.

Latch operations in asynchronous logic feedback loop for fast time-over-threshold (TOT) encoding

The proposed synchronous Front-End solution introduced some remarkable features. On the one hand, hit generation is synchronized with an external 40 MHz master clock, preventing time-walk issues in the time-stamp assignment. On the other hand, the usage of a low-gain preamplifier coupled to a positive feedback stage provides fast response and high-resolution performance for voltage comparisons, with the capability of discriminate very low charge-induced signals above the nominal threshold. Moreover, discrete-time comparators can implement autozeroing techniques for the local threshold adjustment, hence high-precision calibration schemes can be defined according to on-line machine operations, avoiding the usage of local D/A converters. As a matter of fact, synchronous Front-End architectures naturally fits into a bunched-experiment environment.
A further interesting and very promising feature offered by a synchronous approach resides in the possibility of performing a fast time-over-threshold (TOT) encoding by turning the track-and-latch comparator into a local oscillator. This is discussed in the following.

Over the years, different solutions have been proposed within the pixel ASIC community to perform amplitude measurements using the TOT technique. In a first generation of pixel systems TOT information is retrieved on a time-stamp based approach [Peric 2006, Mazza 2012]. A free-running counter placed in the chip periphery is used to generate a time-stamp word that is distributed to all pixels. Gray encoding is adopted to minimize the switching activity and data transmission errors. When a particle hit is detected, a dedicated digital control logic registers leading-edge and trailing-edge occurrences of the hit pulse provided by the Front-End discriminator and stores the corresponding time-stamp values in banks of D-type latches or RAM cells. The TOT information is finally obtained as difference between two time-stamp counts. The main drawback of this approach resides in the large power consumption introduced by the propagation of the time-stamp to all channels. This can pose severe limitations for the maximum clock frequency that can be used to generate and distribute the time-stamp. Hence time resolution can be an issue.
Power consumption can be reduced introducing local counters, avoiding time-stamp distribution. If speed and resolution are not tight constraints, a clock signal can be propagated to all pixels. The clock can be provided off-chip or generated in the chip periphery with clock-multiplication techniques from a lower frequency external clock. TOT digitization is then obtained by counting the number of clock cycles within the hit pulse window [Hemperek 2009].
To overcome speed and power limitations, efficient solutions necessary require the usage of locally generated high-frequency clock signals by means of CMOS oscillators. In these systems, clocks in the 500-700 MHz range are generated at the pixel level by means of Voltage-Controlled Oscillators (VCOs) placed in each pixel or shared among pixels [Llopart 2007, Gromov 2010, Poikela 2014, Valerio 2014]. In order to mitigate frequency differences due to PVT variations, a Phase-Locked Loop (PLL) circuit located at the chip periphery provides a control voltage which regulates the actual frequency of individual oscillators [Fu 2014].
More complex systems for precise time and energy measurements have been designed including on-pixel Time-to-Digital Converters (TDCs) as well [Martoiu 2009, Rolo 2013].

Discrete-time comparators can be easily turned into oscillators by means of asynchronous logic. This technique is widely employed in the design of modern high-speed, power-efficient charge redistribution Successive Approximation Register (SAR) A/D converters. In such systems, track-and-latch comparators or single stage latches are coupled to asynchronous logic in feedback loop in order to internally generate the necessary clock signals for SAR operations [Liu 2010]. This has suggested that high-frequency self-generated clocks for fast TOT digitizations would be available in 65 nm CMOS process as part of the proposed synchronous Front-End architecture. A simple testbench has been adopted to validate the feasibility of such a solution.

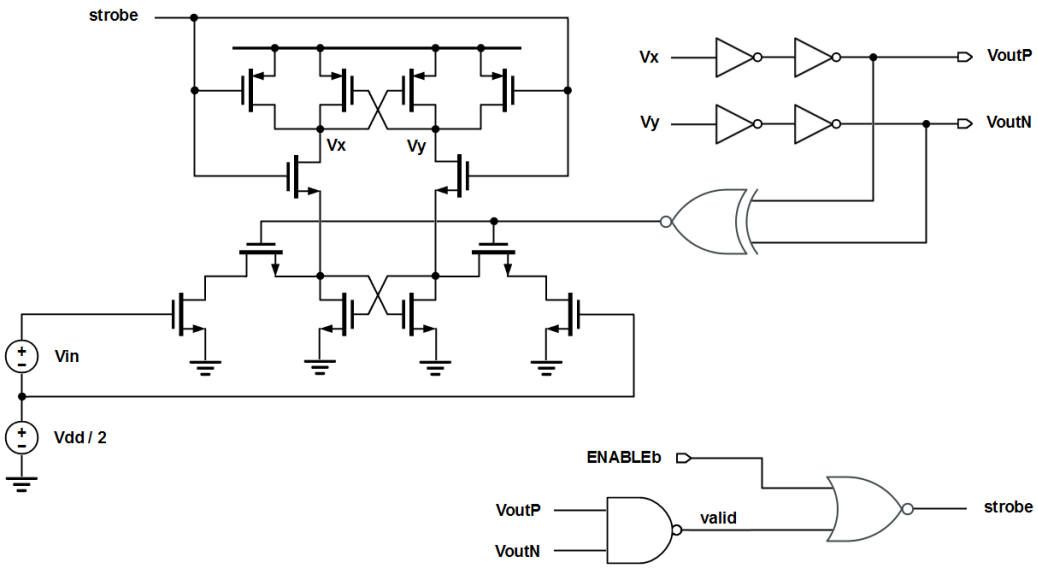


Figure 2.100: Testbench schematic to validate the feasibility of turning the synchronous comparator into a local oscillator using an asynchronous logic feedback loop around the latch.

Referring to the schematic shown in Figure 2.100, for simulation purposes an active-low external signal *ENABLEb* is introduced to control latch operations. An asynchronous logic feedback loop is established on the *strobe* signal path, generating a *valid* signal from *VoutP* and *VoutN* outputs and feeding it back to the latch. The XNOR gate ensure low-power operations breaking current paths after positive feedback transients have completed, as discussed. As shown in Figure 2.101, at the beginning of a transient simulation *ENABLEb* is tied high, forcing *strobe* at the output of the NOR gate to the ground rail. Positive-feedback in the latch is disabled and internal nodes V_X and V_Y are forced to the supply voltage. Both *VoutP* and *VoutN* outputs are tied high as well, thus *valid* at the output of the NAND gate is low. When *ENABLEb* is pulled down, *strobe* switches from low to high and positive-feedback is enabled. After the regenerative transient, *VoutP* and *VoutN* settle to complementary levels according to the sign of the voltage difference *Vin* presented at latch input nodes. Due to *VoutP* and *VoutN* complementarity, *valid* toggles from low to high and thanks to the inverting behaviour of the NOR gate, *strobe* is again forced to ground. Hence latch internal nodes reset to high, restoring positive-feedback operations. Provided that *ENABLEb* remains low, a new comparison is performed and latch outputs settle again to opposite rails. As a result, the latch has been turned into a compact clock generator.
The resulting oscillation frequency only depends on latch transistor sizing and propagation delays introduced in the asynchronous feedback loop by NAND and NOR gates. As shown in Figure 2.102, for a 100 mV input voltage difference the final optimized circuit provides an intrinsic (no extra delay is inserted in the asynchronous feedback path) oscillation frequency larger than 4 GHz.

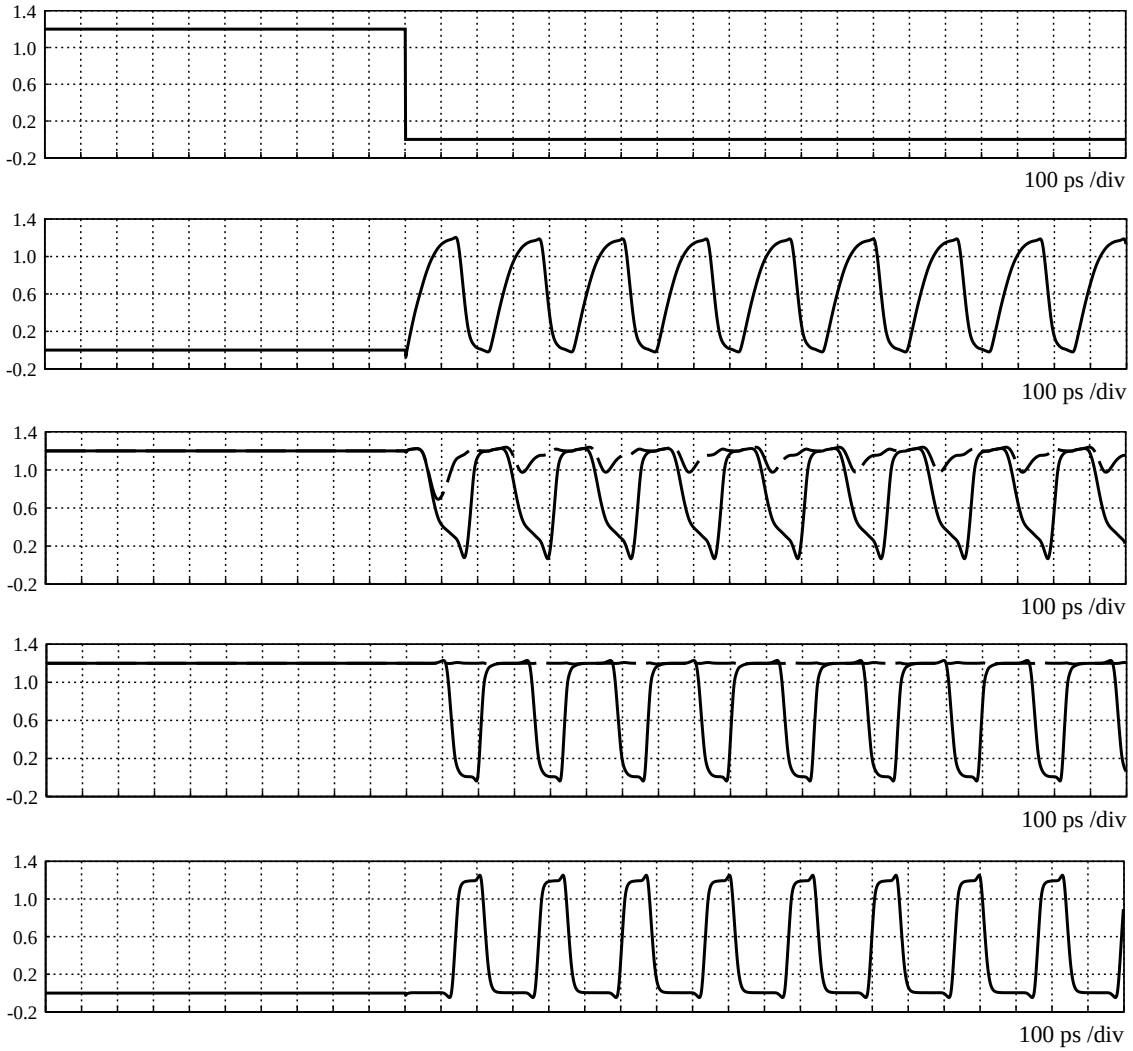


Figure 2.101: Transient simulation for latch operations in asynchronous logic feedback loop. From top to bottom: *ENABLEb*, *strobe*, latch internal nodes V_X/V_Y, digital outputs *VoutP*/*VoutN* and *valid*. Due to logic gates propagation delays the latch turns into a high-frequency clock generator.

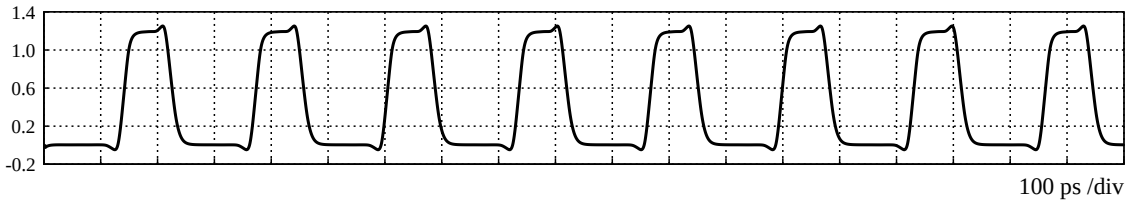


Figure 2.102: Simulated self-generated clock waveform *valid* for the final optimized latch. Assuming a 100 mV input voltage difference the intrinsic oscillation frequency is larger than 4 GHz ($\approx$ 225 ps clock period).

In practice, the final optimized transistor sizing was derived from W/L parametric simulations with the the latch coupled to the above described asynchronous logic. The primary goal in the latch design has been in fact to ensure that a symmetrical and reliable *valid* clock waveform is generated by the latch/oscillator in perspective of a fast TOT digitization with such a solution. In particular, due to parasitic capacitances most critical devices are NMOS input transistors M1-M2 and PMOS switches M9-M10. Both PMOS and NMOS switches can introduce in fact significant capacitance at positive feedback nodes if oversized. This leads to slower charging and discharging processes during the regenerative transient, which in turn generates asymmetries in the shape of the *strobe*, hence in the duty cycle of the *valid* clock. Furthermore, the oscillation is affected by frequency variations due to mismatches and PVT variations. Thereby special attention has been devoted in characterizing the frequency stability across MC and PVT corner simulations.
A small sample of transient MC simulations for the *valid* signal is presented in Figure 2.103. As mentioned, by applying a 100 mV voltage difference at latch input terminals the resulting oscillation frequency is larger than 4 GHz. Certainly, the intrinsic frequency varies due to random fluctuations in MOS device parameters. Pixel-to-pixel frequency variations must be therefore quantified. Figure 2.104 shows the distribution of the latch intrinsic frequency as derived from 100 transient MC simulations. Mismatches and process variations introduce a spread of about 340 MHz RMS with respect to a nominal value of 4.1 GHz. The latch turned into an oscillator is therefore fairly stable against pixel-to-pixel frequency variations, at the 8% level.
Frequency variations across different process corners are presented in Figure 2.105. As one can see, large worst-case frequency variations affect the system. Nevertheless, this does not represent a limiting factor since process variations can be always compensated by means of PLL techniques, as already performed in pixel ASICs that employ local oscillators to perform the TOT encoding.
Frequency stability across different operating temperatures is istead of primary importance. On the one hand, during on-line operations the CMS pixel tracker is cooled down. At present, the lowest temperature that the cooling system can deliver to pixels is -30 °C, with a nominal operating temperature of -25 °C. On the other hand, the new pixel ASIC has to be operational in the basic commercial range (0 °C-80 °C) extended to lower temperatures. The circuit response must be therefore as much as possible uniform, from simple room-temperature bench characterizations down to nominal cooled operations. As shown in Figure 2.106, the latch/oscillator exhibits excellent stability across temperature variations from -50 °C to +50 °C, with frequency variations below 200 MHz with respect to 4.1 GHz at 27 °C, less than 5%.
Figure 2.107 presents the simulated intrinsic frequency across ± 200 mV supply voltage variations (more than ± 15%) with respect to a nominal 1.2 V core voltage[10] As expected, the frequency linearly increases with the supply voltage. Simulated voltage variations are actually much larger than the typical maximum allowed supply voltage uncertainty, typically ± 5%. Assuming ± 60 mV DC variations around 1.2 V, frequency variations are of the order of 300 MHz. Certainly the usage of on-chip Low Drop-Out (LDO) voltage regulators to ensure supply voltage stability and a careful power distribution scheme that minimizes voltage drops along power supply lines are mandatory.
Eventually, the dependence of the intrinsic frequency as a function of the latch input voltage difference has been investigated as well. This is shown in Figure 2.108, where voltage differences have been applied around a 600 mV quiescent point, half the supply rail. The frequency if stable for input differences larger than 200 mV. The characteristic is affected instead by a non linear frequency decrease for voltage differences below 200 mV. As already discussed for the basic bistable element, the time required to produce full logic levels after positive feedback is triggered in the circuit depends on the initial condition. Indeed, the usage of a low-gain preamplifier stage coupled to the latch mitigates such a non linearity, always presenting to latch input nodes amplified values.

[10] As discussed later, the latch has been connected to the *digital* supply rail *VDDD* in order prevent the analogue counterpart *VDDA* to be corrupted by the intense digital switching activity during fast TOT encoding. A supply voltage of 1.2 V has been adopted for both the analogue and digital components. Indeed, the choice of a scaled digital supply voltage of 800 mV is under discussion within the RD53 collaboration.

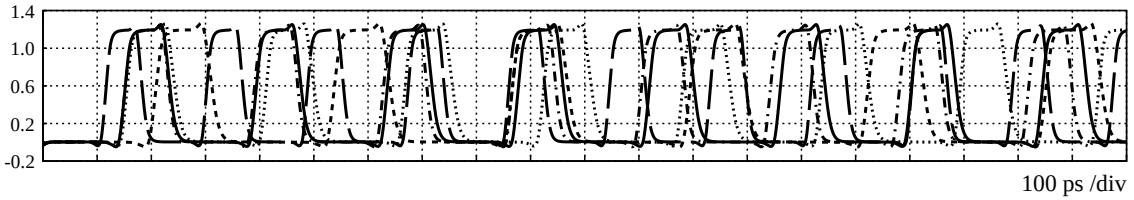


Figure 2.103: Self-generated clock signal *valid* for different transient MC runs. The intrinsic oscillation frequency varies across iterations due to random fluctuations introduced for MOS device parameters.

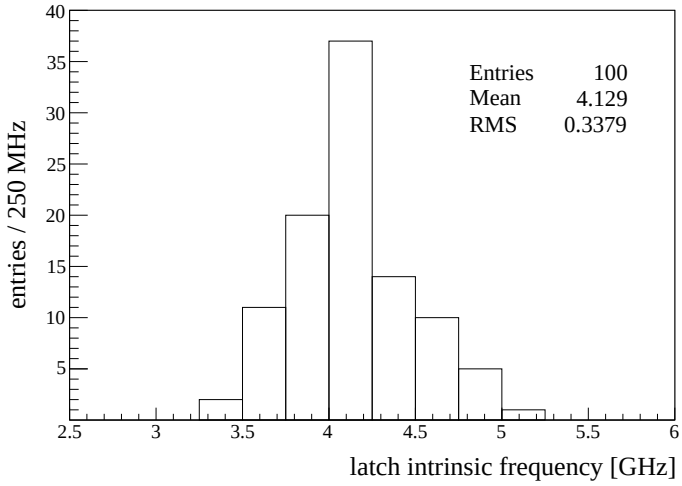


Figure 2.104: Distribution of the intrinsic oscillation frequency across 100 transient MC iterations.

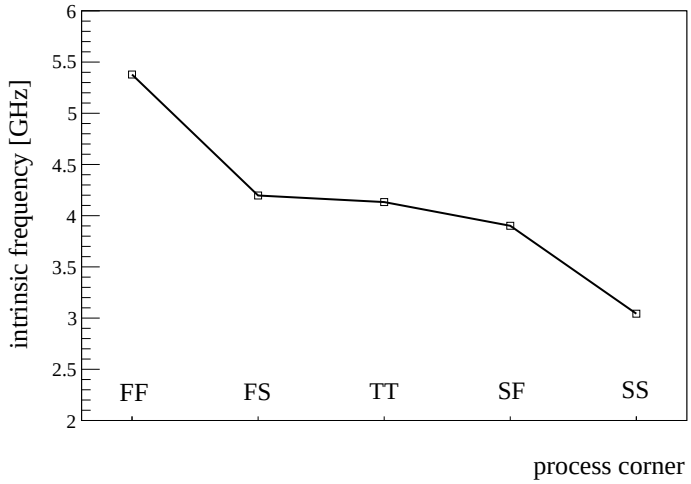


Figure 2.105: Simulated latch intrinsic frequency across different process corners.

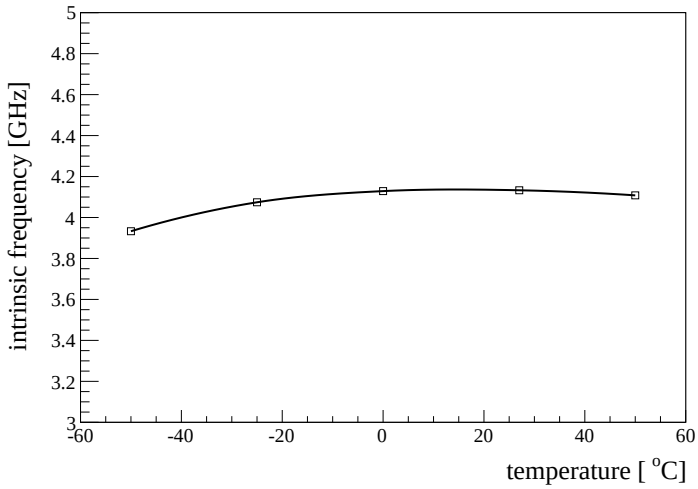


Figure 2.106: Simulated latch intrinsic frequency across ± 50 °C temperature variations.

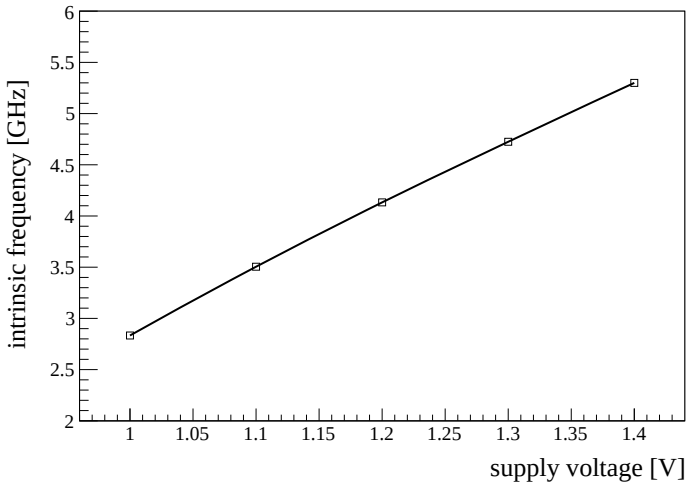


Figure 2.107: Simulated latch intrinsic frequency across ± 200 mV DC supply voltage variations.

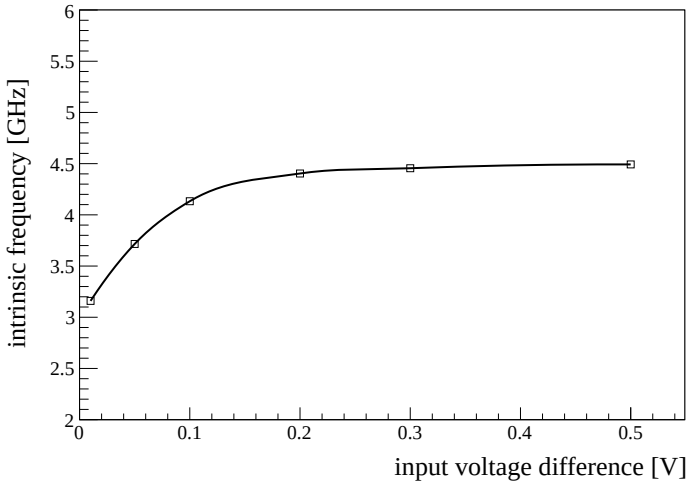


Figure 2.108: Simulated latch intrinsic frequency as a function of the input voltage difference.

Thanks to speed offered by a 65 nm CMOS technology, clock frequencies that can be obtained with such a technique are in the 4 GHz range. Note that beside the clear advantage of a fast charge encoding, the usage of a local high-frequency clock introduces an additional benefit. In fact, the frequency content of the digital noise due to a such high-frequency switching activity is beyond the limited bandwidth of the Front-End amplifier, which in turn would result into a better immunity of the analogue waveform against digital-induced disturbances. Certainly, the intrinsic frequency can be then reduced and tuned by introducing an appropriate delay element in the *valid* asynchronous feedback path, as depicted in Figure 2.109.

Delay lines are fundamental constituents of modern high-speed integrated circuit and different CMOS implementations can be found in literature [Mahapatra 2002].
The choice of the actual delay circuit topology represents a key point in order to guarantee low power operations, stability and adequate linearity for TOT encoding. As an example, delays can be obtained using an inverter chain with MOS shunt-capacitors. However, due to the large power consumption introduced by charging and discharging processes this is not a feasible solution.
On the one hand, the limited available area dictates that the number of stages in the delay line must be minimized. On the other hand the usage of a variable-delay element allows to tune the frequency of the self-generated clock, increasing the flexibility of the system. In practice, current starving is the natural choice for implementing a low-power, compact and flexible design, including the possibility of employ PLL techniques to compensate frequency differences due to PVT variations. Hence a voltage-controlled delay line was inserted in the logic feedback loop. As a result, the overall system becomes a compact VCO and configurable high-speed TOT measurements can be performed up to the GHz level by means of a synchronous Front-End architecture.
A description of the on-pixel digital control logic designed to support latch operations and perform fast TOT digitizations is remanded to Section 7.7. A schematic view of the circuit topology adopted for the implementation of the delay element as part of the final control logic is shown in Figure 2.110. As one can see, current starving is simply performed with a series-connected NMOS device placed in between a CMOS inverter. This introduces an asymmetry between pull-up and pull-down transitions, hence two cascaded starved inverters are required. A further output inverter provides necessary drive strength and fanout for subsequent logic gates placed in the digital part. The control voltage $Vctrl$ is provided by a current mirror driven by an external current $Ictrl$. The usage of a frequency control performed in the current-domain is mandatory in order to minimize frequency variations. The usage of a global DC control voltage would be in fact unreliable since any small variation in the gate-source voltage of NMOS starving devices would lead to large variations in their on-resistance, thus in the oscillation frequency. Note that during any low-to-high or high-to-low transitions, one of the two inverters of the delay element is discharged trough the controlled device, with a delay which is inversely proportional to the discharging current. The other one is charged instead through a simple PMOS devices. Neglecting the contribution of the output inverter, the delay of starved inverters is the sum of a controlled delay (pull-down transition) and a simple CMOS inverter delay (pull-up transition).

Sample transient simulations for the *valid* clock waveform by varying the control current $Ictrl$ are presented in Figure 2.111. As one can see, a flexible range of frequencies is achievable, from 250 MHz up to 1 GHz in figure. The plot in Figure 2.112 shows instead the relationship between the control current and the effective control voltage provided to NMOS starving devices. Frequency variations across different process corners are presented instead in Figure 2.113.

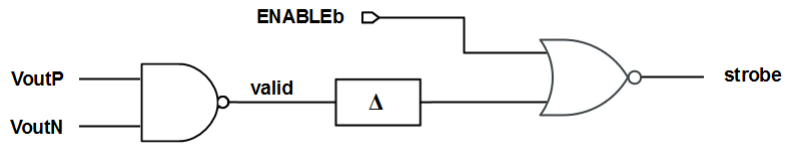


Figure 2.109: Insertion of a delay element in the *valid* asynchronous feedback path in order to reduce the intrinsic frequency of latch turned into an oscillator.

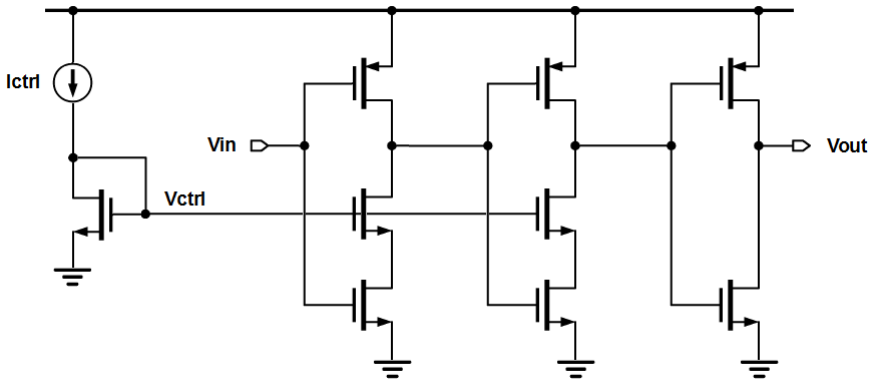


Figure 2.110: Practical transistor level implementation for the delay element. Current starving is performed with NMOS-only devices controlled by a current mirror.

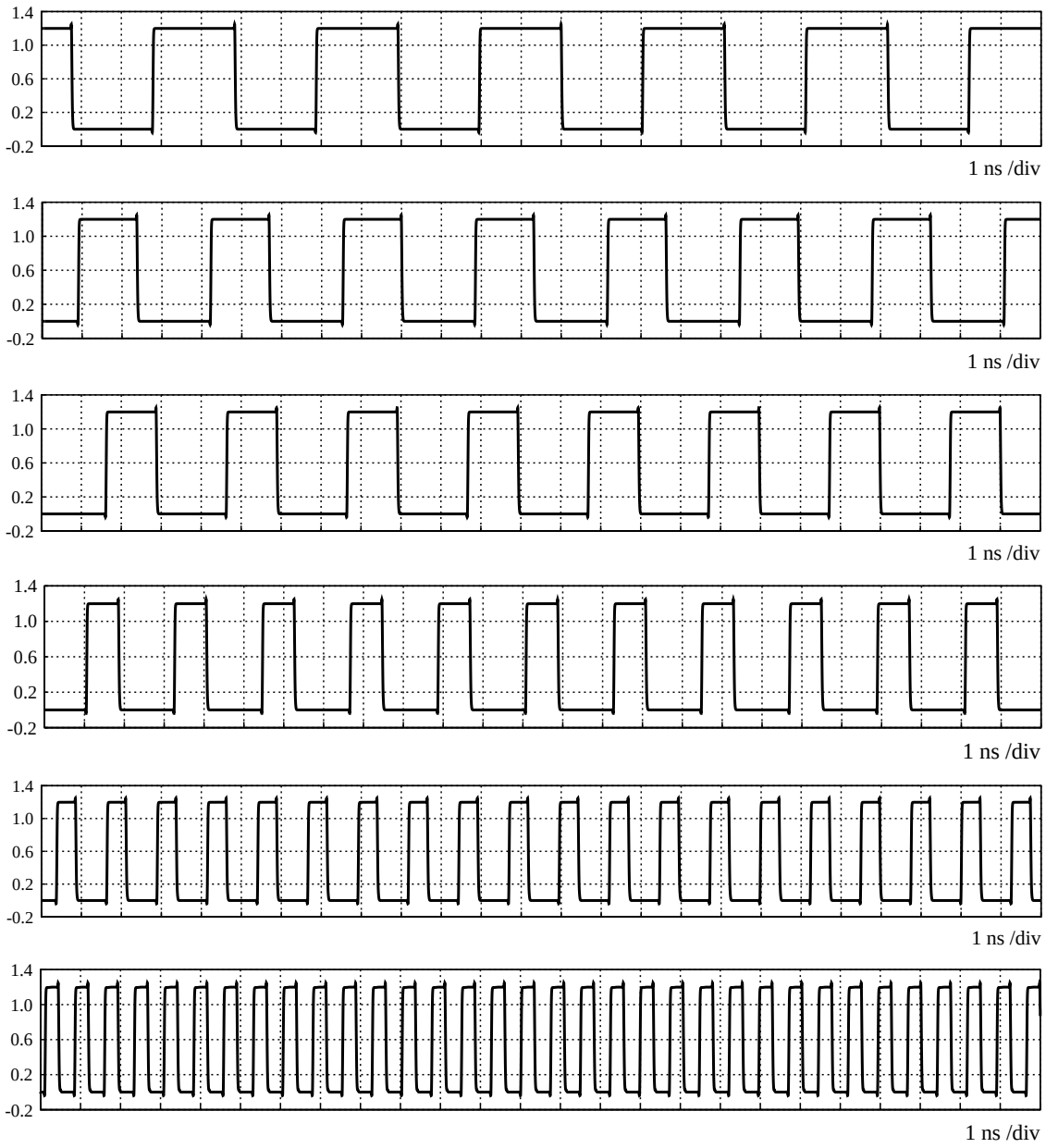


Figure 2.111: Sample of transient simulations for the *valid* clock signal by varying the control current. Selectable frequency values from 250 MHz up to 1 GHz are available thanks to the usage of a variable-delay element.

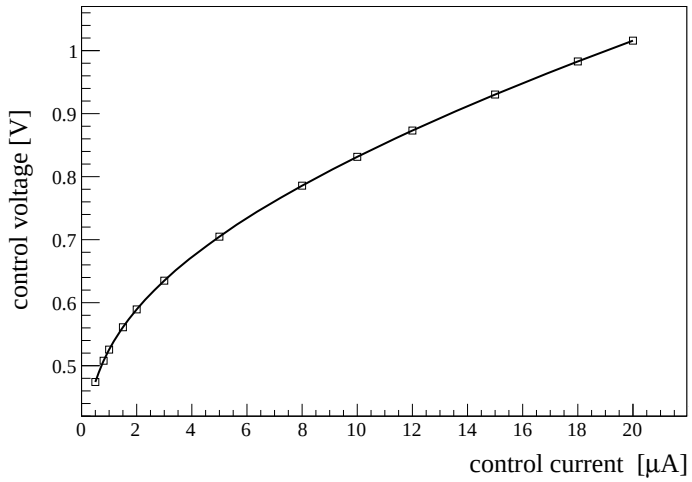


Figure 2.112: NMOS starving devices gate voltage as a function of the control current.

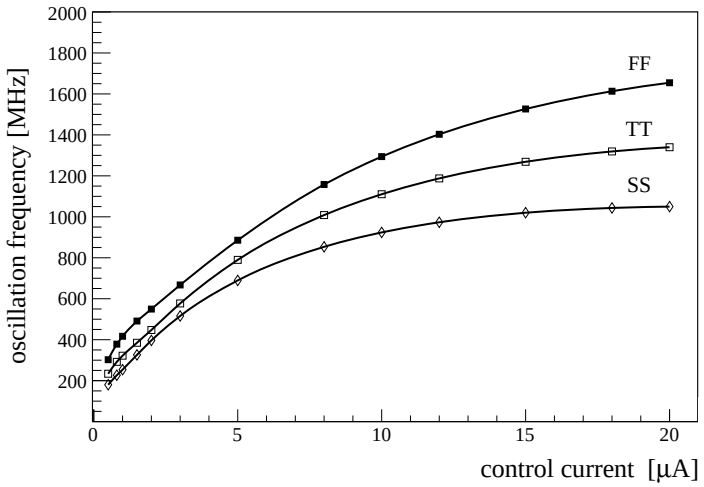


Figure 2.113: Oscillation frequency as a function of the control current across different process corners.

2.6 Analogue pixel cell layout

Figure 2.114 presents the final layout of the analogue pixel cell [Monteil 2014, Demaria 2014]. In agreement with specifications, it occupies a total area of 26 μm $\times$ 50 μm, about half the pixel size. Referring to the numbering scheme adopted in figure, the building blocks are: charge sensitive amplifier (1), Krummenacher feedback network (2), test charge injection circuit (3), capacitors used to mimic the sensor (4), comparator low-gain preamplifier (5), autozeroing capacitors (6), digital CMOS switches and buffers for necessary ϕ-signals (7), positive feedback latch (8), delay line (9) and AC coupling components between the Front-End amplifier and the overall discriminator.
The right side of the cell interfaces with the remaining on-pixel digital circuitry. Certainly the CSA and the feedback network are the most sensible analogue block. Hence they have been placed on the opposite side, whereas all switching components of the synchronous comparator fit on the right side. Furthermore, in order to prevent analogue power/ground lines to be corrupted by any digital activity, all switching blocks have been connected to digital supply rails.

With a traditional bottom-up design approach, each building block has been implemented as a full-custom block in respect of general analogue integrated circuit layout techniques such as gate fingering and source/drain sharing [Razavi 2000, Hastings 2001, Saint 2002].
All devices have been instantiated as parametrized cells from the 65 nm device library. Manual placement and routing have been then performed with the aid of schematic-driven and design-rule driven features. Dummy transistors have been inserted wherever needed to increase matching for most sensible devices and symmetrical structures. For better substrate noise isolation, all NMOS transistors involved in the Front-End amplifier signal processing are Deep N-Well (DNW) devices. Furthermore, an extensive usage of guard ring paths and substrate contacts in each block ensures better substrate homogeneity and minimize the susceptibility against latch-up. MOS transistors with the gate connected to a global bias line or to an external digital control signal are susceptible to antenna effects. Hence diodes have been inserted in order to prevent antenna-rules violations. No enclosed layout structures have been adopted instead. As described in the next chapter, due to the specific prototyping option adopted for the first submission, any design submitted using such an option must include the maximum number of metal layers (metal stack) offered by the fabrication technology for interconnections, resulting at the end in some unnecessary redoundance. A relaxed manual routing has been therefore available, without tight constraints on the number of metal layers used for the interconnections. Nevertheless, for global bias lines and external configuration bits a consistent scheme in the usage of metal layers along vertical and horizontal directions was chosen in perspective of chip integration. Via-doubling has been included as well to improve yield, as usually suggested by Design For Manufacturability (DFM) rules. Accurate DRC and LVS checks constantly validated the design of each block and interconnections between building blocks. For post-layout simulations, parasitic contributions have been extracted at the end of the pixel cell integration, including the bump-bonding pad contributions.
All capacitors have implemented as Metal-Oxide-Metal (MOM) devices offered by the 65 nm library. With such a choice, no low-level routing can be performed in correspondence of these passive devices, resulting into a significant area consumption. As an alternative, more expensive Metal-Insulator-Metal (MIM) can be adopted, with the advantage of a more compact layout thanks to a higher position in the metal scheme. Indeed, one has to consider that 25 fF, 50 fF and 100 fF input capacitors have been included only for test purposes to mimic a sensor. The effective active area is much smaller than 26 μm $\times$ 50 μm, with the clear possibility of leaving additional space for the on-pixel digital control logic in a second prototyping iteration.

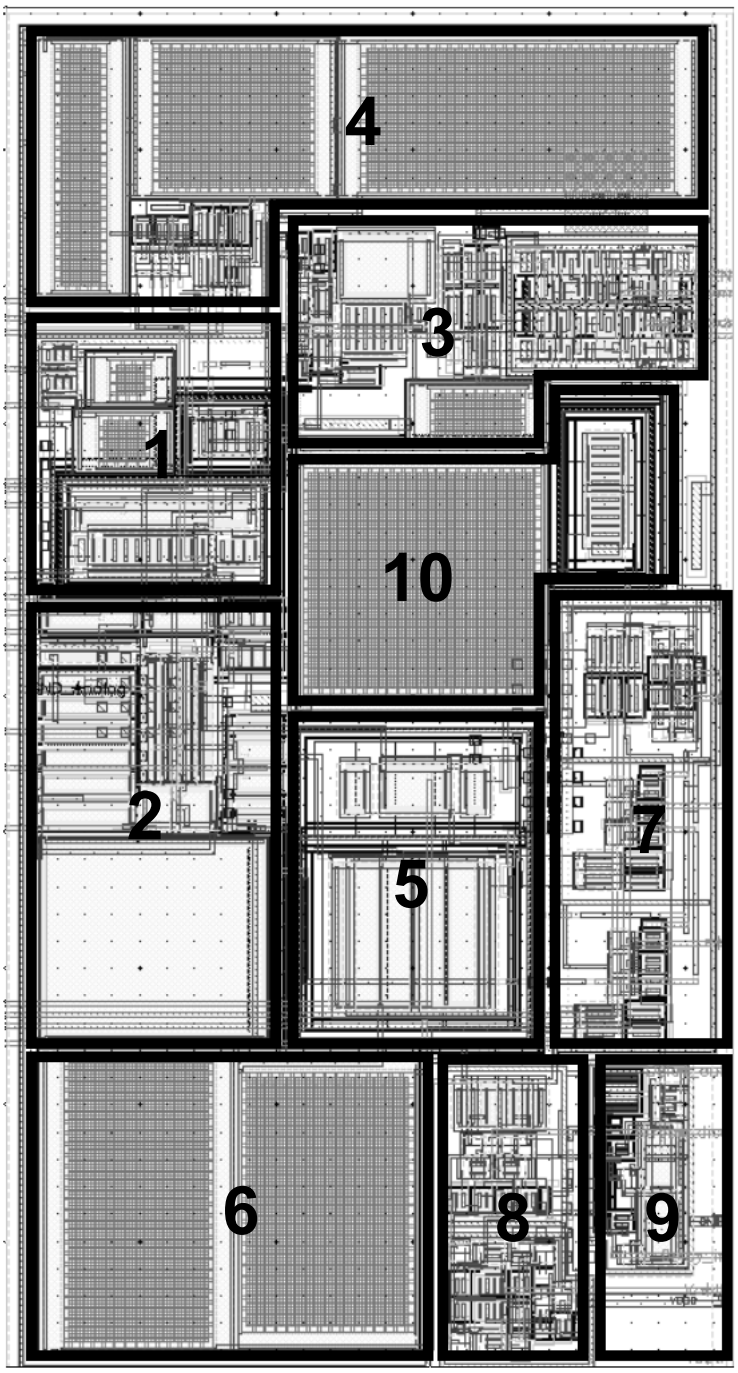


Figure 2.114: Complete analogue pixel cell layout, 26 μm $\times$ 50 μm. CSA (1) and Krummenacher feedback (2). Test charge injection circuit (3) with shunt input capacitors (4). Discriminator preamplifier (5). Autozeroing capacitors (6) and ϕ-signals CMOS switches (7). Latch (8) and delay line (9). AC coupling between Front-End amplifier and discriminator (10).

2.7 Latch control logic

Complexity and functionalities implemented by a on-pixel digital control logic really depends on the target application. Basic features include at least some processing for the hit signal coming from the Front-End discriminator and banks of D-type latches (registers) or RAM cells to store charge encoding data and configuration bits commonly employed in a pixel system for masking, test charge injection and local threshold adjustment. The choice of a specific readout architecture is basically dictated by the foreseen hit rates and trigger aspects instead. One would certainly minimize the data transmission from pixels towards the chip periphery. Nevertheless, area and power budget constraints can pose severe limitations for the total amount of digital intelligence and temporary local data storage that can fit into a single cell or in a pixel region. The analogue power consumption is essentially determined by the signal-to-noise ratio (SNR) and is easily predictable once the analogue blocks of the Front-End system have been defined. In contrast, the digital power only depends on the switching activity, that can be quite different even for the same architecture working under different occupancy conditions. Moreover, the usage of automated synthesis and place-and-route (PNR) tools can lead to significant area and power contributions due to unnecessary buffering if speed constraints are overestimated. Further limitations can be addressed by radiation hardness requirements, demanding for the usage of specific Single Event Upset (SEU) protections and error correction techniques that introduce redundancies and additional logic in the pixel cell for most important configuration bits and state flags. Certainly, the absence of a on-pixel D/A converter for channel-to-channel threshold corrections is a fundamental advantage offered by an autozeroed comparator solution, avoiding the necessity of dedicated SEU-tolerant registers to store the configuration bits for the digital trimming and increasing the available area for temporary data storage and signal processing that can be put into the digital part.
Once specifications are defined, detailed simulations supported by synthesis and PNR explorations are necessary in order to choose the best compromise among different pixel cell architectures. The overall concept, implementation and optimization of the digital component represent therefore a large design effort. As mentioned in Chapter 1, this is now part of a dedicated research program within the RD53 collaboration framework [Marconi 2014, Conti 2015]. The choice of a 65 nm CMOS process gives a concrete chance to implement new interesting solutions, benefiting of performance increase for digital integrated circuits in terms of speed and integration densities offered by technology scaling. However, in perspective of a first INFN submission the primary goal has been to validate the feasibility of performing fast TOT measurements by means of synchronous Front-End operations. Thus only a bare minimum digital circuitry has been included in each pixel cell, remanding to a second iteration the development of a substantial amount of synthesized logic.

The implemented control logic only provides the necessary support to turn the latched comparator of the proposed synchronous Front-End into a local oscillator, generating high-frequency clock signals for fast TOT digitizations.
The basic idea is to use a multiplexor that toggles the discriminator *strobe* control signal between some slow and fast operating modes. According to the synchronous Front-End approach, during normal operations (slow mode) the comparator latch receives an external 40 MHz master clock and samples the CSA output at the peaking time. If no signal is found above the threshold voltage, then the latch continues in receiving the 40 MHz clock. If a signal exceeds the threshold indeed, the multiplexor switches such that an asynchronous logic feedback loop is established, turning the discriminator into a local oscillator (fast mode) and providing a self-generated, high-frequency clock signal that continues in sampling the voltage difference at latch input nodes. The oscillation can be then tuned by means of a voltage-controlled delay line, as discussed in previous sections. After the analogue signal returns to the baseline and goes below the threshold voltage the feedback path must be broken, turning the latch back into slow operations or more realistically introducing some idle or busy states that wait for data readout.

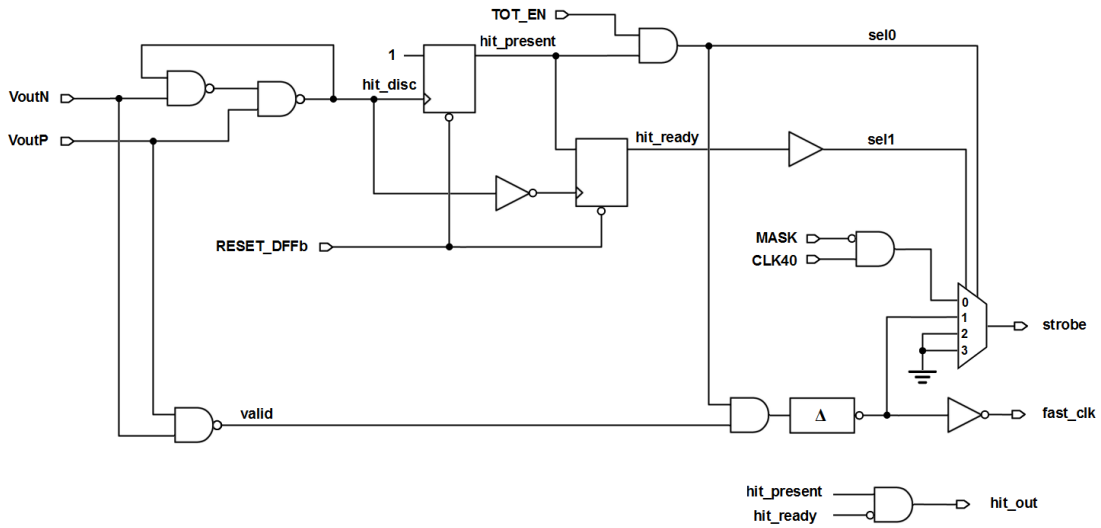


Figure 2.115: Schematic of the proposed latch control logic. The delay box Δ is implemented as a voltage-controlled delay line to tune the frequency of the resulting self-generated clock for fast TOT digitizations. Note the inverting behaviour of the delay line for proper latch/oscillator operations.

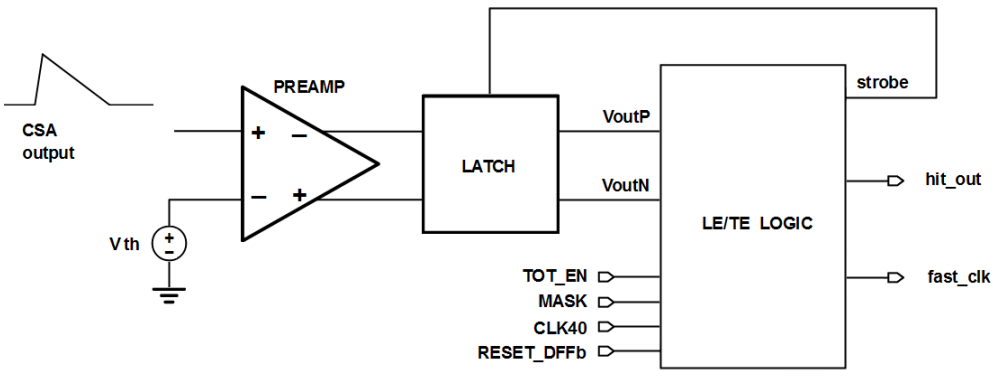


Figure 2.116: Block diagram of the synchronous comparator coupled to its controller.

The digital circuitry implemented to achieve the above described functionality is presented in the schematic of Figure 2.115. Figure 2.116 shows instead a block diagram of the synchronous comparator coupled to its controller. The control logic receives as inputs the output signals *VoutP* and *VoutN* generated by the latch in the analogue part. As discussed, depending on the latch decision *VoutP* and *VoutN* toggle at each strobe cycle. In case a particle hit has been detected above the threshold voltage, a CMOS digital pulse *hit_disc* is obtained from *VoutP* and *VoutN* transitions by means of a NAND-based S/R latch. Furthermore, the *valid* signal is generated.
At the core of the logic resides a 4:1 multiplexor whose output is fed to the comparator as *strobe*, providing the latch main control signal. The least significant input (0) receives the 40 MHz master clock *CLK*40 for normal operations. A *MASK* configuration bit has been included to perform pixel masking if required. When *MASK* is asserted the master clock is not propagated to the multiplexor, hence the latch is always disabled. The second input of the multiplexor (1) is used to close the feedback loop in order to turn the latch into a local oscillator, receiving a delayed *valid* signal generated from discriminator outputs and feeding it back to the latch. The remaining two inputs are tied low and are used to implement in a compact way two different idle states. In both cases the synchronous discriminator is disabled because no switching signal is propagated to the latch. The most significant input (3) is used to stop operations after a hit has been detected if fast TOT digitization has been enabled by means of a *TOT_EN* configuration bit. The remaining input (2) is used to implement a second idle state if a binary-only operating mode is selected indeed. That is, if *TOT_EN* is set to low the latch never can turn into an oscillator, the functionality of the 4:1 multiplexor reduces to a simpler 2:1 multiplexor and *strobe* can only be *CLK*40 or tied low. In order to guarantee low-power operations, proper vetoing has been included in the feedback loop such that the high frequency *valid* signal is fed to the delay line only if the TOT mode is enabled and a hit has been detected. Control signals *sel*0 and *sel*1 for the multiplexor are essentially generated by means of state FlipFlops. As usually performed in most of pixel controllers, the digital pulse *hit_disc* is fed to a couple of complementary edge-triggered D-type FlipFlops (DFFs) that register leading-edge and trailing-edge transitions, resulting into two state flags *hit_present* and *hit_ready*. The digital TOT information can be finally obtained using a simple binary counter that receives as input the *fast_clk* signal produced during feedback operations. If required, an additional count-enable control signal can be generated from leading-edge and trailing-edge hit flags. As determined from CAD simulations, extremely fast TOT digitizations with 5-8 bit resolutions can be achieved using self-generated clocks up to GHz frequencies.

The proposed pixel control logic certainly has some relevant limitations. At first, no temporary data buffering has been implemented. A simple external *RESET_DFFb* signal is used to properly initialize and reset state FlipFlops, thus limiting the processing of the entire system to a single event. Only one hit pulse at a time can be generated by the synchronous comparator and registered by the logic. Hence a new hit can not be generated until a new external *RESET_DFFb* cycle is provided to DFFs. As a result, *RESET_DFFb* defines the overall acquisition time in perspective of simple bench characterizations at the oscilloscope. In a more complete design instead, after the hit information has been retrieved and stored, some busy flag must inform a suitable control logic located in the pixel or at the chip periphery that data are ready, waiting for further instructions. At some time, specific reset commands must be internally generated to properly reset *hit_present* and *hit_ready* flags such that the base 40 MHz clock can be again propagated to the comparator latch. Additionally, no dedicated control logic has been included in the pixel cell for generating necessary ϕ-signals that trigger the offset compensation in the discriminator. In the submitted designs in fact, ϕ-signals have been simply generated by means of asynchronous delays and coincidences at the chip periphery.

TOT_EN	*hit_ready*	*hit_present*	*sel*1	*sel*0	*strobe*
0	0	0	0	0	*CLK*40
0	0	1	0	0	*CLK*40
0	1	1	1	0	0
1	0	0	0	0	*CLK*40
1	0	1	0	1	*fast_clk*
1	1	1	1	1	0

Table 2.7: Logic values assumed by internal flags and *strobe* signal under slow and fast operating modes. Logic 0 is digital ground *GNDD* and logic 1 is the digital supply voltage *VDDD*.

Steady logic values assumed by internal control signals under slow and fast operating modes are summarized in Table 2.7. A sample Verilog HDL behavioural simulation of the functionality provided by this solution is presented in the timing diagram of Figure 2.117.
Transient analyses reported in Figure 2.118 shows instead a full transistor-level simulation of the digital control logic coupled to the analogue Front-End chain when $TOT\_EN = 0$ is selected. Nominal values in terms of feedback capacitance (4 fF), detector capacitance (100 fF) and feedback current (40 nA) have been assumed for the analogue Front-End. The analogue waveform has been obtained by injecting a test charge of 4 ke^- at the CSA input node, thus requiring proper timing between *RESET_DFFb* and the *TestP* signal that triggers the calibration circuit. Furthermore, the charge injection must be synchronized with the 40 MHz master clock. The hit pulse is generated at the peaking time of the CSA output and its duration is an integer number of *CLK*40 cycles. As already discussed, in such a configuration a slow TOT encoding can be simply performed by means of the 40 MHz clock. More realistically, $TOT\_EN = 0$ can be used to disable the binary counter as well, thus reducing the Front-End to a true binary-only system.
Under the same analogue Front-End configuration, transient simulations in Figure 2.119 have been obtained enabling latch operations as a local oscillator, thus setting $TOT\_EN = 1$. Furthermore, different self-generated clock waveforms obtained by trimming the control current for the delay line are presented in Figure 2.120. As one can see, flexible charge encoding resolutions with locally generated clocks of 200-600 MHz are available thanks to the usage of a variable-delay element in the asynchronous logic feedback path around the latch.

Despite its simplicity, practical design, simulation and layout of this solution introduced a few important issues. On the one hand, due to the initial unavailability of a tapeout design kit in 65 nm technology all required CMOS logic gates have been re-designed from scratch [Uyemura 2002, Rabaey 2003, Kang 2003]. Such a design effort has been shared among different design teams of the CHIPIX65 collaboration, developing a full-custom basic digital library and providing schematic, symbol and layout views for each cell. Moreover, transistor-sizing guidelines proposed by the RD53 collaboration upon preliminary radiation-hardness qualification results have been satisfied. These cells were used in other different designs as well. On the other hand, due to the intrinsic asynchronous nature of the digital circuit, special attention was paid in timing verification. The entire logic has been assembled and simulated into the full-custom design environment, optimizing the interfacing with the comparator latch. Buffers were inserted wherever needed in order to prevent any race conditions and glitches in the logic feedback loop that establishes around the latch, simulating the circuit behaviour across PVT corners and MC runs.

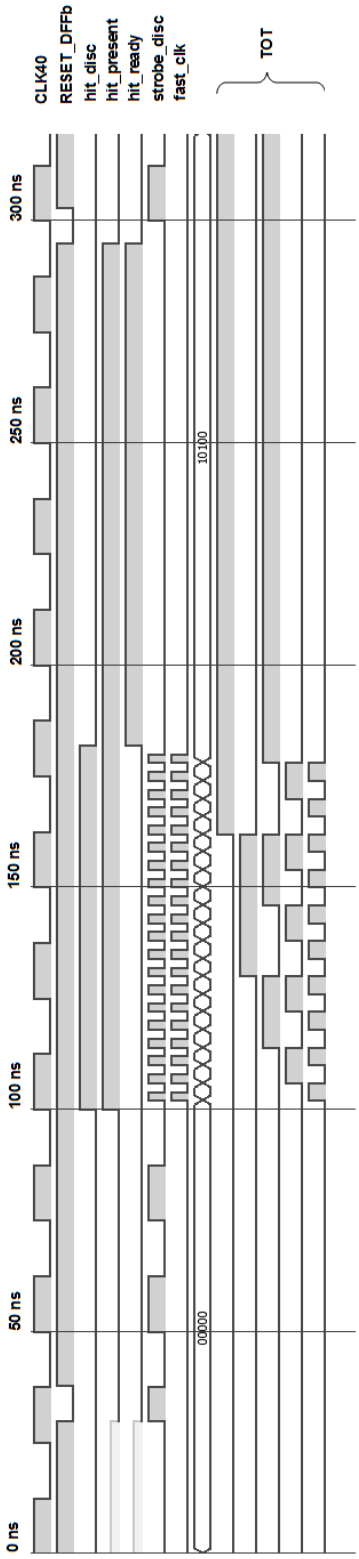


Figure 2.117: Verilog HDL behavioural simulation of the functionality provided by the latch control logic. The generation of the hit pulse is synchronized with a positive edge of the master clock, then leading-edge/trailing-edge state flags are used to control multiplexor selection bits and switch between slow, fast and idle operations. A 5-bit resolution TOT word is finally retrieved with a 5-bit binary counter.

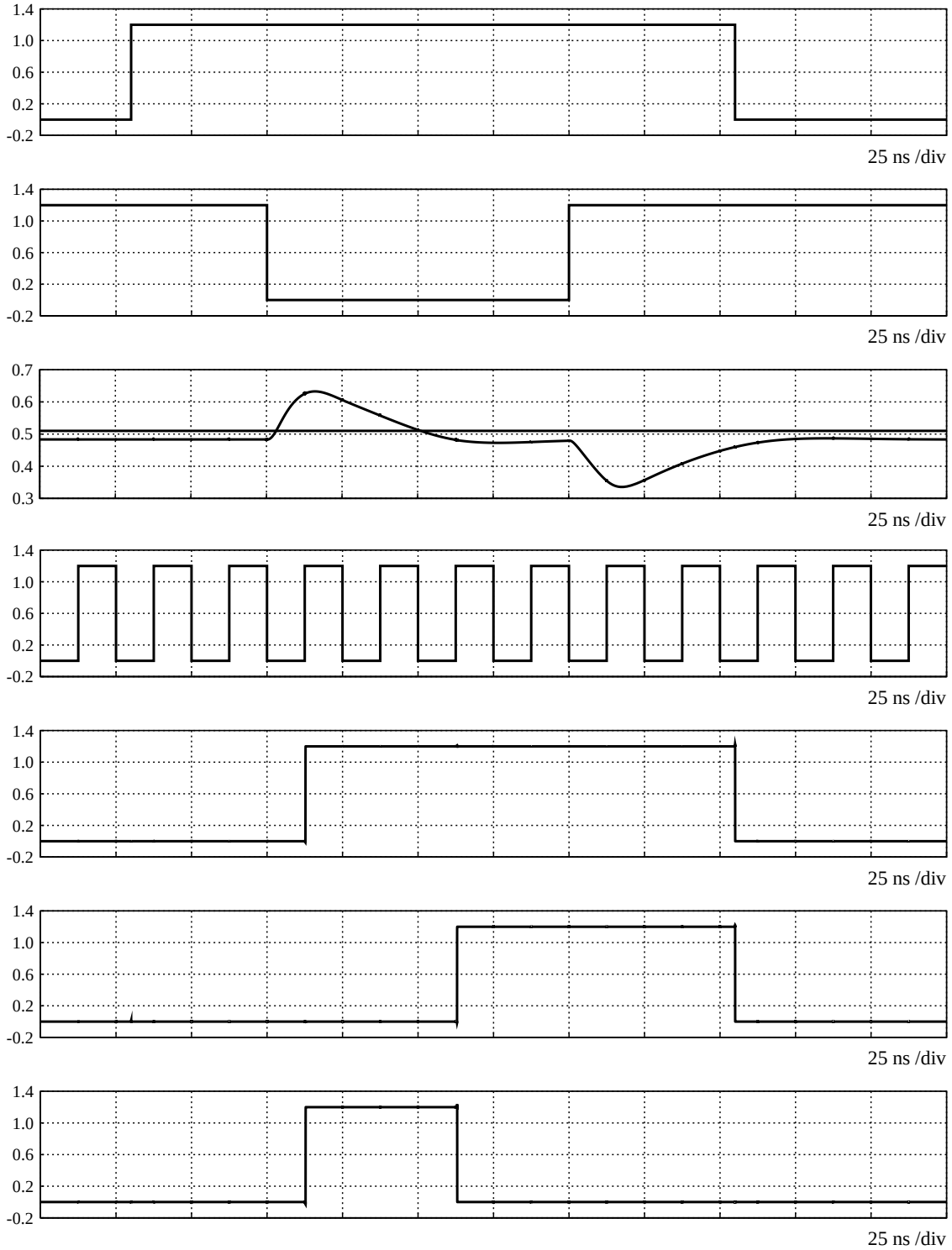


Figure 2.118: Full transistor-level simulation of the latch control logic coupled to the analogue Front-End chain when $TOT\_EN = 0$. From top to bottom: global reset *RESET_DFFb*, *TestP*, CSA analogue waveform and global threshold, $CLK40$, *hit_present*, *hit_ready* and *hit_disc*. The threshold voltage is 10 mV above the baseline.

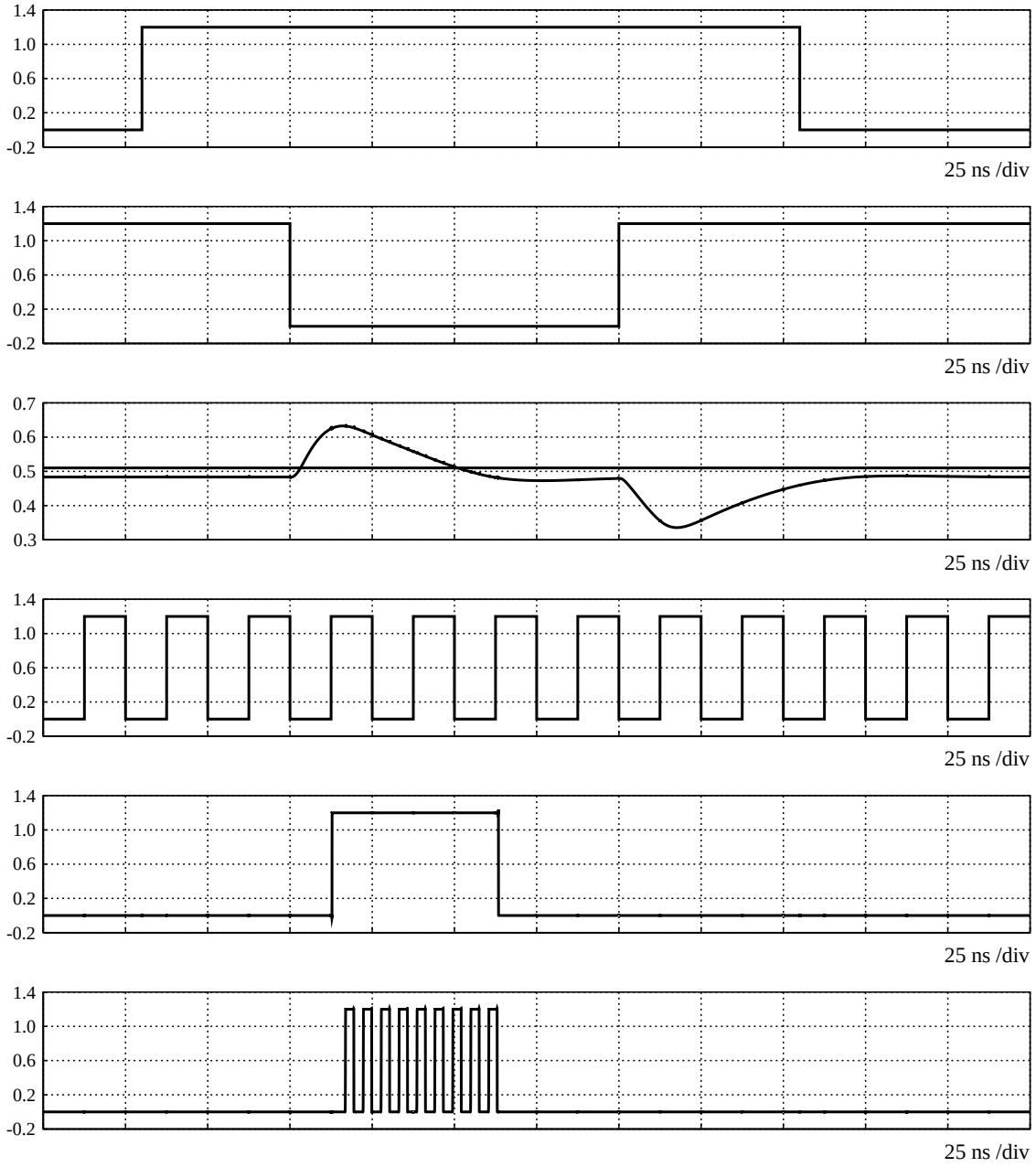


Figure 2.119: Full transistor-level simulation by enabling latch operations as a local oscillator. From top to bottom: *RESET_DFFb*, *TestP*, CSA analogue waveform and threshold, *CLK*40, *hit_disc* and *fast_clk*.

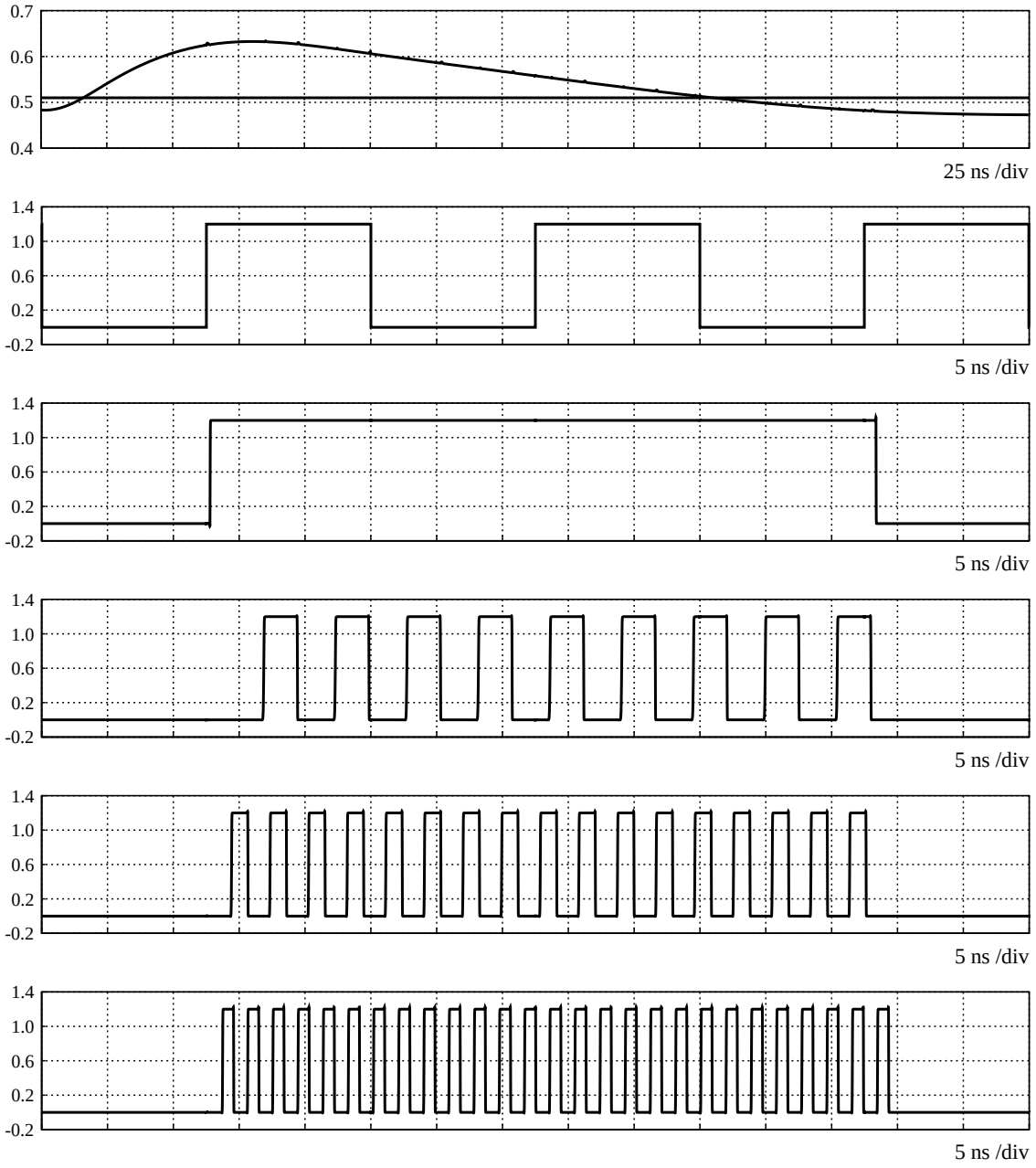


Figure 2.120: Simulated self-generated clock waveforms for different values of the control current in the delay line. From top to bottom: CSA analogue waveform and global threshold, $CLK40$, *hit_disc* and three different *fast_clk* signals by increasing the control current (50 nA, 150 nA and 300 nA).

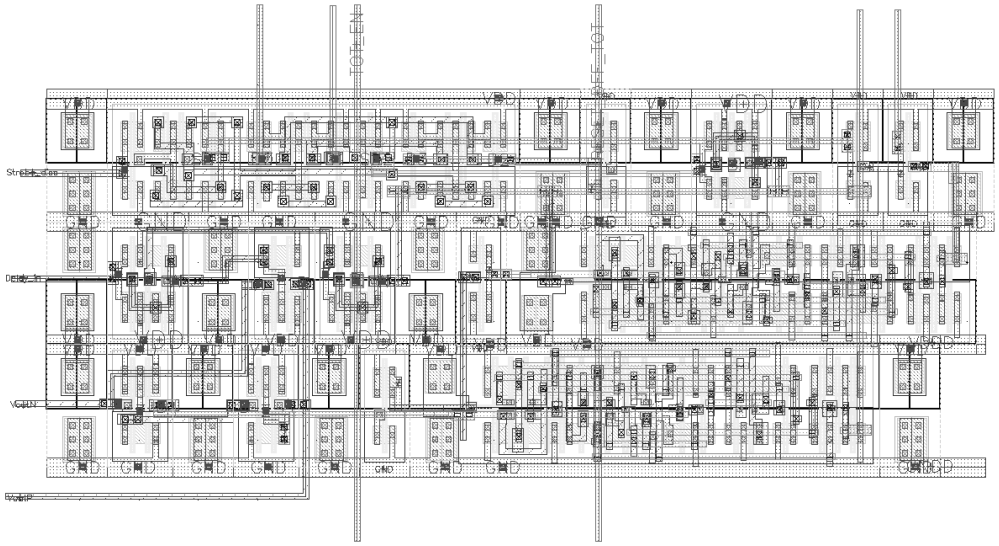


Figure 2.121: Latch control logic final layout (approx. 15 μm $\times$ 8 μm). Standard-cell based design using full-custom cells and manual placement and routing. The voltage-controlled delay line lies in the analogue part.

The final layout of the latch control logic is presented in Figure 2.121. It has been implemented according to a standard-cell design approach [Sicard 2007] without the support of any automated procedures. Full-custom logic gates have been laid out as 2.4 μm fixed-height, variable-width cells enclosed within 0.33 μm height horizontal metal stripes, resulting into larger cells with respect to those provided by 65 nm digital libraries. However, area was not an issue due to the limited number of gates and with about half of the total pixel size assigned to digital components. Cells were arranged into three rows, following a schematic-driven placement. Power/ground rails abutment and vertical cell mirroring have been adopted for a more realistic and compact design. Full-custom tap cells were also inserted between adjacent logic gates to provide well and substrate contacts to minimize latch-up susceptibility. Furthermore, custom tie-high and tie-low cells has been designed to implement 0/1 constant logic values, avoiding direct gate connections to the power/ground networks in order to increase the reliability against Electrostatic Discharge (ESD) surges. A relaxed manual routing has been performed, without tight constraints on the number of metal layers used for the interconnections. Nevertheless, for global signals and external configuration bits a consistent scheme in the usage of metal layers along vertical and horizontal directions was chosen in perspective of chip integration. Moreover, via-doubling has been included to improve yield as usually suggested by Design For Manufacturability (DFM) rules. Accurate DRC and LVS checks validated the final design and parasitic contributions have been extracted.
The layout of the digital block occupies a total area of about 15 μm $\times$ 8 μm and interfaces to latch outputs in the analogue part. One notable aspect is that the layout of the layout of the delay line has been integrated into the analogue part. Furthermore, in order to prevent analogue power/ground supply rails to be corrupted by the intense digital switching activity during fast TOT encoding, both latch and delay line were connected to digital power/ground lines. Hence the latch coupled to its voltage-controlled delay line de facto represents a true digital block embedded in the analogue Front-End.

Further details on digital components will be introduced in Chapter 8, describing the assembling of Torino INFN small pixel arrays that were put on silicon as part of the October 2014 CHIPIX65 submission. A first matrix contains 8 × 8 pixels. A simple pixel addressing provides for all cells full-analogue access at the oscilloscope to both the CSA waveform and one selectable discriminator output (the hit pulse or the self-generated clock) with no support for a digital data readout. A second array of 8 × 12 pixels has been equipped instead with a shift-register based digital serial readout and configuration, retrieving TOT information in digital form by means of on-pixel binary counters. The basic functionality of the hit logic has been also re-adapted for a third small pixel matrix of 8 × 12 pixels implementing a continuous-time Front-End solution proposed by Pavia INFN and equipped with the same serial readout.

In the following, most significant performance simulation results in terms of TOT linearity and stability across process corners and MC runs are presented. Figure 2.122 shows the number of TOT counts as a function of the input charge for three different process corners TT, FF and SS. The Front-End amplifier has been configured with nominal values in terms of feedback capacitance, input detector capacitance and total feedback current. Under these assumptions, the pulse duration for a maximum signal of interest of 30 ke^- is rough 300 ns. Hence the delay line was tuned such that a self-generated clock of about 650 MHz is fed to an 8-bit binary in order to retrieve a maximum TOT of 255 counts within this time scale. As one can see, TOT counts exhibit high linearity with the input charge. Certainly, due to uncompensated frequency variations the number of counts varies across different corners, as expected. For a fixed 4 ke^- input charge instead, variations in the number of TOT counts across 100 different transient MC runs are presented in Figure 2.124 and 2.124. With respect to a nominal value of 40 counts, the spread in the number of TOT counts is about 15%. Certainly the usage of a PLL in order to mitigate frequency variations will be addressed only after an experimental validation of the feasibility and reliability of the proposed TOT solution.

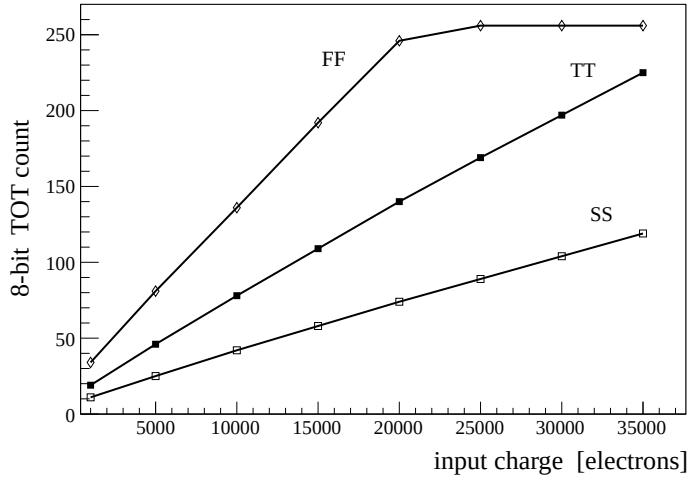


Figure 2.122: Number of TOT counts as a function of the input charge for different process corners. The frequency of the self-generated clock has been tuned to 650 MHz (TT).

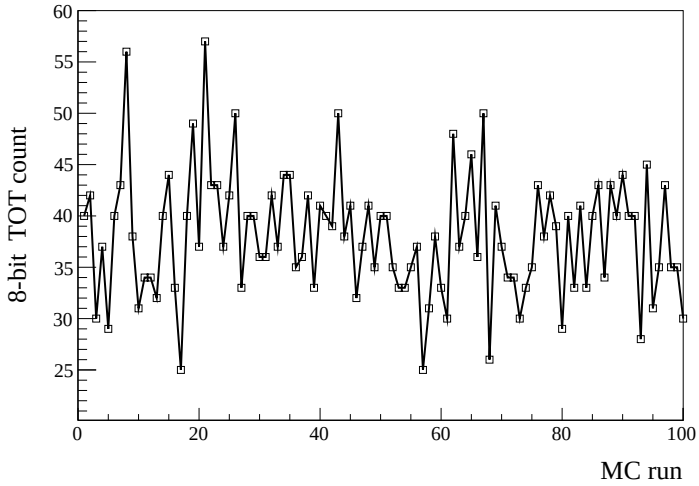


Figure 2.123: Number of TOT counts across 100 different transient MC runs assuming 4 ke^- input charge, 100 fF input capacitance, 4 fF feedback capacitance and 40 nA feedback current. The frequency of the self-generated clock is 650 MHz (TT).

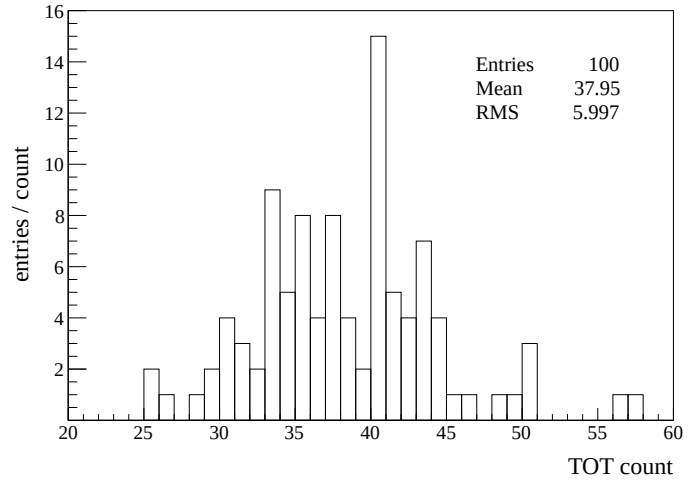


Figure 2.124: Distribution of the number of TOT counts presented in Figure 2.124. Frequency variations due to random fluctuations in MOS device parameters introduce a spread in the number of counts of the order of 15% RMS.

2.8 Summary

The design of a synchronous pixel Front-End chain in 65 nm CMOS technology has been discussed. The primary goal of the overall architecture optimization was to meet all CMS-specific performance requirements and general guidelines defined within the RD53 collaboration for the pixel upgrades at HL-LHC. The usage of a discrete-time comparator introduces several advantages. The hit generation is synchronized with a 40 MHz clock. Pixel-to-pixel threshold variations have been compensated by means of Output-Offset Storage, avoiding the necessity of a on-pixel D/A converter for digital trimming. The latch can be easily turned into a local oscillator, as extensively performed in modern SAR A/D converters. Transient simulations suggest that 5-8 bit TOT measurements can be retrieved using locally generated clock signals up to GHz frequencies thanks to speed offered by a 65 nm technology node. Most important Front-End parameters are summarized in Table 2.8.

Parameter	Value
pixel size	50 μm $\times$ 50 μm
total analogue area	26 μm $\times$ 50 μm
active analogue area	26 μm $\times$ 39 μm
analogue supply voltage	1.2 V
total DC analogue current	3.75 μA (CSA + PREAMP)
static analogue power consumption	4.5 μW at 1.2 V supply voltage
dynamic power consumption	$\approx$ 1 μW
CSA feedback capacitance	selectable, 2.5 fF, 4 fF or 6.5 fF
Krummenacher feedback current	1 nA - 100 nA
nominal peaking time	12.5 ns at 4 fF feedback capacitance and 40 nA
minimum detectable charge	$\approx$ 900 e^-
charge-to-voltage linear range	8-10 ke^-
TOT linearity	up to 40 ke^-
PSRR	$<$ 0 dB at all frequencies
noise	ENC $\approx$ 90 e^- RMS at 100 fF input capacitance
latch dynamic offset	22 mV, $\approx$ 75 e^- RMS
TOT resolution	5-8 bit
self generated clock frequency	selectable, 100-900 MHz

Table 2.8: Summary table for most important synchronous Front-End parameters.

Chapter 3

The first CHIPIX65 submission

In this chapter, a description of the first INFN CHIPIX65 submission in 65 nm CMOS technology is given. Thanks to a special prototyping option offered by the foundry access service for small ASIC designs, three different silicon dies of area 1.96 mm $\times$ 1.96 mm have been submitted by the CHIPIX65 collaboration to the manufacturer on October 2014. They were received back from the foundry for test measurements and bench characterizations at the beginning of 2015. According to the aim of this work, most of the contents of the chapter is dedicated to design aspects with personal contributions.

Keywords: IC submission, IP block, pixel matrix, floorplan, I/O cell, pad frame, analogue readout, power distribution, clock tree, chip integration, ESD, wire-bonding, segmented DAC, decoder, Verilog, STD cell, synthesis, PNR

3.1 Introduction

The first submission represented a fundamental milestone for the CHIPIX65 project as well as for the Torino CMS tracker group. Without doubt it offered a great human experience in terms of sharing efforts, knowledge and design expertise among the different INFN units involved in 65 nm design activities. Despite a few designs were already put on silicon in the past by some INFN groups now part of CHIPIX65, the submission validated up to the foundry transfer all aspects of both full-custom and digital design flows attached to the 65 nm CMOS process. Moreover, INFN is one of the founding members of the CERN/RD53 international collaboration. The first CHIPIX65 submission offered therefore an important opportunity for promoting first Italian contributions and design efforts in 65 nm CMOS technology within this framework.
The submission included one die dedicated to preliminary pixel electronics prototypes (single-channel test structures and small pixel matrices implementing different Front-End solutions) and two dies dedicated to more general purpose analogue, digital and mixed-signal building blocks, commonly referred to as Intellectual Property (IP) blocks[1]. Personal contributions to the October 2014 submission are both in the design, assembling and sign-off of pixel Front-End chips and to IP block design.

[1] In semiconductor industry and electronic design automation, the term Intellectual Property core (IP core or IP block) refers to a re-usable analogue, digital, mixed-signal block or full-chip design that can be included as a building block into ASIC designs or FPGA logic designs as well as in PCB design. Example IPs are A/D and D/A converters, Phase Locked Loops (PLLs), high-speed I/O interfaces, RAM banks, microprocessors, discrete integrated circuits. Depending on the type of the block, IPs can be in form of a physical layout (hard IP) or a synthesizable HDL code or a gate-level netlist (soft IPs). Similarly to software, IP blocks are released under specific licenses and are copyrighted by a silicon manufacturer or by who designed the block itself. Hence the usage of IPs may require the payment of fees and is subjected to severe legal issues and limitations imposed by the IP owner.

After a short description of the prototyping strategy adopted by the CHIPIX65 collaboration and of the different designs included into the first submission, most of the chapter is dedicated to the assembling of two pixel Front-End ASICs. Common aspects involved in a chip-integration (design floorplan, I/O pad frame assembling and power partitioning, clock distribution, ESD protection, wire-bonding floorplan and pad assignment) up to final full-chip DRC/LVS and sign-off checks before the foundry transfer are therefore discussed.
The last part of the chapter is dedicated to personal contributions to IP block design instead. It describes a standard-cell based implementation of a binary-to-thermometer column/row decoder for a segmented 10-bit current steering D/A converter. This block was then put on silicon as part of a dedicated IP block chip. All aspects of the design flow are considered, covering the optimized synthesis of a behavioural Verilog HDL description of the block, the automated place-and-route (PNR) in the digital design environment and the final design import and verification back into the full-custom design environment. The design of this simple block validated for the first time in Torino the automated digital implementation flow attached to the 65 nm CMOS technology.

3.2 Prototyping strategy

Several considerations and practical design aspects determined the prototyping strategy and time schedule for a first submission using 65 nm CMOS.
As a matter of fact, it was of primary importance for the CHIPIX65 collaboration to submit both first prototypes of pixel Front-End electronics and of IP block designs, providing a necessary step towards the design of a future more complex hybrid pixel Front-End ASIC demonstrator and a first iteration for RD53 IP blocks with strong INFN commitment. On the one hand, special attention has been devoted to the design of first pixel Front-End structures, exploring as much as possible innovative solutions and different architectures for the analogue signal processing and for the charge encoding. RD53 input specifications and design guidelines have been satisfied, as well as more CMS-specific pixel detector upgrade requirements for HL-LHC. On the other hand, the implementation, optimization and radiation qualification of reliable building blocks with specific functionality for services, configuration and data readout play a fundamental role in the design of any complex pixel ASIC system. Hence large design efforts have been focused to IP blocks with INFN contribution as part of the RD53 research program.

The choice of the prototyping option adopted by the collaboration introduced practical constraints in terms of available area, layout and time schedule. Different fabrication solutions are offered for manufacturing by foundry access services. IC fabrication costs are extremely high in terms of wafer and mask processing. Thus foundries usually offer prototyping and small production volumes through Multi-Project Wafer (MPW) runs, integrating on the same silicon slice a certain number of different designs from various teams, research institutes and universities such that costs are shared among participants. With a minimum chunk of area of the order of 10-12 mm^2, MPW is usually the standard approach adopted by the HEP VLSI design community.
However, as a result of a special prototyping solution offered by the foundry access service for small ASIC designs with lower prices with respect to traditional MPW runs, three different chips of area 1.96 mm $\times$ 1.96 mm have been submitted on October 2014 to the silicon manufacturer by the CHIPIX65 collaboration. Such a solution introduced some important design aspects. On the one hand, any design submitted using this option must include the maximum number of metal layers (metal stack) offered for interconnections by the chosen 65 nm CMOS process, resulting at the end in some unnecessary redoundance. On the other hand, a drawback in the choice of this option is in terms of availability, being limited to only twice a year and thus posing tight constraints in the time schedule. The number of required chips has been carefully investigated with extensive preliminary floorplan studies, exploring advantages and disadvantages according to the foreseen number of IP blocks ready for the submission and pixel Front-End structures to be included.

IP block	Specifications/notes	INFN commitment
4x bandgap references	BJT/diode/MOS based	Bergamo/Pavia
1x bandgap reference	1 V supply voltage	Milano
2x current D/A converters	10-bit segmented architecture, synthesized or full-custom decoding logic	Bari/Torino
3x SRAM arrays	64 kb/bank, full-custom DICE	Milano
2x Ser/Des	1.6 GHz, 20-bit, SEU-tolerant, synthesized	Pisa

Table 3.1: INFN IP blocks included in the first October 2014 CHIPIX65 submission. Additional designs from CERN and other groups part of the RD53 collaboration are not listed.

These studies demonstrated that a single IP blocks chip would have introduced severe limitations for the maximum number of available I/O pads for each block, thus significantly reducing most of the testability. Hence the collaboration opted for two different chips dedicated to IP designs and one chip to integrate different pixel Front-End solutions.
A first chip, referred to as CHIPIX_BIAS and shown in in Figure 3.1, was dedicated to biasing integrated circuits. It included different versions of temperature and supply independent bandgap voltage references and current-steering D/A converters [Johns 1996, Razavi 2000]. The available silicon area was also shared to include IP designs from other teams part of the RD53 collaboration (CERN, CPPM and Prague University), offering a good feasibility example of sharing design efforts and manufacturing costs across CHIPIX65 and RD53 collaborations. Biasing circuits are fundamental building blocks of any integrated system, allowing to establish precise and reliable DC operating conditions in terms of voltage or current. The investigation of the effects of high radiation levels on such circuits is therefore of primary importance in determining the reliability of the 65 nm CMOS technology against extreme radiation levels foreseen at HL-LHC.
A second chip of IP blocks, referred to as CHIPIX_SRAM and shown in Figure 3.2, includes two standard-cell based prototypes of 20-bit, 1.6 GHz, Single Event Upset (SEU) tolerant high-speed Serializer/Deserializer (Ser/Des) blocks [Athavale 2005] and three banks of 256 $\times$ 256 (64 kbit) CMOS Static-RAM (SRAM) cells [Kang 2003, Rabaey 2003] radiation-hardened by design and implemented as Dual-Interlocked Storage Cells (DICE) [Calin 1996, Lee 2011]. The chip aims to investigate performance degradation and susceptibility to radiation-induced Single Event Effects (SEEs) of digital integrated circuits in the chosen 65 nm CMOS fabrication process.
A summary of INFN IP blocks put on silicon in the first CHIPIX65 submission is presented in Table 3.1. Further ongoing design activities on additional blocks with INFN commitment (e.g. A/D converter, SLVDS drivers, low-power clock drivers etc.) have already started and are foreseen to be manufactured throughout 2015.

The third chip included in the submission, referred to as CHIPIX_VE1, has been entirely dedicated to first pixel Front-End electronics solutions. The following sections describe fundamental design aspects and personal contributions to the top-level chip integration and to the assembling of two small pixel matrices proposed by the Torino INFN group.

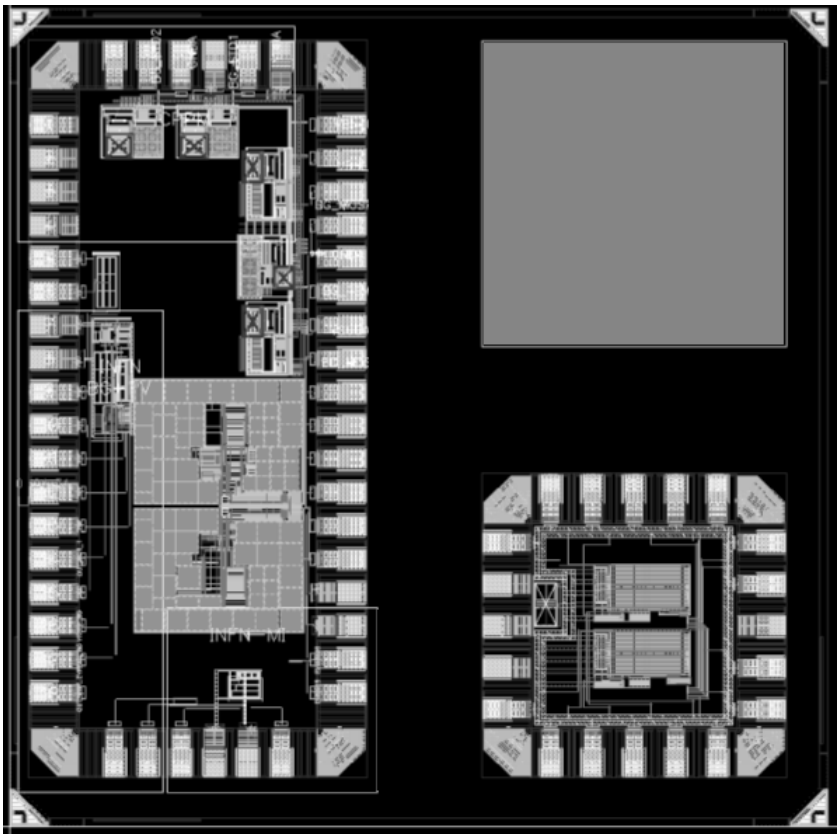


Figure 3.1: CHIPIX_BIAS, 1.96 mm × 1.96 mm. On the same silicon die have been included three independent chips. The left chip hosts different versions of bandgap voltage references, whereas two current-steering D/A converters are placed in the bottom-right chip. The empty space at the top-right hosts an additional D/A converter chip designed by another group of the CERN/RD53 collaboration. Personal contributions to the IP blocks submission involve the design of a binary-to-thermometer column/row decoder for the DAC chip. Note that at the end of the manufacturing process the silicon die was cut and each chip has been *packaged*. The gap among adjacent chips is therefore required in perspective of die cutting.

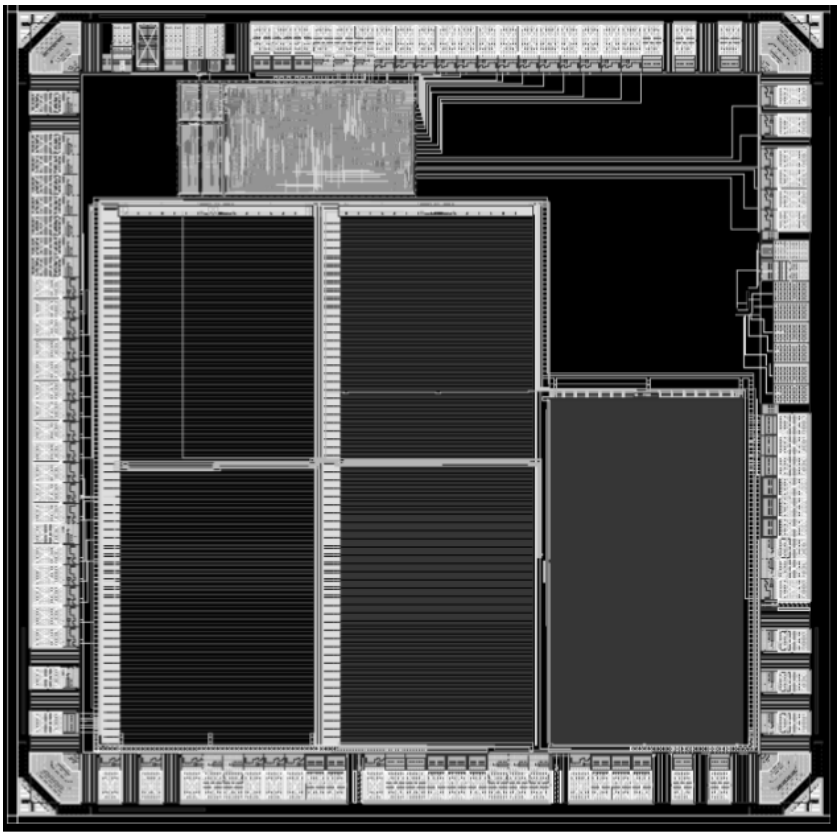


Figure 3.2: CHIPIX_SRAM, 1.96 mm × 1.96 mm. The chip contains two SEU-tolerant high-speed Ser/Des (top block) and three banks of radiation-hard SRAM DICE cells (bottom).

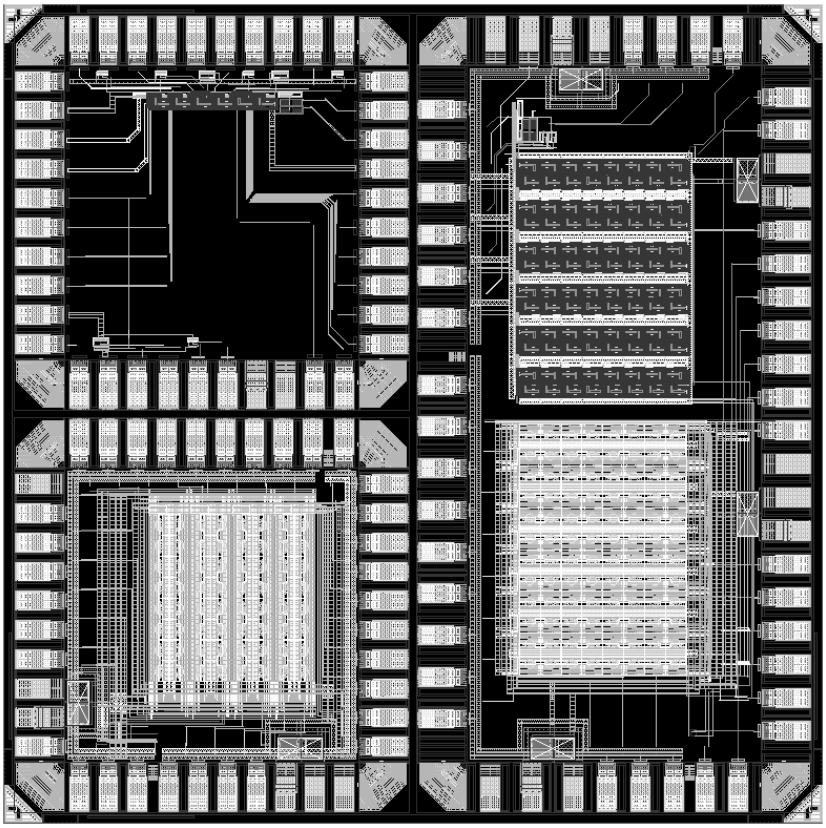


Figure 3.3: CHIPIX_VFE1, 1.96 mm × 1.96 mm. On the same silicon die have been included three independent chips implementing pixel Front-End electronics in 65 nm CMOS technology. CHIPIX_VFE1/PV (top-left) CHIPIX_VFE1/TO (bottom left) CHIPIX_VFE1/2x2. The chip integration has been carried out in a close collaboration among Pavia, Pisa and Torino INFN groups.

3.3 CHIPIX_VFE1 pixel Front-End prototypes

The final layout of CHIPIX_VFE1 is shown in Figure 3.3. The die integrates on the same silicon area of 1.96 mm × 1.96 mm three totally independent chips, each one containing different pixel Front-End electronics prototypes. All design activities were carried out in a close collaboration among Pavia, Pisa and Torino INFN groups.
A first chip, referred to as CHIPIX_VFE1/PV in the following and placed at the top-left, has been dedicated to single-channel analogue Front-End test structures designed by Pavia INFN [Gaioni 2014, Ratti 2014]. The implemented Front-End chain is a continuous-time solution consisting of a charge sensitive amplifier with Krummenacher feedback and a compact current-based comparator for the hit discrimination. A 4-bit D/A converter is included for the threshold adjustment.
At the bottom-left is placed a first small pixel matrix designed by Torino INFN . The chip, referred to as CHIPIX_VFE1/TO, contains an array of 8 × 8 cells with 50 μm × 50 μm pixel size. Each cell implements the synchronous Front-End architecture and necessary hit control logic already discussed in Chapter 2. As described in more details later in the chapter, all pixels of the matrix can be electrically stimulated and both the CSA analogue output and a selectable digital information provided by the synchronous discriminator (hit pulse or self-generated fast clock) are accessible for all pixels by using output lines shared among columns and a simple serial pixel addressing.
The largest chip placed on the half-right, referred to as CHIPIX_VFE1/2x1, incorporates instead two independent small pixel arrays of 8 × 12 pixels each one. These pixel matrices have been equipped with additional on-pixel digital circuitry. An 8-bit time-over-threshold count is in fact registered in each pixel with a binary counter. Moreover, digital serial readout and configuration are performed by means of shift-registers. Such a digital part, designed by Pisa INFN, is essentially common to both pixel arrays. The top matrix, assembled by Pavia INFN, uses the continuous-time Front-End solution implemented in single test structures and provides TOT measurements using an external clock. The bottom matrix, assembled by Torino INFN, uses instead the synchronous solution with the addition of digital readout and configuration.
Personal contributions in the chip integration and sign-off are described in the following sections.

3.4 Floorplan and pad frames assembling

In a first prototyping phase the number of testable circuit optimizations must be maximized. A preliminary top-level design partitioning and floorplan has been therefore essential in determining the maximum number of I/O pads available for each sub system.
According to the adopted prototyping option, the overall silicon area is 1960 μm × 1960 μm. However, any designs must fit within a dedicated boundary structure provided by the foundry access service. A seal-ring is in fact necessary to ensures mechanical-stress reliability. It is used to prevent cutter machines to break the die during silicon cutting (dicing), as well as to avoid moisture penetration [Chen 2005]. In the chosen technology, such a ring has a width of 20 μm. Since the effective available silicon area is reduced of 40 μm in each direction, a design cannot be larger than 1920 μm × 1920 μm. Indeed, an additional 5 μm space has been left from the internal edge of the seal-ring in order to further increase reliability, whereas a minimum 20 μm spacing between adjacent pad frames has been considered [Bonacini 2014]. As a result, an effective area of 1910 μm × 1910 μm was assumed for the initial partitioning of the silicon die in three independent regions. Note that no packaging has been required for CHIPIX_VFE1 chips.

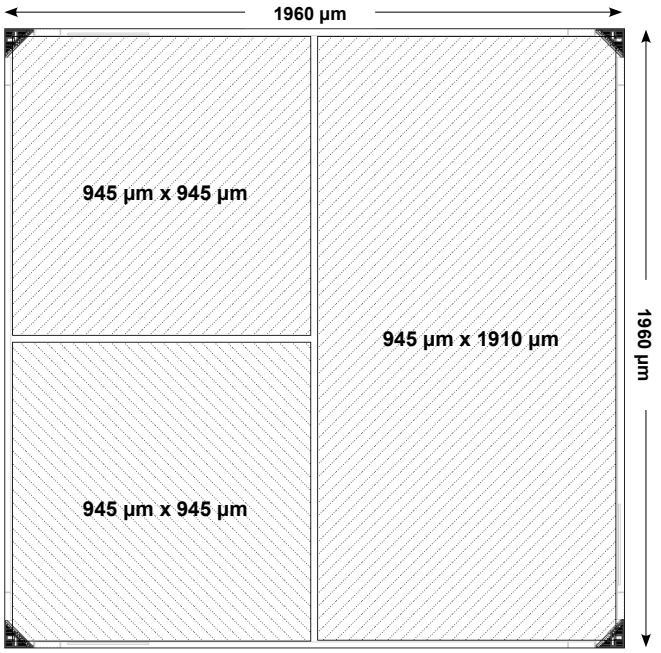


Figure 3.4: CHIPIX_VFE1 floorplan. The overall silicon area is 1960 μm × 1960 μm. The usage of a 20 μm width seal-ring reduces the effective available area to 1920 μm × 1920 μm. An extra 5 μm gap has been left between the internal edge of the seal-ring and each design. Furthermore, 20 μm spacing is introduced between adjacent designs.

A top-level design floorplan is depicted in Figure 3.4. On the left side, two square regions of 945 μm × 945 μm each one host CHIPIX_VFE1/PV and CHIPIX_VFE1/TO chips, whereas on the half-right a rectangular area of 945 μm × 1910 μm has been devoted to CHIPIX_VFE1/2x1. Starting from these specifications, a realistic estimate for the maximum number of available I/O pads per chip has been derived taking into account the dimensions of I/O cells and in respect of wire-bonding constraints.
Given the limited amount of digital circuitry included in first pixel Front-End prototypes, only simple bidirectional and unbuffered analogue I/O cells have been used to assemble the pad frames of all three chips, without the need of more complex programmable digital I/O cells [Kang 2003, Weste 2011]. As depicted in Figure 3.5, such circuits only contain basic protecting diodes against electrostatic discharge (ESD) damages. They are usually implemented with a couple of large complementary diode-connected MOS transistors. I/O power and ground rails for ESD devices are then provided by dedicated I/O cells. In order to guarantee adequate drive strength, proper buffering must be taken into account when I/O cells are connected to core internal signals.
The usage of differential digital output drivers would have been instead preferable, in particular in perspective of a full-analogue characterization of the high-frequency self-generated clock for fast TOT measurements. Unfortunately, no Low-Voltage Differential Signalling (LVDS) output pads nor Scalable counterparts (SLVDS) were available for the first submission in 65 nm. All digital input and output signals are therefore CMOS standard.

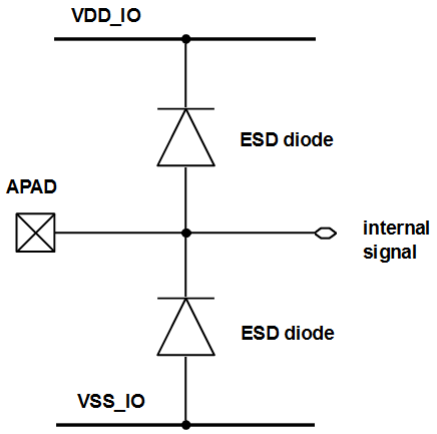


Figure 3.5: Generic analogue I/O circuit. The cell only contains protecting diodes against ESD surges, without additional output drive strength capabilities. Dedicated I/O power and ground rails are required for ESD devices.

Cells that comes with the 65 nm analogue I/O library are 120 μm hight and 50 μm width. The minimum spacing between two adjacent I/O pads has been discussed in close collaboration with the staff of the Torino INFN wire-bonding laboratory [Pini 2014]. According to bonding machine capabilities, the minimum pitch between adjacent wires has been fixed to 70 μm, thus resulting into a 20 μm minimum space constraint within two consecutive pads. Taking into account the size of corner cells, a maximum number of 40 pads has been identified for each 945 μm $\times$ 945 μm design, 10 pads per side. Certainly the pad count more than doubles for the 945 μm $\times$ 1910 μm chip. Indeed, due to an initial unavailability of schematic and full layout views of standard cells and I/O cells in 65 nm technology, the assembling of basic I/O rings has been more challenging. Without enough information to properly work in a full-custom design environment, a preliminary pad frame skeleton has been generated with the aid of a fully automated place-and-route tool, mainly to avoid wrong cell orientation issues due to black-box cells appearance. A top-level Verilog[2] HDL wrapper has been used to describe the pad ring in form of a gate-level netlist, instantiating necessary corner cells, power/ground cells and multiple general-purpose I/O cells. As shown in Figure 3.6, full-chip floorplan and script-based cell placement were then performed in respect of the above minimum pitch constraints. Moreover, filler cells have been automatically added in order to ensure electrical continuity for I/O supply and ground rails. At the end, the ring has been exported back into the full-custom design framework, providing a reliable skeleton for all subsequent customizations and modifications. With such an approach in fact, proper cell orientation is guaranteed if an automated cell placement is adopted. Wire-bonding pads have been manually placed on the top of each cell instead. The final I/O pads floorplan with cell abstract views is shown in Figure 3.7. The effective available core area inside each pad frame is essentially determined by the size of pads, resulting therefore into three pad-limited designs. Power partitioning and distribution for analogue and digital domains as well as secondary ESD protection placement will be addressed later in the chapter.

[2] IEEE std. 1364-2001, *IEEE Standard Verilog Language Reference Manual*, 2001

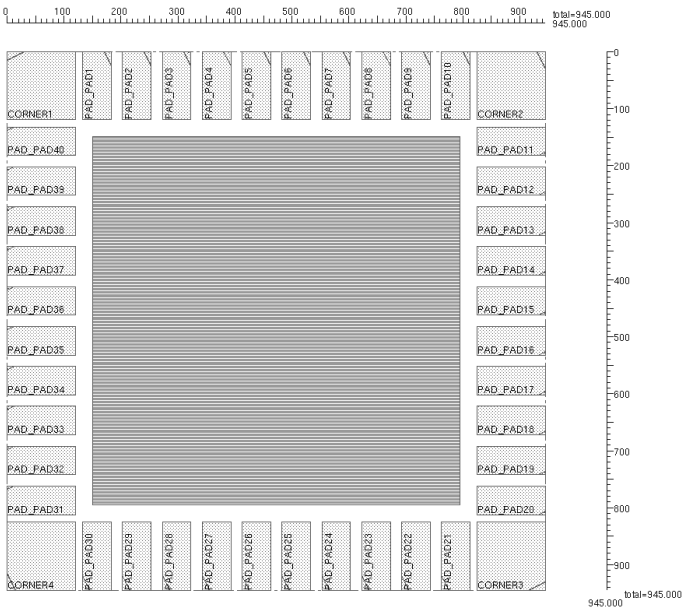


Figure 3.6: Preliminary pad frame skeleton for a 945 μm $\times$ 945 μm chip, generated from a Verilog HDL netlist with the aid of an automated digital place-and-route tool. The effective maximum core area of 705 μm $\times$ 705 μm available inside the ring is determined by the size of the I/O cells (pad-limited design). Filler cells are not shown.

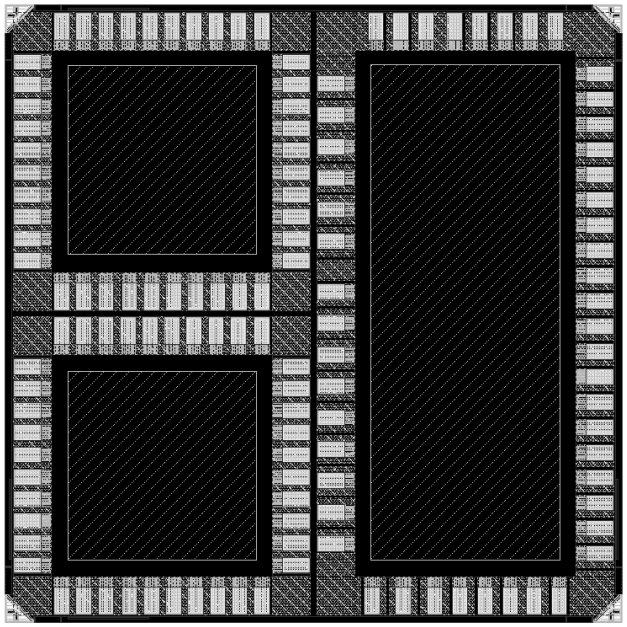


Figure 3.7: Final CHIPIX_VFE1 floorplan. Abstract views.

3.5 CHIPIX_VFE1/TO assembling

CHIPIX_VFE1/TO is a first small pixel matrix implementing the proposed synchronous Front-End architecture and necessary hit control logic to support latch operations for fast TOT measurements. The array contains 64 cells arranged into 8 rows × 8 columns. Pixels have been logically partitioned into 16 groups of 4 cells each one. Thanks to the usage of a simple shift-register based pixel addressing, both the CSA analogue waveform and a selectable discriminator digital waveform (the hit pulse or the self-generated fast clock) are accessible at the oscilloscope for all 16 pixel groups, 4 cells at a time. The chip has been assembled by following a traditional Analogue-on-Top (AoT) approach in the full-custom design environment, with constant support of schematic information. Only the preliminary skeleton of the pad frame has been automatically generated into the digital design environment with the aid of a place-and-route tool, as already discussed. In the following, details about the assembling of the pixel matrix and other aspects of the chip integration such as pad assignment, clock and power distribution are given. Necessary schematic level considerations are discussed before the actual chip layout implementation.

Pixel addressing and analogue readout

The CHIPIX_VFE1/TO chip implements a full-analogue readout of all pixel cells in the matrix, targeting to evaluate synchronous Front-End performance with simple waveform characterizations at the oscilloscope and maximum statistics. Additional circuitry and some fundamental choices adopted in the assembling of the chip are therefore specific to achieve such a functionality.
According to a maximum number of 10 pads foreseen for each side of the chip, a preliminary pad assignment attempt demonstrated that no more than 8 pads would have been available for direct analogue probing, corresponding to 4 pixels at a time, two output waveforms per pixel[3] (the CSA output and one discriminator output). Most of the pads are devoted in fact to power/ground inputs, bias currents, DC voltages, digital control signals and configuration bits. In particular, CHIPIX_VFE1/TO does not include an internal digital serial configuration, hence all selection bits for the pixel Front-End require dedicated pads as well. Indeed, the analogue read out of 4 channels at a time naturally fits with the choice of a square pixel region of 2 × 2 pixels as fundamental replica unit-cell for the layout assembling of the overall pixel matrix, as described shortly later. Since I/O cells are 120 μm height, the maximum core area available inside a pad ring of 945 μm × 945 μm is 705 μm × 705 μm. Thereby, with a pixel size of 50 μm × 50 μm, an array of 8 × 8 pixels occupies 400 μm × 400 μm, leaving enough space at the matrix periphery for all required bias cells and input/output buffers.

As depicted in Figure 3.8 the 8 × 8 pixel matrix has been logically partitioned into 16 groups of 4 cells each one. Following the traditional approach adopted in pixel ASICs, two adjacent pixel columns form a double-column of 8 × 2 pixels in order to share analogue/digital output lines, bias and digital control signals. For each double-column, 4 metal stripes along the column are used to transmit the CSA analogue waveforms of 4 pixels towards the matrix periphery and then to output pads $ANAOUT0$-$ANAOUT3$ after proper buffering. Similarly, 4 lines per double-column are used to transmit a selectable discriminator digital output to 4 output pads $DIGOUT0$-$DIGOUT3$.

[3] As already mentioned, all pad frames assembled for CHIPIX_VFE1 prototypes only contain *analogue* I/O cells. The distinction between analogue and digital inputs/outputs only refers to different signal types and separate power domains, but no configurable *digital* I/O cells have been used, nor LVDS/SLVDS differential cells.

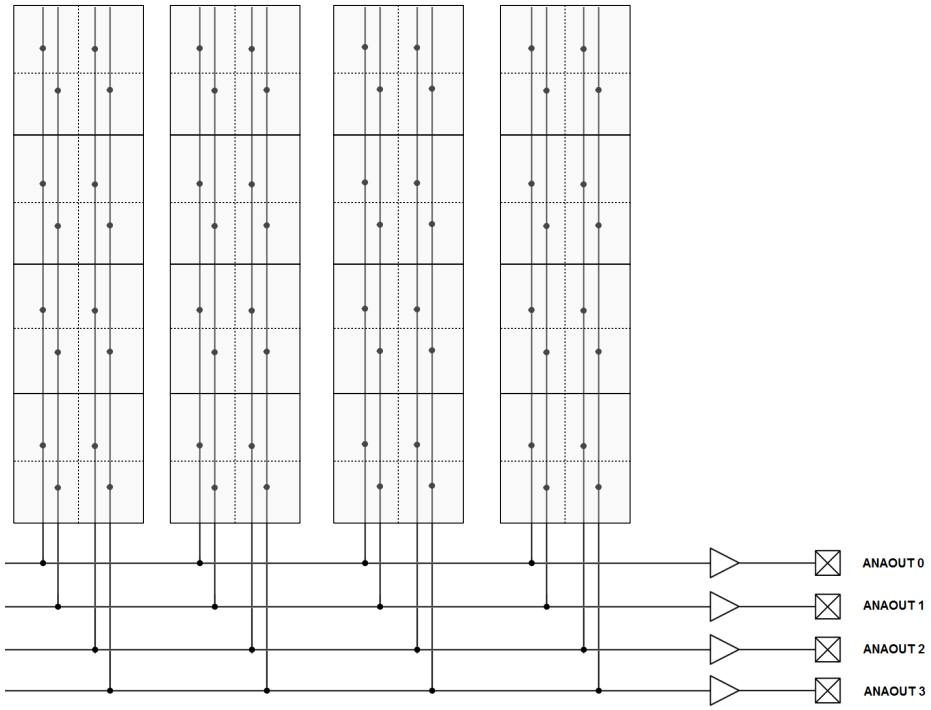


Figure 3.8: CHIPIX_VFE1/TO full-analogue readout scheme. The 8 × 8 matrix is partitioned into 16 groups of 4 pixels each one. Adjacent pixel columns form double-columns of 8 × 2 pixels. For each double-column, 4 vertical metal stripes transmit the CSA output waveform of a group of 4 pixels towards the periphery and then to *ANAOUT*0-*ANAOUT*3 pads. Necessary buffering is performed at the chip periphery before pads. The same scheme (not shown in figure) is adopted to transmit 4 discriminator outputs to *DIGOUT*0-*DIGOUT*3 pads.

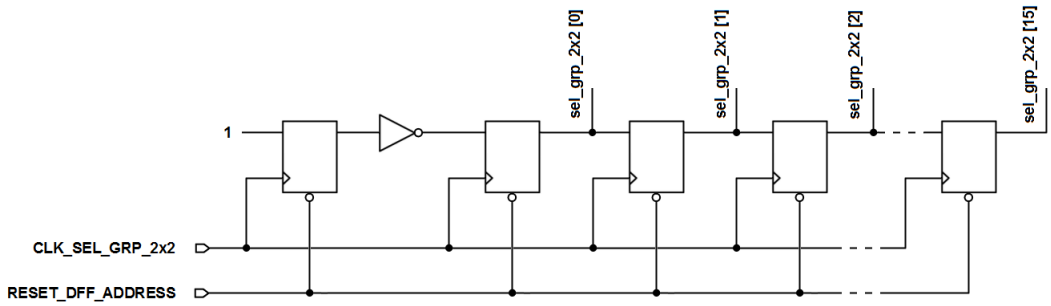


Figure 3.9: Addressing scheme to select among pixel regions of 2 × 2 cells. Control signals *CLK_SEL_GRP_*2x2 and *RESET_DFF_ADDRESS* are provided off-chip using dedicated input pads. As described in the next chapter, these signals are generated with simple push-buttons on the test board. Logic one is fixed with a tie-high cell.

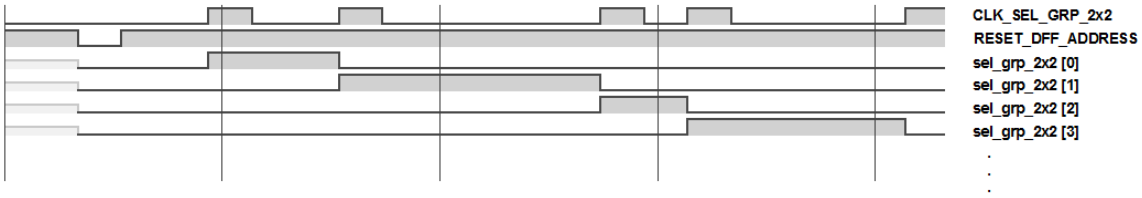


Figure 3.10: Behavioural simulation for pixel regions selection (arbitrary time scale). External control signals *CLK_SEL_GRP*2x2 and *RESET_DFF_TOT* are supposed to be generated with push-buttons on a test board.

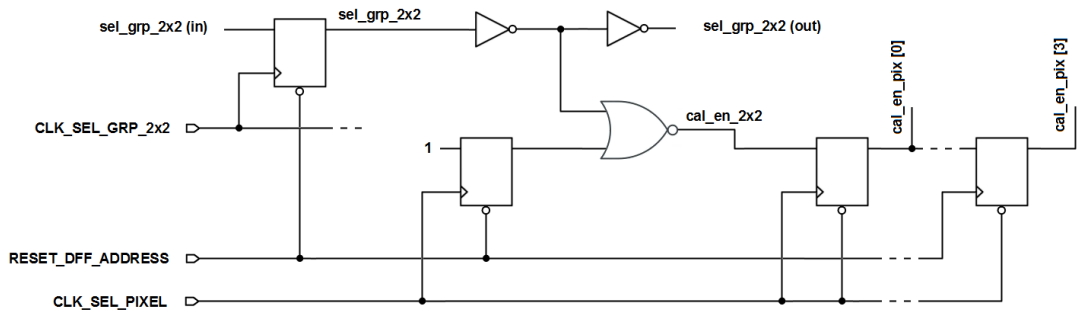


Figure 3.11: Addressing logic implemented in each 2 × 2 pixel region. The additional 4-bit shift register with one-hot outputs is used to switch the test charge injection in one pixel at a time among 4 selected cells.

The multiplexing of 16 pixel regions to 8 output pads is accomplished with a pixel addressing scheme based on shift registers with one-hot encoding. As shown in Figure 3.9, a first 16-bit shift register with one-hot outputs (one FlipFlop in each pixel group) is used to propagate a *sel_grp_*2x2 flag which selects a group of 4 cells inside the matrix. An example timing diagram is depicted in Figure 3.10. External *CLK_SEL_GRP_*2x2 and *RESET_DFF_ADDRESS* control signals are required in order to select the desired pixel region and properly reset FlipFlops respectively. Both digital signals must be provided off-chip using dedicated input pads. As described in the next chapter, these signals have been generated with simple push-buttons on the test board.
The *sel_grp_*2x2 logic level available in each pixel region is then used for two purposes. On the one hand, it activates a second 4-bit shift register with a similar one-hot encoding placed in each pixel region and used to address the test charge injection. As shown in Figure 3.11, when a 2×2 pixel group is selected (hence *sel_grp_*2x2 is high for that pixel region) the test charge injection circuit can be enabled in one pixel at a time by means of an auxiliary *CLK_SEL_PIXEL* signal that shifts a *cal_en_pix* flag among the 4 selected pixels. An additional input pad is therefore required for this purpose. In order to minimize the number of pads used for the overall pixel addressing, FlipFlops are initialized using the same global *RESET_DFF_ADDRESS*. On the other hand, *sel_grp_*2x2 is used to drive a few additional switches, buffers and logic gates such that both the CSA analogue output *ANAOUT* and a discriminator digital output *DIGOUT* of all selected 4 cells can be transmitted towards the matrix periphery. As shown in schematic of Figure 3.12, in the analogue part *sel_grp_*2x2 enables CMOS switches and a sufficiently large source follower used to provide necessary buffering to drive the parasitic capacitance ($\approx$ 150 fF from extracted layout parasitics) of the entire column output line. A fixed bias current of 6 μA flows in the source follower and is generated by means of a local current mirror placed in each 2×2 pixel region[4]. Referring instead to schematic of Figure 3.13, *sel_grp_*2x2 is fed to the on-pixel digital part and implements the *MASK* bit functionality. That is, in CHIPIX_VFE1/TO pixels a 40 MHz master clock *CLK*40 is distributed to all cells, but the discriminator is activated only if the pixel region is selected. A 2:1 multiplexor is used to identify the digital output of interest for the discriminator (hit pulse or self-generated fast clock) according to the value of the *TOT_EN* configuration bit. In particular, if *TOT_EN* is set to low the latch never turns into a local oscillator. The selected output signal is therefore the hit waveform, with a pulse duration which is an integer number of *CLK*40 clock cycles. The self-generated clock waveform is available when *TOT_EN* is set to high instead. A 4-stages tri-state digital buffer is used to drive the column output line up to pad buffers.

The additional circuitry placed in the analogue cell introduces extra noise contributions when activated. The actual Front-End performance in terms of ENC and minimum detectable charge are therefore modified by the presence of the source follower and CMOS switches. Furthermore, due to layout parasitics and crosstalk effects, digital outputs transmitted towards the periphery can inject noise and corrupt the analogue waveforms. This is an important point in perspective of test measurements. Two global configuration bits *ANAOUT_EN* and *DIGOUT_EN* have been therefore added at the cost of two additional pads. These configuration bits allow to enable/disable the transmission of Front-End outputs towards the pads without preventing normal synchronous operations. With such a solution, analogue and digital output waveforms are totally independent. The effective discriminator threshold can be extracted from the binary-only information by means of threshold scans and S-curves [Spieler 2005, Rossi 2006] while keeping off the analogue buffer. Furthermore, the analogue waveform can be precisely characterized by completely turning off the whole digital part (without providing the 40 MHz master clock to the chip) or during normal synchronous operations but without crosstalk effects from digital outputs.

[4]The DC current required for the buffer certainly increases the static power consumption well above the maximum power budget of 6 μW/pixel defined for the analogue Front-End. Nevertheless, this is only a prototyping feature, hence it does not contribute to the effective power compute for the pixel Front-End.

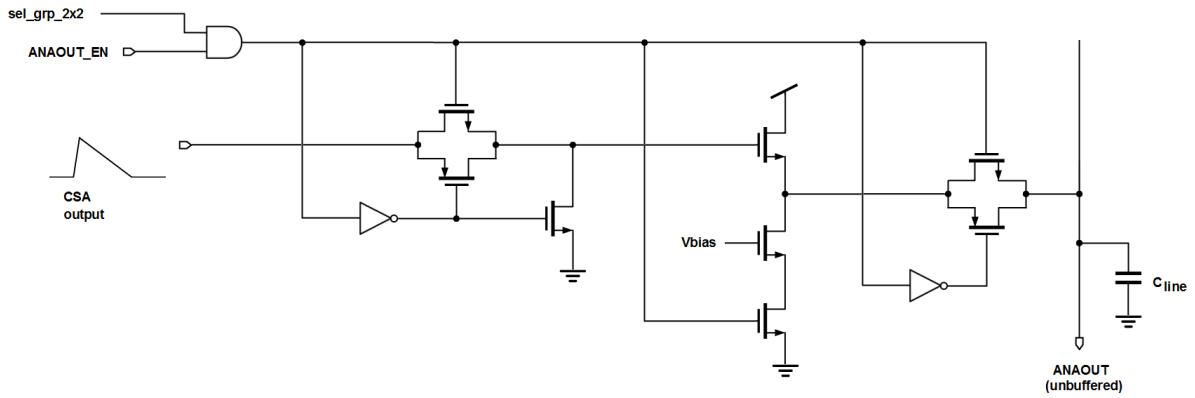


Figure 3.12: Additional circuitry in the analogue part to implement a full-analogue readout of the CSA waveform. Necessary buffering to drive the parasitic capacitance of the entire column line is provided by a source follower. The buffer is biased with a fixed current of 6 μA generated by a local current mirror placed in each 2 $\times$ 2 pixel region.

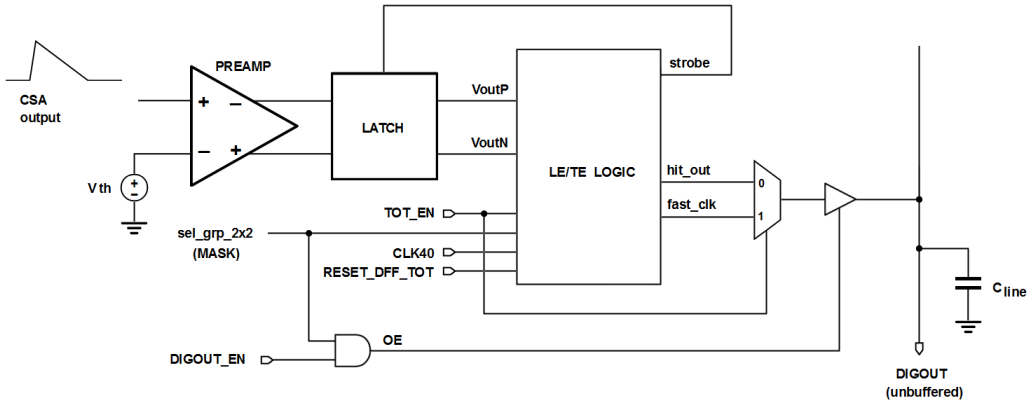


Figure 3.13: Additional logic in the digital part to implement a full-analogue readout of one selectable discriminator waveform. The output can be either the hit pulse or the self-generated clock depending on the value set for the *TOT_EN* configuration bit. The latch control logic has been described in the previous chapter.

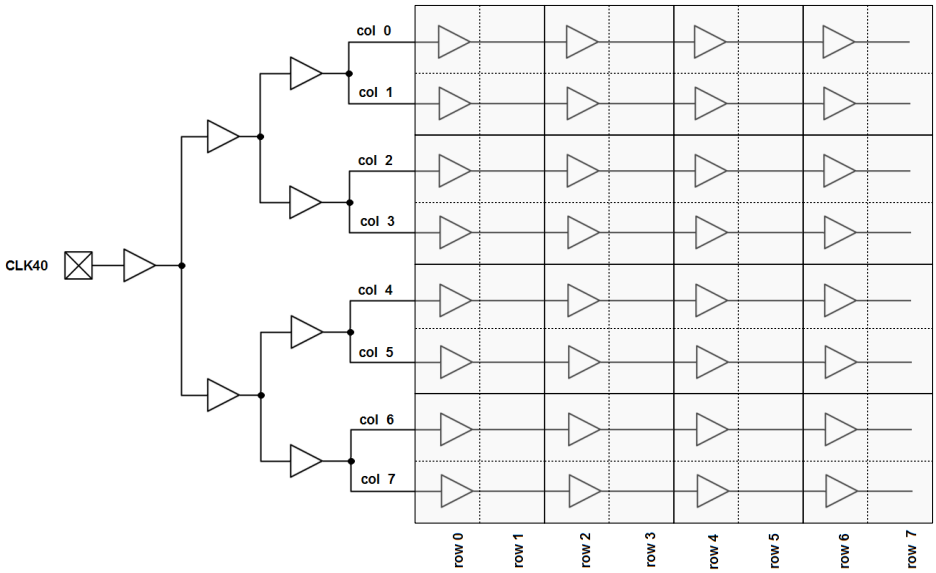


Figure 3.14: Clock distribution network.

Clock and digital control signals distribution

A nominal 40 MHz master clock *CLK*40 must be available in each pixel cell for synchronous latch operations. A schematic view of the adopted clock distribution network is presented in Figure 3.14. As usually performed in pixel ASICs in which a clock signal is propagated to all pixels, *CLK*40 was distributed along pixel columns from pixel to pixel, exploiting the delay introduced by local buffers placed every two consecutive cells. Transient simulations for the clock skew are presented in Figure 3.15. This simplifies the topology of the clock distribution tree, without the need of automated routing engines and extensive clock buffering. This also reduces the power dissipation. Most important, a non-zero skew avoids a simultaneous switching of all pixels in the matrix, which in turn would affect the stability of power supply rails. As mentioned, all digital signals are CMOS single-ended, remanding to a second iteration the usage of differential implementations. The same network has been replicated in order to distribute to all pixels a *TESTP* signal, required to trigger the test charge injection circuit included in each pixel cell, and a *RESET_DFF_TOT* signal, necessary for the on-pixel control logic.
All ϕ-signals for the offset compensation in the synchronous comparator are generated instead from a unique *PHI_CAL* external signal, which must be provided off-chip and is buffered towards the 4 double-columns. A simple circuitry placed at the beginning of each double-column is then used to generate ϕ_{1A}, ϕ_{1B} and ϕ_2 signals by means of asynchronous delays and coincidences. Transient simulations for the resulting control signals are presented in Figure 3.16. With the same principle, these signals are then propagated with non-zero skew along pixel columns from pixel to pixel. No extra buffering has been introduced instead for *CLK_SEL_GRP*2x2, *CLK_SEL_PIXEL* and *RESET_DFF_ADDRESS* signals used to address the pixel group to be read out and the pixel in which enable the test charge. As anticipated, all these signals have been generated with simple push-buttons on a test board, hence their very low frequency and low duty cycle does not require a dedicated distribution network.

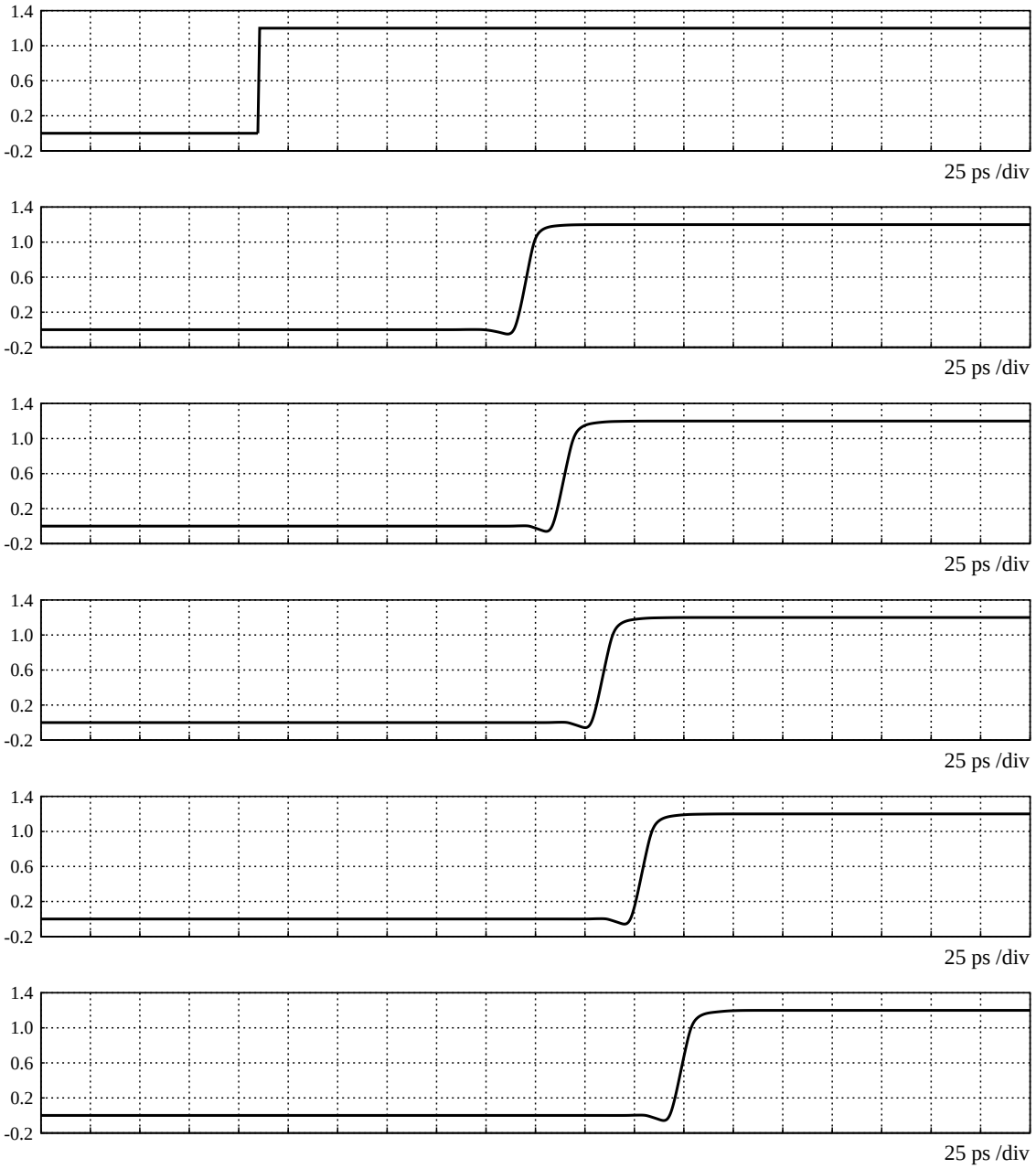


Figure 3.15: Simulated clock skew along a pixel column. From top to bottom: external $CLK40$ clock source (input pad), begin of column and pixel groups (one buffer per 2 pixels). The clock delay between two consecutive pixel regions is about 25 ps.

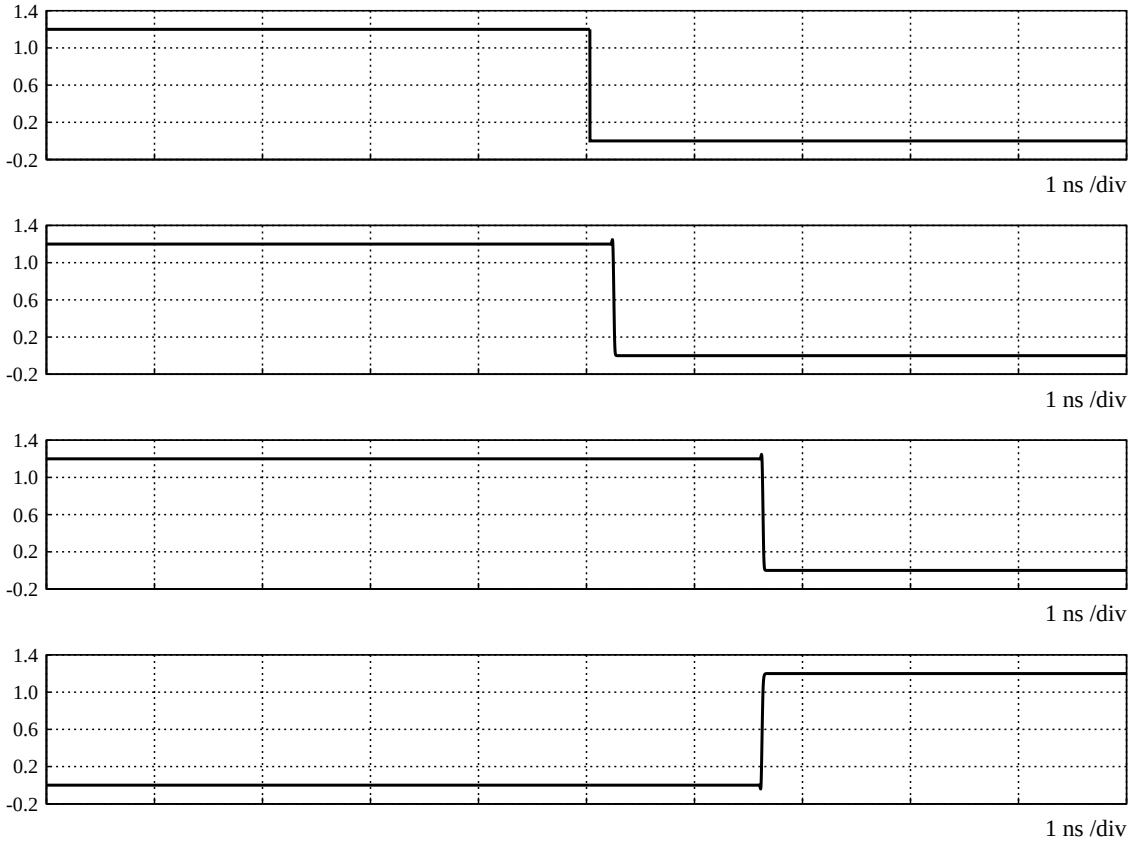


Figure 3.16: Timing for ϕ-signals. From top to bottom: *PHI_CAL* at the end of column (buffered), ϕ_{1B}, ϕ_{1A} and ϕ_2.

Bias generation and power distribution

The analogue Front-End requires specific bias currents to establish proper DC operating conditions for MOS transistors. In perspective of the first prototyping phase, all these currents have been generated with basic current mirrors placed at the periphery of the core 8 × 8 pixel matrix using diode-connected devices. Steady voltages in cascode current mirrors have been fixed instead by means of simple MOS voltage dividers. More realistic solutions using D/A converters and bandgap references will be introduced in a second iteration instead.
As already discussed in Chapter2, the core charge sensitive amplifier requires two independent bias currents $IbiasP1$ and $IbiasP2$ with nominal values 500 nA and 1.5 μA respectively, whereas a nominal 500 nA $IbiasSF$ flows in the NMOS output source follower. The feedback network implements a NMOS Krummenacher scheme. A total current I_{Krum} in the 1-100 nA range is required. The low-gain differential preamplifier of the synchronous discriminator uses a nominal tail current of 1 μA and two auxiliary lateral shunt current sources on 400 nA each one to increase the gain. A unique bias cell with current mirrors is used to properly obtain these current values starting from a single current $IbiasDisc$ of about 400 nA. The current flowing in the additional source follower included in each pixel cell for a full-analogue readout of the CSA output have been fixed instead to 6 μA. However, further buffering is performed at the chip periphery in order to drive the load capacitance of a test probe (up to pF) connected to $ANAOUT$ output pads. Large source followers have been therefore placed at the chip periphery. These buffers are biased with a selectable current $IbiasBuff$ in the 100-200 μA range. Finally, the frequency of the self-generated clock signal for fast TOT measurements is determined by a bias current $IctrlTOT$ in the 1-20 μA range, which is used to control the delay line inserted in the asynchronous logic feedback loop.
Each mentioned bias current requires a dedicated input pad connected to a bias generation circuit. In the preliminary prototyping phase, simple voltage dividers placed on a test board can be used to establish a voltage reference to which a fixed bias resistor $Rbias$ is connected. As shown in Figure 3.17, this accomplished with commercial external discrete components placed on the custom test board designed to host the chip. Test points are necessary in order to measure the voltage drop across the bias resistor and compute the current flowing into the pad. The actual values chosen for trimmers and resistors depend on the bias current. An important point is that for a given current, a single diode-connected bias cell serves a double-column of 8 × 2 pixels. Thus, 4 different bias cells placed at the periphery are actually connected in parallel to the same pad, and the current provided by the external circuit must be 4 times larger than the actual nominal value.
The proposed analogue Front-End also requires 4 selectable DC voltages: the reference voltage for the Krummenacher feedback, the DC level for the test charge injection circuit, the global threshold for the discriminator and a common-mode voltage to accomplish the offset compensation. As a result, 4 additional pads *VREF_KRUM*, *CAL_LEVEL*, *VTH_DISC* and *VREF_DISC* respectively have been assigned for this purpose. All necessary bias currents and DC voltages with relative pad names are summarized in Table 3.2. Figures 3.18-3.24 show instead for each bias current the relationship between the current flowing into the chip and the test point voltage at the internal terminal of the bias resistor. More details about the test board are given in Chapter 4.

Analogue power/ground rails $VDDA/GNDA$ and digital counterparts $VDDD/GNDD$ are totally independent networks. They have been distributed along pixel columns as vertical metal stripes and using the less-resistive top-metal offered by the chosen 65 nm fabrication technology. The most notable aspect of the power distribution strategy resides in the choice of not separating the analogue substrate from the analogue ground *GNDA*, without the usage of an independent substrate rail. This choice has been primary motivated by the limited number of pads available to implement all required configuration bits, bias and DC voltages, analogue/digital outputs and digital control signals.

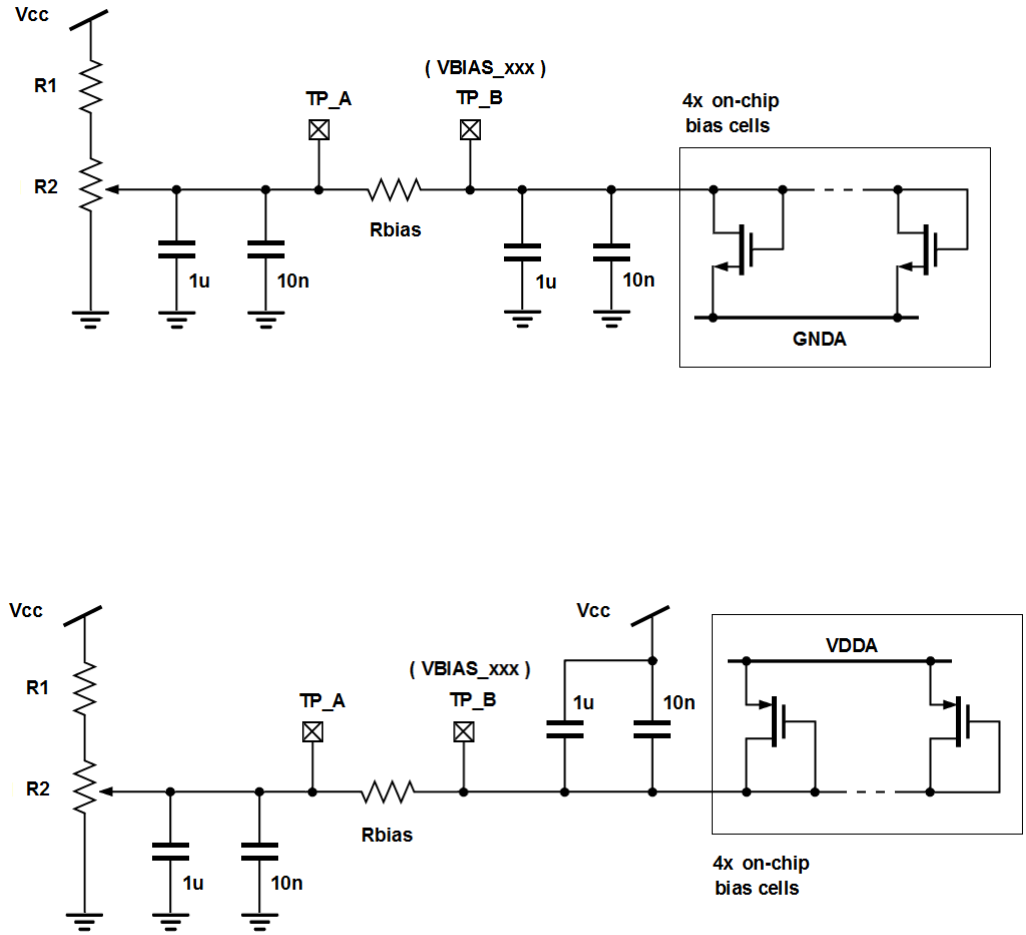


Figure 3.17: All necessary bias currents are generated with diode-connected devices at the chip periphery. A simple voltage divider on the test board is then used to drive a bias resistor connected to the chip. Actual values for trimmers and resistors depend on the required bias current.

Pad name/test point	Specifications
*VBIASP*1	core amplifier $IbiasP1$, 4x 500 nA
*VBIASP*2	core amplifier $IbiasP2$, 4x 1.5 μA
VBIAS_SF	core amplifier $IbiasSF$, 4x 500 nA
VBIAS_FEED	total feedback current, 4x 1-100 nA
VBIAS_BUFF	pad analogue buffers, 4x 100-200 μA
VBIAS_DISC	discriminator bias current, 4x 380 nA
VCTRL_TOT	delay line $IctrlTOT$, 4x 1-20 μA
VREF_KRUM	Krummenacher feedback DC voltage, 490 mV nominal value
CAL_LEVEL	test charge injection circuit DC level, 10-600 mV
VTH_DISC	global threshold, 450-600 mV
VBL_DISC	discriminator common-mode voltage, 490 mV nominal value

Table 3.2: CHIPIX_VFE1/TO input pads dedicated to bias currents and steady DC voltages.

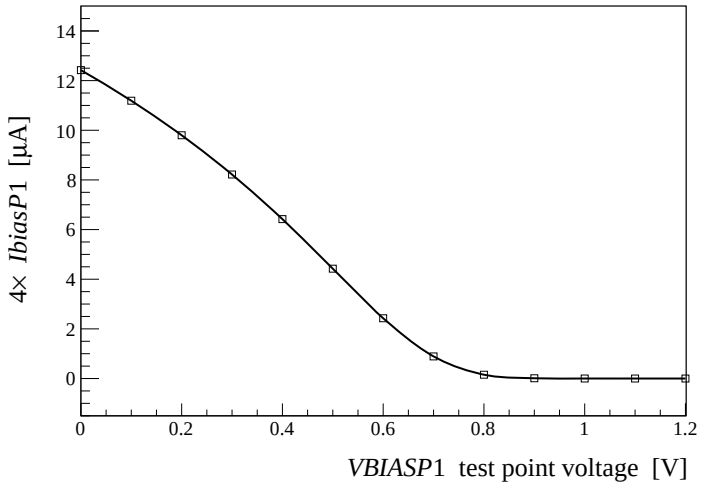


Figure 3.18: CSA total bias current versus *VBIASP1* test point voltage.

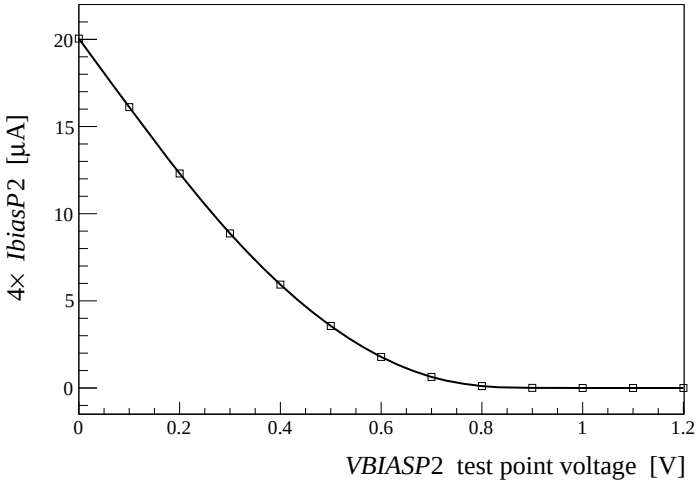


Figure 3.19: CSA total bias current versus *VBIASP2* test point voltage.

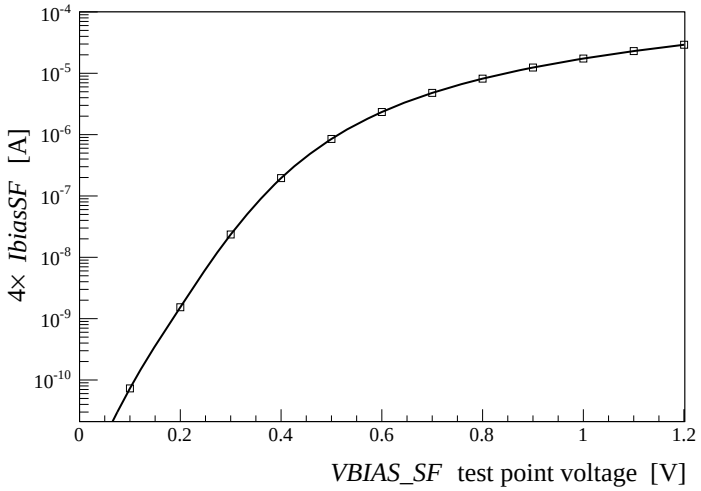


Figure 3.20: CSA output SF total bias current versus *VBIAS_SF* test point voltage.

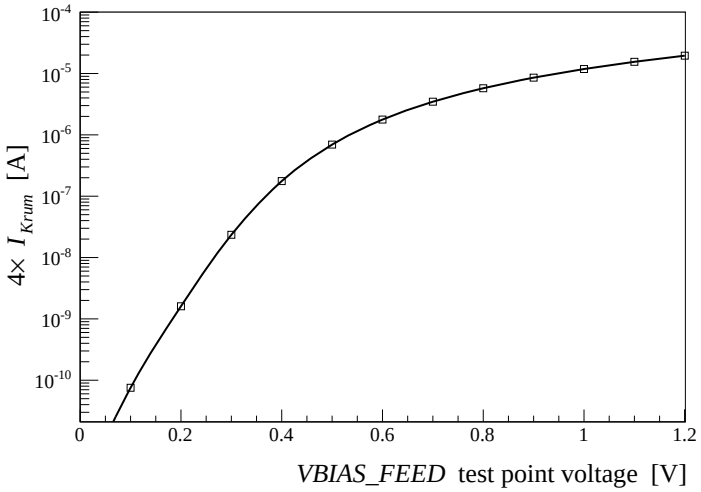


Figure 3.21: Krummenacher feedback total bias current versus *VBIAS_FEED* test point voltage.

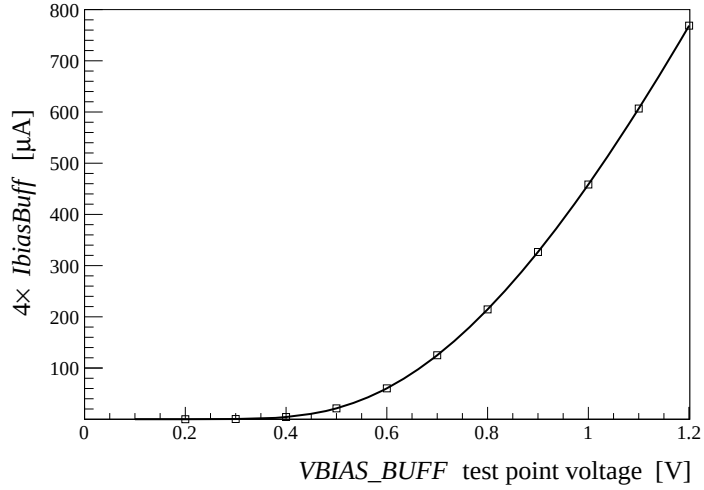


Figure 3.22: Output buffers total bias current versus *VBIAS_BUFF* test point voltage.

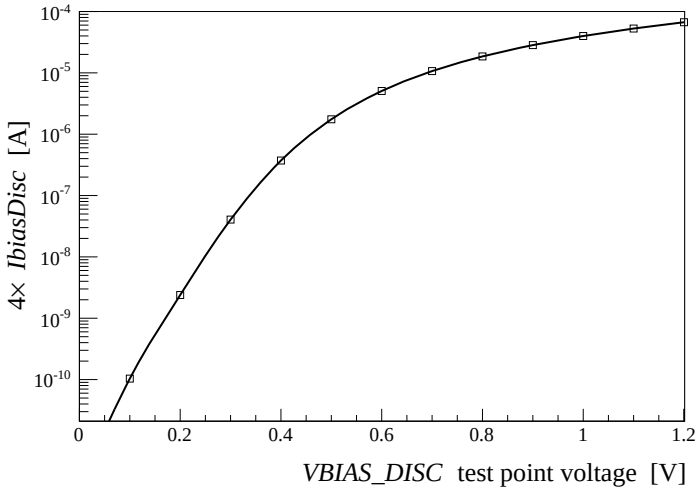


Figure 3.23: DISC total bias current versus *VBIAS_DISC* test point voltage.

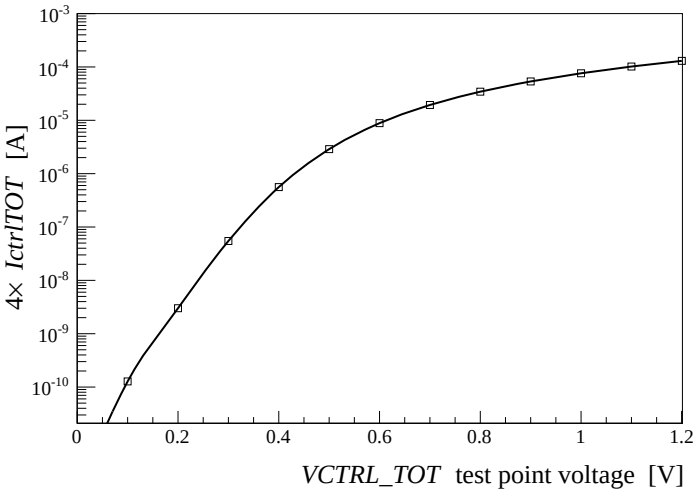


Figure 3.24: Total TOT control current versus *VCTRL_TOT* test point voltage.

Core pixel matrix layout assembling

In this section, the assembling of the overall layout of the 8 × 8 core pixel matrix is described. As mentioned, the chip integration has been accomplished using a traditional AoT approach, with the constant support of a hierarchical design entry at the schematic level and extensive DRC and LVS cross checks at each step. Thanks to the repetitive structure of a pixel array, the layout has been essentially obtained by replicating reference cells at different levels of the design hierarchy.
As a first step, the layout of the basic pixel cell has been assembled by interfacing the analogue part with the digital control logic. Furthermore, a bump-bonding pad connected to the input transistor of the Front-End chain has been included in each pixel cell. The layout is shown in Figure 3.25. The pad is an octagon of about 24 μm diameter (9.215 μm edge length) implemented with specific mask layers according to flip-chip design rules of the 65 nm CMOS process. A via-stack across the full metal scheme connects the pad to a lower routing metal layer and then to the CSA input node. The pad is placed at the geometrical centre of a 50 μm × 50 μm square pixel cell. The assembled matrix exhibits therefore a regular 50 μm pitch bump pattern. With such a choice, half pad lies in the pixel digital part. As a result, a careful shielding is required under the pad in order to avoid digital-induced noise to couple with the input node through parasitic capacitances. Due to the absence of a significant amount of on-pixel digital circuitry, a selectable 2 GHz, 17-stages ring oscillator was added in order to characterize the noise that might be injected in the analogue part by an intense digital switching activity. As depicted in Figure 3.26, the ring oscillator can be activated using an external configuration bit *SUBSTRATE_NOISE*, thus requiring a dedicated additional pad. Figure 3.27 shows the basic layout of a pixel unit cell. The digital part includes the latch control logic already discussed in Chapter 2 and necessary pixel addressing, multiplexing and output buffering for a full-analogue readout of the CSA output and of a discriminator information, as discussed. The ring oscillator is placed under the bump pad. As one can see, about half of the digital pixel cell is empty. Moreover, full-custom cells adopted to implement the digital circuitry are much larger than STD cells provided by 65 nm logic libraries.
Indeed, the fundamental replica unit-cell adopted for the assembling of pixel double-columns is represented by a 100 μm × 100 μm pixel region of 2 × 2 pixels. As shown in Figure 3.28, pixel cells have been mirrored with both upside-down and left-right symmetry. On the one hand, this choice leads to a better isolation of the Front-End amplifier from the digital part. On the other hand, a perfectly symmetric analogue layout reflects the mainstream floorplan adopted within the RD53 collaboration for a pixel region of 2 × 2 pixels in order to have *analogue islands* embedded into a fully-digital environment [Mekkaoui 2014]. A dedicated internal routing has been therefore performed in order to address the analogue readout and the test charge injection among mirrored pixels of a group of 2 × 2 cells. Empty STD cell rows have been filled instead with buffers for the master clock and for other digital control signals, with dummies and with custom decoupling capacitors. A top-level routing using vertical metal stripes was then added for necessary bias lines, digital control signals, global configuration bits and analogue/digital outputs. Finally, analogue and digital power/ground have been routed using the less-resistive top-metal offered by the chosen 65 nm fabrication technology, as mentioned.
The layout of a double-column of 8 × 2 pixels has been then assembled by replicating 4 pixel regions, as shown in Figure 3.29. All necessary bias cells and end-of-column buffers have been placed at the periphery. Eventually, the layout of a pixel matrix of 8 × 8 pixels has been obtained by replicating and abutting 4 pixel columns. The final layout is presented in Figure 3.30. At this level of the design hierarchy, only a custom routing among pixel double-columns was necessary in order to propagate the *CLK_ADDRESS_*2x2 selection clock.

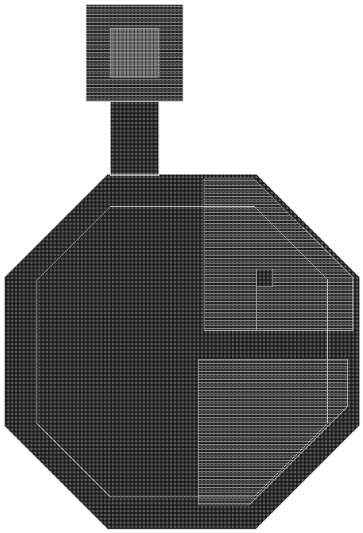


Figure 3.25: Bump bonding pad, 24.08 μm diameter. Such a full-custom cell has been designed according to flip-chip design rules of the 65 nm CMOS process.

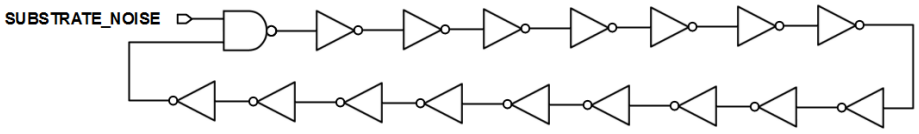


Figure 3.26: A selectable 2 GHz ring oscillator is included in each pixel cell in order to characterize the noise that might be injected in the analogue part by an intense digital switching activity.

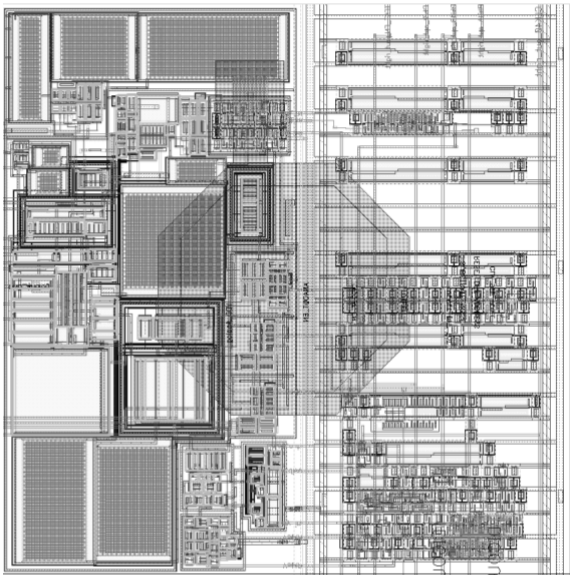


Figure 3.27: Pixel unit cell layout, 50 μm $\times$ 50 μm. The digital part includes necessary latch control logic, output buffering and pixel addressing logic. The bonding pad is placed at the centre of the pixel cell. A selectable 2 GHz ring oscillator is placed under the bump pad for noise characterization. Empty STD cell rows have been then filled with buffers, dummies and decoupling capacitors.

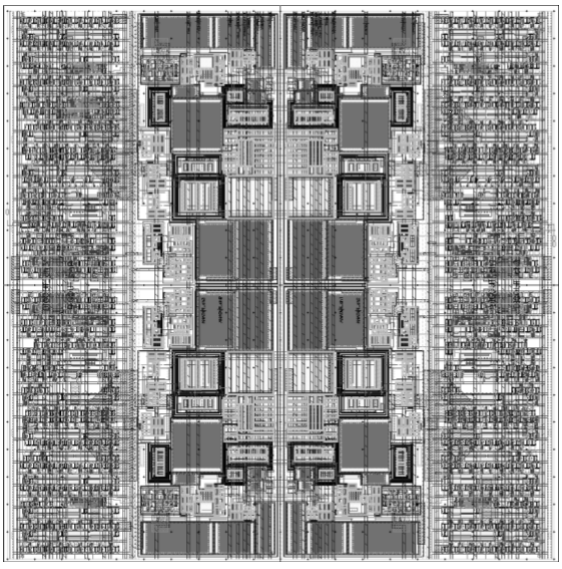


Figure 3.28: Complete layout of a 2 $\times$ 2 pixel region, 100 μm $\times$ 100 μm. The layout exhibits both upside-down and left-right symmetry and represents the fundamental replica unit-cell adopted in the assembling of 8 $\times$ 2 double-columns. Only lower routing metal layers are shown.

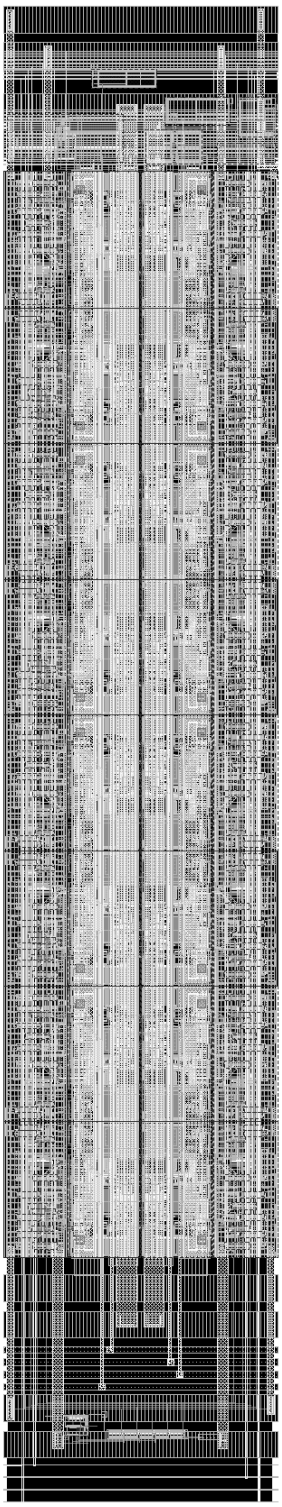


Figure 3.29: CHIPIX_VFE1/TO pixel double-column 8 × 2 layout.

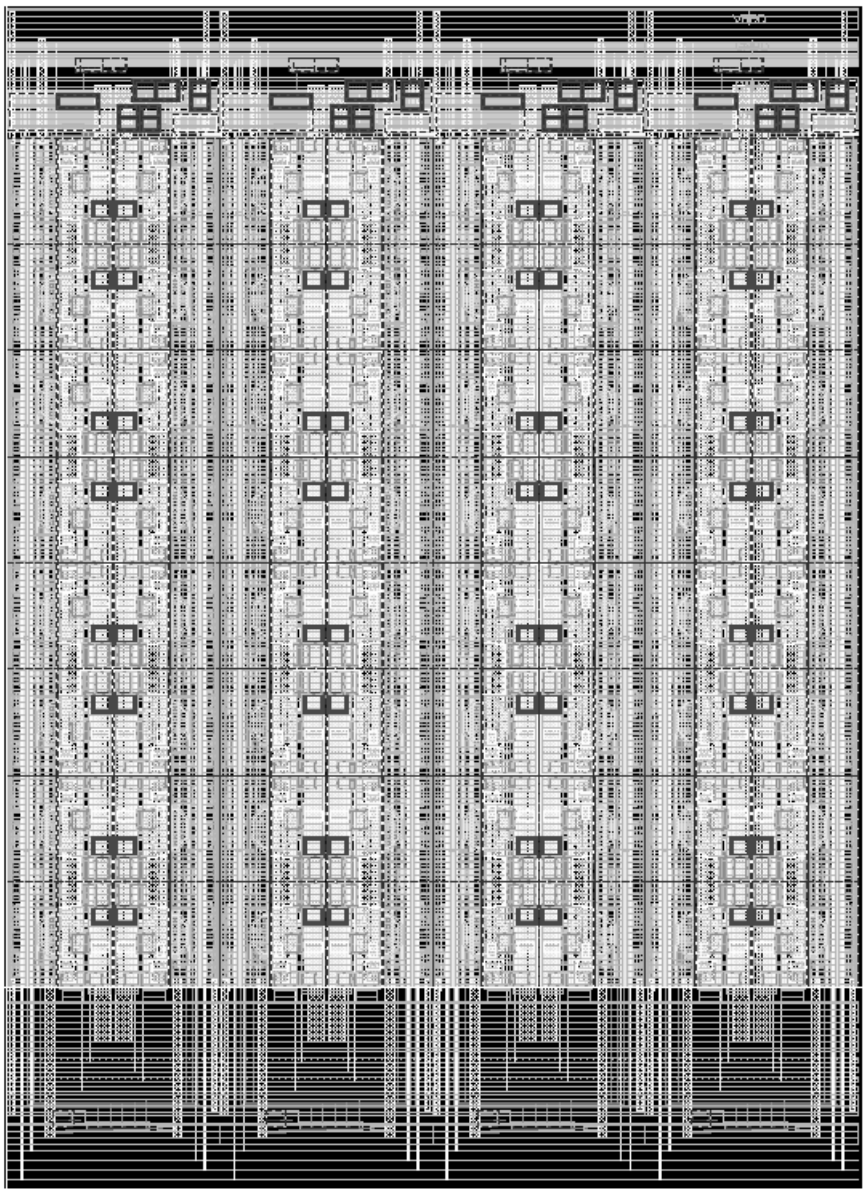


Figure 3.30: CHIPIX_VF1/TO core pixel matrix 8 × 8 layout.

I/O power partitioning and final pad frame assembling

As already discussed, without full layout views of logic cells and I/O cells initially available in 65 nm, a preliminary skeleton of the pad frame was generated with the aid of an automated place-and-route tool in order to prevent placement issues due to black-box cells appearance. The ring has been then exported from the digital environment to the full-custom design framework for necessary further customizations and integration with the core pixel matrix. The floorplan determined a maximum number of 40 pads available for the CHIPIX_VFE1/TO chip. The final pad assignment was essentially driven by power partitioning and wire-bonding considerations.
Analogue and digital power domains have been kept separated and dedicated power/ground cells from the 65 nm analogue I/O library have been instantiated in the pad frame. Indeed, different I/O power distribution schemes can be adopted. Each I/O cell includes in fact power/ground rails for internal ESD diodes. In general, the supply voltage required by I/O cells is totally independent from the core voltage. Most of commercial integrated circuits commonly employ different power domains for core transistors and I/O devices. As an example, I/O cells can be powered at 2.5 V/3.3 V while the internal core has a supply voltage of 1.2 V. Certainly, the usage of a higher I/O voltage increases the maximum current that can be absorbed by protecting devices in case of a static discharge occurs from the external world. Similar considerations apply also to ground rails, with the possibility of having totally independent grounds for core transistors and I/O cells or coupling ESD devices to the analogue ground for example. Despite usually not recommended, the internal ground of I/O cells can be also left floating. A reach variety of I/O power/ground cells is therefore available in modern CMOS fabrication technologies.

As a matter of fact, due to the absence of a digital serial configuration for pixels, a significant number of pads in CHIPIX_VFE1/TO is simply devoted to global configuration bits. The I/O power distribution has been therefore simplified, as summarized in Table 3.3. On the one hand, the pad frame has been partitioned into two semi-rings such that two independent supply voltages are used by ESD devices, corresponding to two separated analogue and digital domains. On the other hand, pads within these two I/O power domains have been connected to same 1.2 V core voltages of the internal matrix, renouncing to the usage of a higher bias for pads. Specific power cells which are internally connected to supply rails of pads ESD devices have been therefore chosen for *VDDA* and *VDDD*. Furthermore, two dedicated power-cut cells have been inserted such that I/O cells in the analogue semi-ring are powered at *VDDA*, whereas I/O cells in the digital counterpart are powered at *VDDD*. Core analogue and digital ground networks *GNDA* and *GNDD* have been instead decoupled by the I/O ground. Two dedicated cells provide in fact *GNDA* and *GNDD* without any internal connection to ground rails of ESD devices. An independent ground is fed instead to I/O cells with two specific ground cells *VSSA* and *VSSD*. Despite their different names, the pad frame has a unique internal ground rail, hence these two pads are actually shorted. In order to guarantee a uniform discharge current path for all I/O cells around the frame, these pads have been placed almost equidistant. Filler cells have been then added to ensure electrical continuity for I/O internal rails among adjacent pads. Finally, according to design-for-reliability (DFR) rules suggested by the manufacturer, two additional power clamp cells offered by the 65 nm analogue I/O library have been connected to each couple of power/ground pads *VDDA*/*GNDA* and *VDDD*/*GNDD*. These cells are used to specifically protect core power/ground networks from ESD surges. A detailed view of the final I/O power assignment scheme is presented in Figure 3.31. Note that power-cut cells only break I/O supply rails but guarantee continuity for the pad ring internal ground connected to *VSSA*/*VSSD*.
With a few modifications, the same I/O power distribution scheme has been adopted for the assembling of the CHIPIX_VFE1/2x1. A different solution was chosen for CHIPIX_VFE1/PV test structures instead [Gaioni 2014].

Pad name	Power/ground functionality
VSSA, *VSSD*	internal ground for I/O cells
VDDA	1.2 V core analogue voltage and supply voltage for I/O cells in the analogue domain
VDDD	1.2 V core digital voltage and supply voltage for I/O cells in the digital domain
GNDA	core-only analogue ground
GNDD	core-only digital ground

Table 3.3: I/O power distribution scheme adopted for the CHIPIX_VFE1/TO chip.

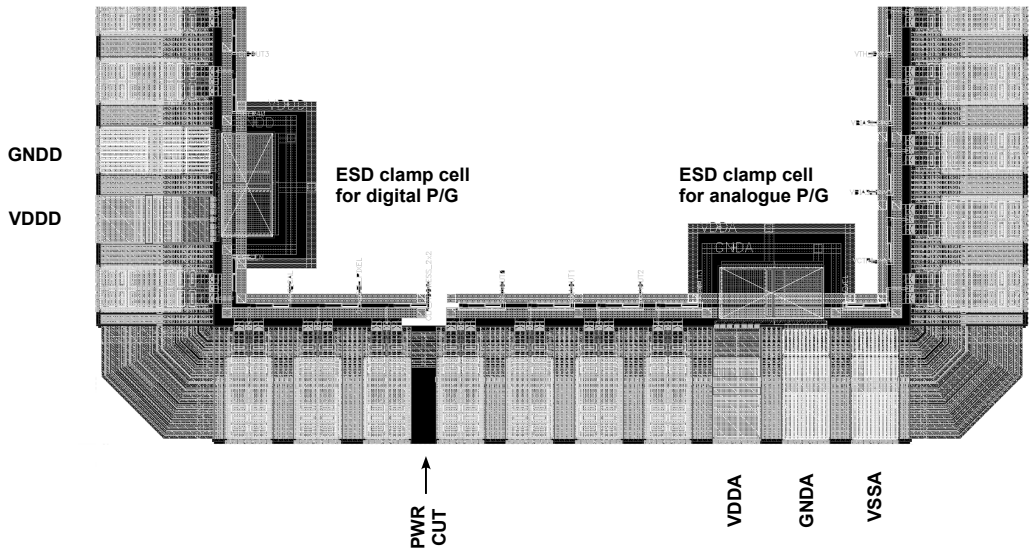


Figure 3.31: Final I/O power assignment. When available, layout views have been substituted to black-box abstract cells. *VSSD* and the second power-cut cell are not shown.

The actual pad assignment was chosen instead upon a wire-bonding floorplan. The position of the CHIPIX_VFE1/TO chip at the bottom-left of the overall CHIPIX_VFE1 area introduces in fact a few additional considerations. Figure 3.32 shows the custom chip footprint designed by Torino INFN for the CHIPIX_VFE1/TO test board [Rotondo 2014]. In order to take into account the inductive behaviour of bonding wires, all bias currents and analogue/digital steady voltages have been assigned to pads on top and right sides of the chip, which are supposed to have longer wires and therefore a larger inductance. Switching signals and power/ground cells have been assigned to bottom and left sides instead, which foresee shorter bonding wires, hence lower impedance. Certainly, most critical digital switching signals are the master clock $CLK40$, the global reset $RESET\_DFF\_TOT$ for the latch control logic and the timing pulse $TEST\_EN$ for the test charge injection. As already discussed these signals are CMOS, without the usage of differential I/O cells. As usually requested to ensure good signal integrity in printed circuit board design, critical signals or clock traces should be always routed between power/ground lines. Hence these signals have been assigned to pads placed near a power or ground cell in order to guarantee a proper return current path [Kirkpatrick 1989, Montrose 2000, Khandpur 2005].
Referring to the numbering scheme shown in Figure 3.33, the complete CHIPIX_VFE1/TO pad list is presented in Table 3.4. Not surprisingly, apart from necessary bias and output signals most of pads provides some digital functionality (configuration bits, pixel addressing, clock and control signals). Note that only 50 fF and 100 fF capacitance values can be selected to mimic a sensor input capacitance. Indeed, not enough pads have been available to implement a SEL_CIN_25F configuration bit. This bit has been therefore fixed to high at the periphery of the core matrix. As a result, in all CHIPIX_VFE1/TO pixels the 25 fF capacitor placed in each cell is always connected at the input of the Front-End chain. This limitation does not affect instead CHIPIX_VFE1/2x1 pixels, in which a digital serial configuration is performed by means of shift registers, as described in the next section.

As a final improvement against ESD effects, custom secondary ESD protections have been added for all I/O signals in respect of further reliability rules proposed by the manufacturer. As depicted in Figure 3.34, this requires the design of a full-custom cell with proper transistor sizing. The cell consists of a resistor connected in series with a PMOS protection device to the power rail and a NMOS counterpart to the ground rail. The cell is then placed in between a pad and the core signal to be connect to. The pad frame has been therefore equipped with a couple of additional power/ground metal rails placed along each internal side of the I/O ring which secondary ESD cells are connected to [Loddo 2014]. A basic cell was already available from Milano INFN [Stabile 2014] and has been adopted in other designs of the first CHIPIX65 submission as well. In order to reduce the effective resistance, each rail is composed by two different 10 μm width metal layers superimposed and connected in parallel through arrays of via-stacks. As shown in Figure 3.35, these auxiliary rails are interrupted in correspondence of power-cut cells such that analogue and digital power domains are independent also for secondary ESD devices. In perspective of simple characterizations of pixel Front-End waveforms at the oscilloscope, the extra filtering introduced by additional elements must be considered. As derived from simulations, the filtering introduced by a secondary ESD cell is actually negligible. The largest contribution comes instead from output buffers connected to secondary ESD protections. As discussed, they have been implemented as large source followers in order to drive a load capacitance up to pF values, which are typical values of instrumentation probes.

A description of most interesting issues encountered in full-chip verifications and final sign-off checks up to the foundry transfer are remanded to Section 3.7, after a description of the integration of the CHIPIX_VFE1/2x1 chip.

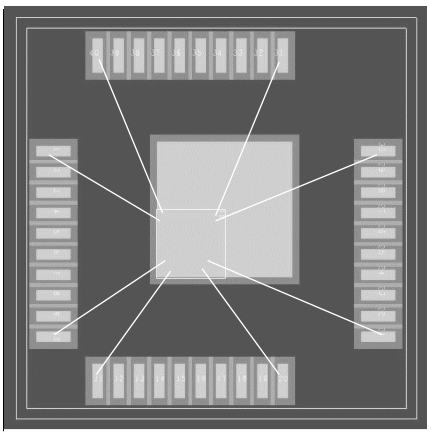


Figure 3.32: CHIPIX_VFE1/TO chip footprint [Rotondo 2014]. The chip is foreseen at the bottom-left, hence pads on the left and bottom sides exhibit shorter bonding wires.

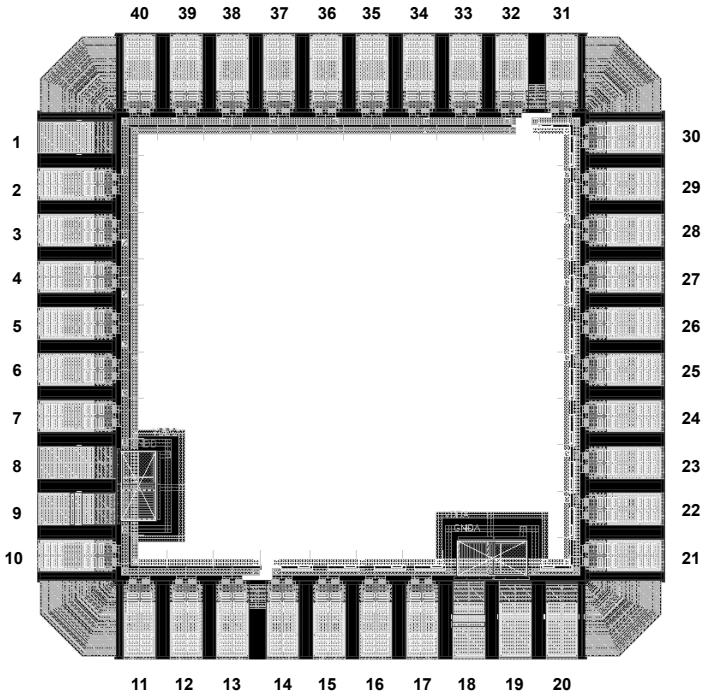


Figure 3.33: CHIPIX_VFE1/TO final pad frame layout. Power cuts are well visible at the top-right and at the bottom-left of the pad frame, in correspondence of two gaps between secondary ESD cells power/ground rails.

Number	Pad name	Functionality/specifications
1	*VSSD*	pad frame internal ground rail, same as *VSSD*
2	*RESET_DFF_TOT*	latch control logic global reset
3	*DIGOUT*0	ch0 discriminator output waveform (hit pulse/fast clock)
4	*DIGOUT*1	ch1 discriminator output waveform (hit pulse/fast clock)
5	*DIGOUT*2	ch2 discriminator output waveform (hit pulse/fast clock)
6	*DIGOUT*3	ch3 discriminator output waveform (hit pulse/fast clock)
7	*CLK*40	40 MHz master clock, CMOS
8	*GNDD*	digital ground, 0 V
9	*VDDD*	digital supply voltage, 1.2 V
10	*TESTP*	timing digital pulse for the test charge injection circuit
11	*PHI_CAL*	main control signal to generate discriminator ϕ-signals
12	*CLK_SEL_PIXEL*	enables the test calibration among a group of 4 pixels
13	*CLK_ADDRESS_*2x2	selects a group of 2 $\times$ 2 pixels to be read out
14	*ANAOUT*0	ch0 Front-End amplifier output waveform
15	*ANAOUT*1	ch1 Front-End amplifier output waveform
16	*ANAOUT*2	ch2 Front-End amplifier output waveform
17	*ANAOUT*3	ch3 Front-End amplifier output waveform
18	*VDDA*	analogue supply voltage, 1.2 V
19	*GNDA*	analogue ground, 0 V
20	*VSSA*	pad frame internal ground rail, same as *VSSD*
21	*CAL_LEVEL*	DC voltage for the test charge injection circuit
22	*VCTRL_TOT*	delay-line control current
23	*VBIAS_BUFF*	on-pixel analogue output buffer bias current
24	*VBIAS_DISC*	discriminator bias current
25	*VTH_DISC*	global threshold voltage
26	*VBL_DISC*	discriminator common-mode voltage during autozeroing
27	*VREF_KRUM*	Krummenacher feedback DC reference voltage
28	*VBIAS_FEED*	Krummenacher feedback bias current
29	*VBIAS_SF*	core amplifier output source follower bias current
30	*VBIASP*2	core amplifier bias current
31	*VBIASP*1	core amplifier bias current
32	*SUBSTRATE_NOISE*	enables/disables the ring oscillator in the digital part
33	*DIGOUT_EN*	enables/disables digital outputs
34	*ANAOUT_EN*	enables/disables analogue outputs
36	*TOT_EN*	enables/disables latch operations as a local oscillator
36	*SEL_CIN*100*F*	connects/disconnects 100 fF input capacitance
37	*SEL_CIN*50*F*	connects/disconnects 50 fF input capacitance
38	*SEL_C*4*F*	connects/disconnects 4 fF feedback capacitance
39	*SEL_C*2*F*	connects/disconnects 2 fF feedback capacitance
40	*RESET_DFF_ADDRESS*	global reset for the pixel addressing logic

Table 3.4: CHIPIX_VFE1/TO final pad list.

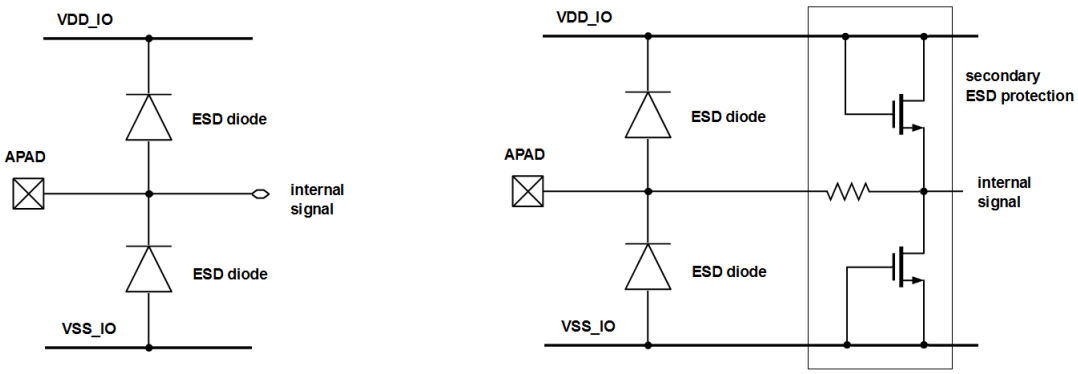


Figure 3.34: Generic analogue I/O cell (left) and addition of a secondary ESD protection (right). The adopted solution uses a 220 Ω POLY resistor connected to 30 μm-width and 1 μm-length devices [Stabile 2014].

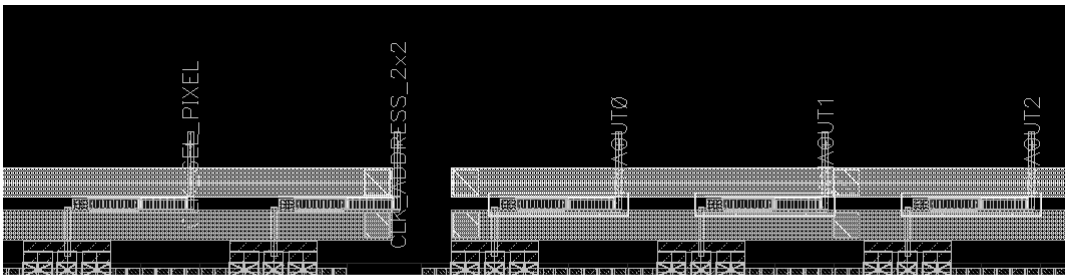


Figure 3.35: Secondary ESD protections placement. Each cell connects to dedicated power/ground metal rails placed along each internal side of the I/O ring. Note the gap between the I/O analogue domain (right) and the digital domain (left).

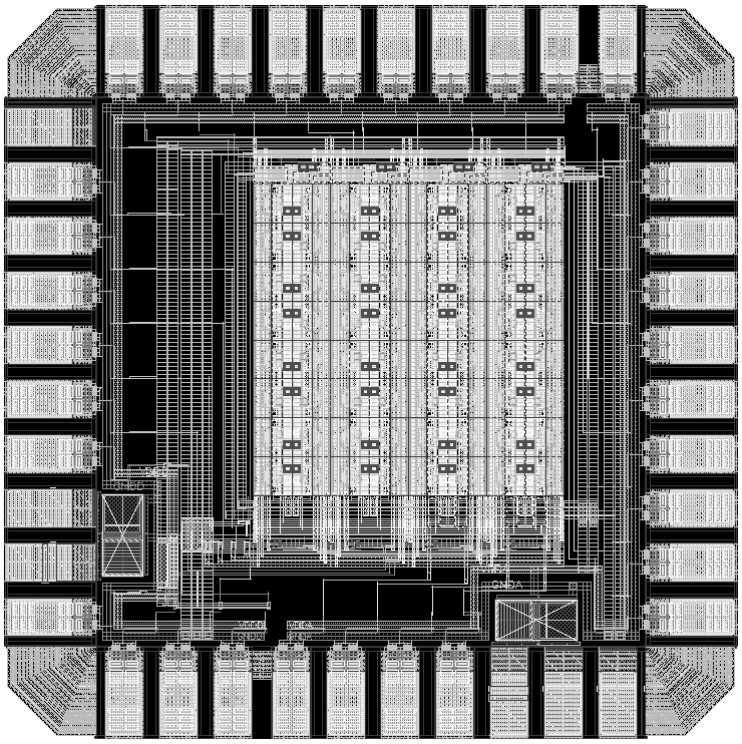


Figure 3.36: Final layout of the CHIPIX_VFE1/TO chip, 945 μm $\times$ 945 μm. 8 $\times$ 8 pixels with synchronous Front-End and full-analogue readout.

3.6 CHIPIX_VFE1/2x1 assembling

CHIPIX_VFE1/2x1 is the largest chip submitted as part of CHIPIX_VFE1 prototypes. The final layout is presented in Figure 3.37. As mentioned, the chip includes two independent small pixel arrays of 50 μm $\times$ 50 μm cells arranged into 8 rows $\times$ 12 columns each one. Pixels in the top matrix implement a continuous-time Front-End proposed by Pavia INFN [Gaioni 2014, Ratti 2014]. The bottom matrix uses instead the synchronous Front-End architecture proposed by Torino INFN and already implemented in the CHIPIX_VFE1/TO chip. Indeed, only the layout of the analogue pixel cell is common to both chips. A different digital part designed by Pisa INFN [Morsani 2014] with personal contributions provides in fact serial readout and pixels configuration by means of shift-registers. Apart from necessary customizations in the interfacing with two different versions of Front-End discriminators[5], serial readout and configuration are common to both Pavia and Torino matrices. The time-over-threshold information is retrieved in digital form with an 8-bit binary counter placed in each pixel cell. An external clock is provided to continuous-time pixels, whereas a local high-frequency clock is available in synchronous pixels by turning the latch into a voltage-controlled oscillator. Only a digital information is extracted from CHIPIX_VFE1/2x1 pixels, without direct probing capabilities for Front-End analogue waveforms. Same top-level considerations already discussed for the assembling of the CHIPIX_VFE1/TO chip in terms of pad frame generation, power partitioning and distribution, digital signals distribution, bias cells and secondary ESD protection placement apply also for the integration of the CHIPIX_VFE1/2x1 chip. Hence they will be only shortly summarized after a schematic-level description of the digital circuitry that provides TOT registering, serial readout and pixels configuration.

Serial readout and configuration

A schematic block diagram of the on-pixel TOT registering logic that couples to synchronous discriminators is shown in Figure 3.37. With a few modifications, the latch control logic is the same already described in Chapter 2 and uses a 4:1 multiplexor that toggles comparator operations. The resulting self-generated clock is fed to an 8-bit synchronous binary counter and a *count_en* signal obtained from leading-edge and trailing-edge hit flags is used to enable the counter and register the TOT count. Furthermore, in order to save power, a *carry_out* signal stops the counter and breaks the asynchronous feedback path around the latch in case of a count saturation is reached. The *MASK* configuration bit is used to enable/disable the propagation of the master clock *CLK*40 to the latch. The *TOT_EN* bit is used instead to enable/disable latch operations as a local oscillator. Moreover, when *TOT_EN* is set to low *count_en* is forced to low as well, hence the counter is disabled and only a binary information is available. Two independent external active-low reset signals *CLRN_CTRL* and *COUNT_RN* are used to properly reset FlipFlops of the control logic and the TOT counter respectively. The major modification with respect to the simpler logic implemented in CHIPIX_VFE1/TO pixels resides in the usage of an external *START* signal that opens a well-defined acquisition time-window before data can be serially read out. That is, pixels can be electrically pulsed within a time-window defined by the *START* command, decoupling configuration, test charge injection and data readout into three totally independent phases. The acquisition timing in CHIPIX_VFE1/TO pixels is instead uniquely defined by the reset signal *RESET_DFF_TOT* used to initialize state FlipFlops that register leading-edge and trailing-edge hit transitions, as discussed. With necessary customizations and simplifications, the same digital circuitry is used to retrieve a TOT count in continuous-time pixels.

[5] Continuous-time with single-ended digital ouput for Pavia, synchronous with differential outputs for Torino

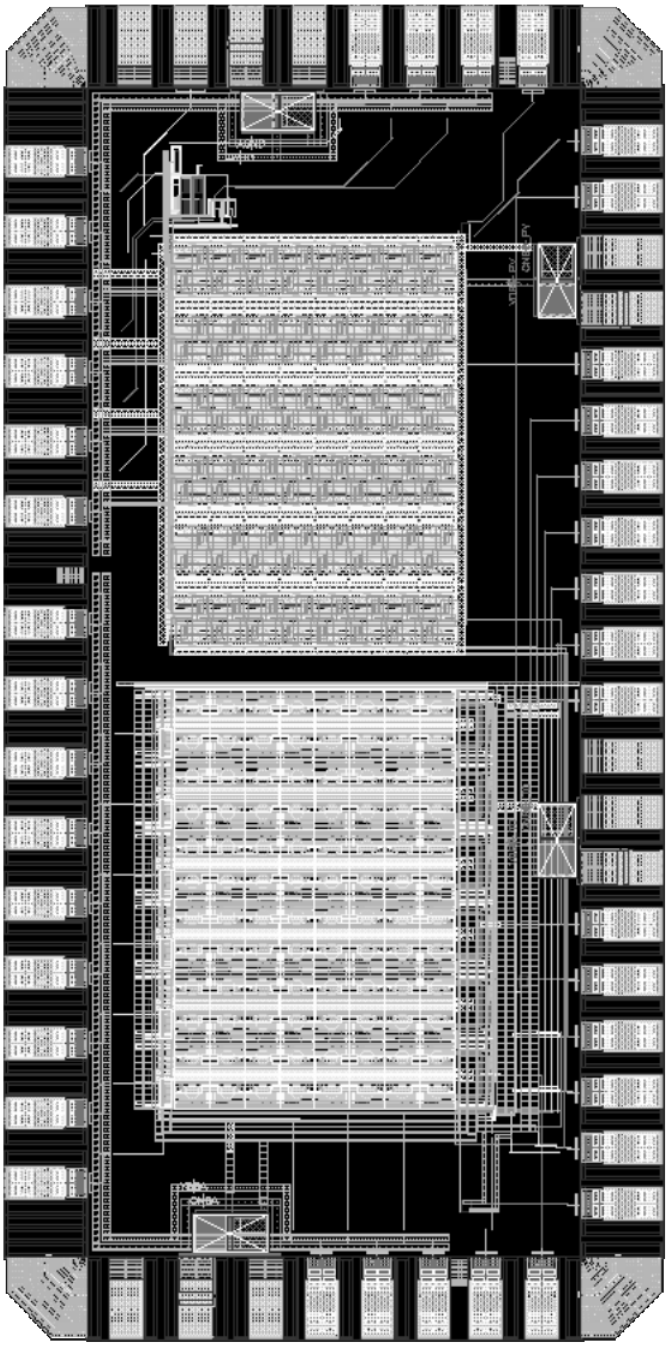


Figure 3.37: CHIPIX_VFE1/2x1 chip, 945 μm $\times$ 1910 μm. The chip includes two independent pixel arrays of 8 $\times$ 12 cells designed by Pavia INFN (top) and Torino INFN (bottom). Serial readout and pixels configuration are provided by a digital part designed by Pisa INFN with personal contributions and common to both matrices.

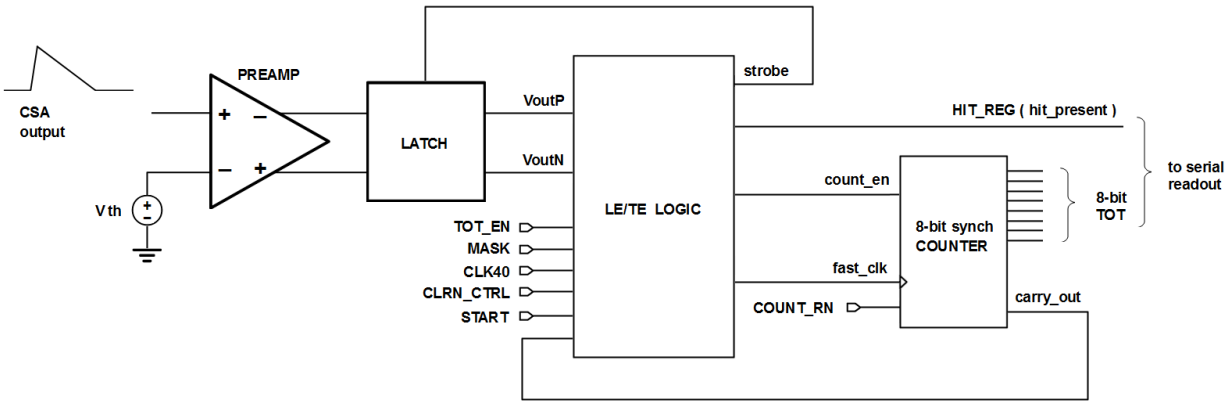


Figure 3.38: Block diagram of the on-pixel control logic implemented to register the TOT information in synchronous pixels of CHIPIX_VFE1/2x1. With a few modifications, the latch control logic uses the same operating principle already described in Chapter 2.

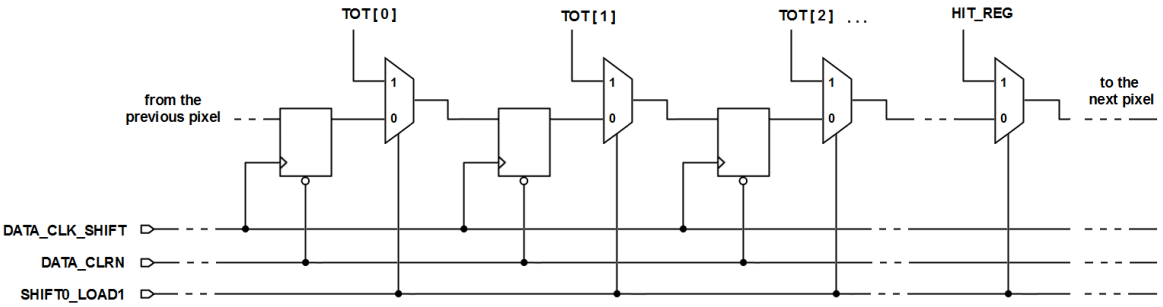


Figure 3.39: The digital serial readout included in CHIPIX_VFE1/2x1 pixel cells uses a 9-bit parallel-in serial-out (PISO) shift register placed in each cell and chained from pixel to pixel. Both the hit binary-only information and a TOT word with 8-bit resolution are presented to the serial readout.

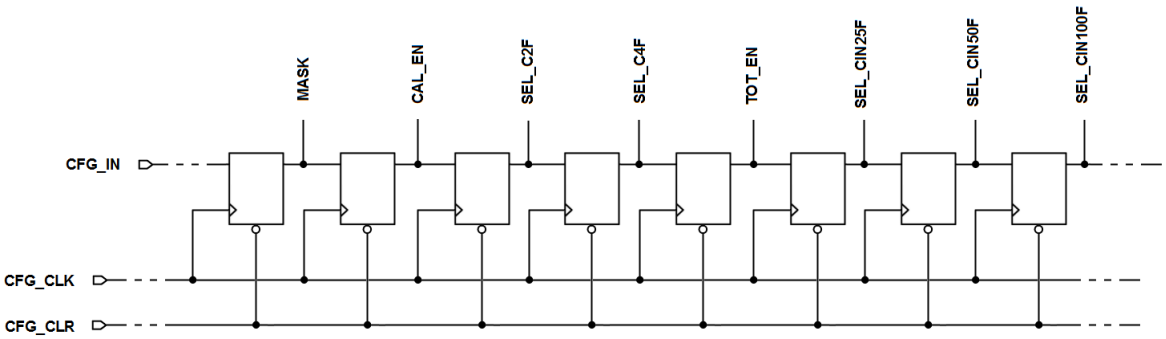


Figure 3.40: Pixels configuration is provided by a serial-in parallel-out (SIPO) shift register placed in each cell and chained from pixel to pixel.

Configuration bit	Functionality
$MASK$	enables/disables the master clock $CLK40$ in the pixel
CAL_EN	enables/disables test charge injection
TOT_EN	enables/disables latch operations as a local oscillator
SEL_C4F	connects/disconnects 4 fF CSA feedback capacitance
SEL_C2F	connects/disconnects 2.5 fF CSA feedback capacitance
SEL_CIN25F	connects/disconnects 25 fF input capacitance
SEL_CIN50F	connects/disconnects 50 fF feedback capacitance
$SEL\_CIN100F$	connects/disconnects 100 fF feedback capacitance

Table 3.5: Configuration bits loaded and stored in each pixel cell implementing the synchronous Front-End architecture. All bits are active-high status flags. Note that no configuration bits are required for the local threshold adjustment.

As shown in Figure 3.39, the digital readout is implemented with a 9-bit parallel-in and serial-out (PISO) shift-register placed in each pixel cell and connected among consecutive pixels to form a chain. An external *SHIFT*0_*LOAD*1 signal defines the functionality of the block. When asserted to high, digital values are loaded in parallel at FlipFlop inputs. Both the binary-only hit information *HIT_REG* (the *hit_present* flag) and the 8-bit TOT count are presented to the readout. When *SHIFT*0_*LOAD*1 is set to low instead, all FlipFlops connect in shift-register configuration and perform a right-shift operation using an external clock *DATA_CLK_SHIFT*. The 8 × 12 pixels are serially read out and an 864-bit data stream is available at the *DATA_SHIFT_OUT* pad. A *DATA_CLRN* reset signal is then used to properly initialize the chain. The same readout has been implemented in both Pavia and Torino matrices.
All necessary configuration bits for the pixel Front-End are stored instead in FlipFlops placed in each pixel cell and chained from pixel to pixel in order to form a serial-in and parallel-out (SIPO) shift-register. As shown in Figure 3.40, a configuration word *CFG_IN* provided off-chip with suitable instrumentation[6] stores the configuration bits for all pixels of a 8 × 12 matrix. A clock signal *CFG_CLK* is then used to serially load the configuration in each pixel cell, whereas a *CFG_CLR* is used to reset the shift-register chain. A unique 12-bit shift-register is placed in each pixel cell and is common to both Pavia and Torino matrices. Then, pixels connect to proper configuration bits depending on the actual Front-End version. Furthermore, *MASK*, *CAL_EN* and *TOT_EN* are used in both Front-End solutions. Thanks to the usage of an autozeroed solution, no configuration bits are required instead for the local threshold adjustment in synchronous pixels, whereas a 4-bit word is fed to an on-pixel D/A converter in continuous-time cells. Configuration bits required in synchronous pixel cells are summarized in Table 3.5. Note that the same configuration bits represent standalone pads or are not available in the CHIPIX_VFE1/TO chip.
For debug purposes, the input of the readout chain and the output of the configuration chain can be accessed using two auxiliary pads *DATA_SHIFT_IN* and *CFG_SHIFT_OUT* respectively. Proper functionality of both shift-register chains can be therefore validated with input test patterns.
As a matter of fact, the support of a pixel-by-pixel digital configuration and readout offers a simple yet flexible solution that maximizes the overall testability. As a most notable example, the test charge injection circuit can be activated in more than one pixel at the same time, allowing for crosstalk studies. On the other hand, only a digital information is provided, thus Front-End performance in terms of noise and hit efficiency can be only extracted with threshold scans and S-curves, as usually performed in binary systems [Spieler 2005, Rossi 2006].

Chip integration

The integration of CHIPIX_VFE1/2x1 has been accomplished in a close collaboration among Pavia, Pisa and Torino INFN ASIC designers. Indeed, core pixel matrices are two totally independent systems in terms of power distribution, bias, test, readout and configuration. The common pad frame has been therefore assembled such that only one pixel matrix at a time can be powered, configured, electrically pulsed and read out.
Apart from necessary routing customizations, the layout of the analogue cell already implemented in CHIPIX_VFE1/TO was adopted for CHIPIX_VFE1/2x1 synchronous pixels. A different custom solution for the layout of the on-pixel digital part has been instead designed from scratch including the above discussed functionalities [Morsani 2014].

[6] Logic State Analyzer (LSA) or Field Programmable Gate Array (FPGA)

In both core matrices, the layout of a 2 × 2 pixel region with upside-down and left-right symmetry was then assumed as fundamental replica unit-cell for the assembling of double-columns, as already performed in CHIPIX_VFE1/TO. Furthermore, both matrices use the same bump bonding pad already designed for CHIPIX_VFE1/TO pixels, placed at the centre of each 50 μm × 50 μm cell. In order to have a uniform 50 μm-pitch bump pattern across the entire chip, pixel matrices have been aligned with 100 μm spacing.
Clocks and other digital signals have been propagated with non-zero skew along columns from pixel to pixel, performing necessary local buffering at the pixel level and from pads to ends of columns. Same bias cells already implemented for CHIPIX_VFE1/TO have been placed at the end of each double-column of 8 × 2 pixels. It is therefore required that external bias currents are 6 times larger than nominal values.

The layout of the pad frame was derived from the one already assembled for CHIPIX_VFE1/TO. Thanks to the support of digital readout and pixels configuration, the pad count did not represented a limiting factor. The final chip only requires 48 pads, despite a few spares have been included. Basically, the same I/O power distribution scheme already discussed for CHIPIX_VFE1/TO was adopted for the CHIPIX_VFE1/2x1 chip. The ring has been partitioned into three sub-rings using power cut cells. I/O analogue and digital power domains are therefore separated. Power cells provide core and I/O supply voltages, whereas core and I/O ground rails use different ground cells. The most notable difference with respect to CHIPIX_VFE1/TO is the usage of totally independent pairs of power/ground cells, both in the common digital part and in each analogue sub-ring. That is, a first pair *VDDD_PV*/*GNDD_PV* provides the digital power/ground rails to the Pavia matrix while keeping the Torino counterpart off. Similarly, a second pair *VDDD_TO*/*GNDD_TO* provides the core digital power to the Torino matrix while keeping the other one off. As already discussed, the ground rail of I/O cells is instead unique. It was connected to *VSSA_TO*, *VSSA_PV* and *VSSD* pads, which are actually internally shorted. For the sake of completeness, an independent substrate rail has been adopted in the Pavia pixel matrix, thus requiring a dedicated *SUBA* pad implemented with a general purpose I/O cell. Finally, clamp cells and secondary ESD protections have been added as already performed in CHIPIX_VFE1/TO. However, due to the necessity of two independent pairs of power/ground cells for the shared digital sub-ring, no secondary ESD protections have been attached to digital pads.
The pad assignment strategy was driven by same wire-bonding considerations already discussed for CHIPIX_VFE1/TO. As a result, almost all digital signals have been placed on the right side of the chip in respect of the same 20 μm minimum spacing constraint already adopted in CHIPIX_VFE1/TO. Most of readout and configuration control signal such as *SHIFT*0_*LOAD*1, *START*, shift-registers serial inputs, clocks and resets are pads common to both matrices. Two independent master clocks *CLK*40 and *CLK_GLOBAL* are available instead for Torino and Pavia matrices respectively. Furthermore, serial outputs have been duplicated as well. Most of pads assigned to bias currents and DC voltages occupy instead the left side of the ring, since they are expected to have longer bonding wires. Furthermore, a larger spacing was used for pads on the right side. Analogue power/ground cells are placed on top and bottom sides.

Figure 3.41 shows the final layout of the Torino pixel matrix. In the next section, the most important issues encountered in the full-chip verification up to the foundry transfer are discussed.

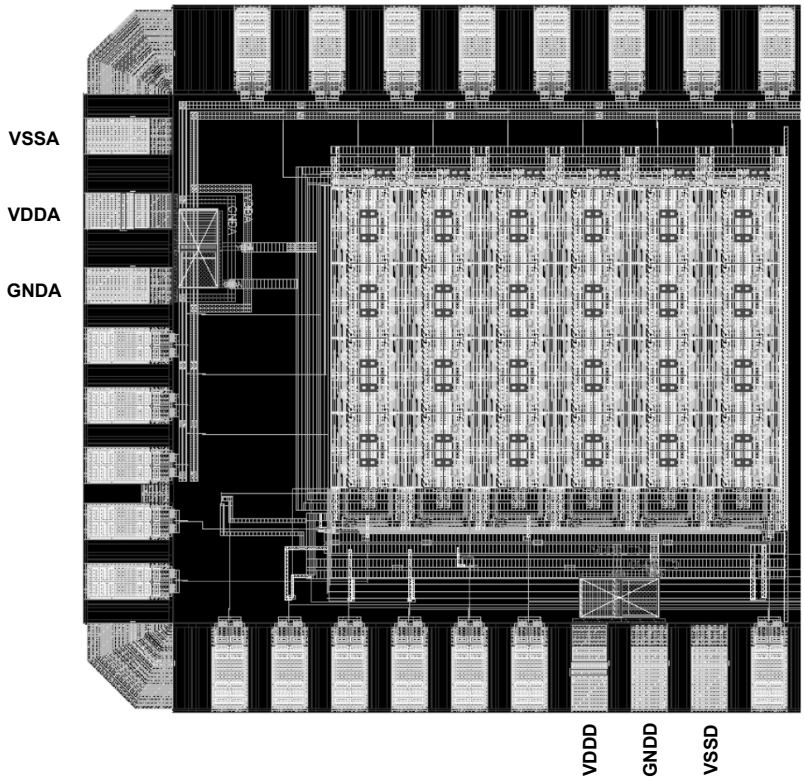


Figure 3.41: CHIPIX_VF1/2x1 core matrix with Torino synchronous pixels. The array contains 96 cells of 50 μm $\times$ 50 μm arranged into 8 rows $\times$ 12 columns. Clamp cells placed in front of power/ground cells indicates the position of *VDDA*/*GNDA* (left in figure) and *VDDD*/*GNDD* (bottom). Totally independent pads are then used to power the Pavia matrix while keeping the Torino counterpart off. Note that no secondary ESD protections are attached to digital pads.

3.7 Sign-off and foundry transfer

After assembling, each chip included in CHIPIX_VFE1 has been validated with extensive design rule checks using full-chip and antenna rule files provided by the manufacturer.
The most notable issue encountered in the sign-off of CHIPIX_VFE1/TO and CHIPIX_VFE1/2x1 prototypes was a different behaviour of the physical verification tool when full-layout views of analogue I/O cells have been available in the chosen 65 nm fabrication technology. The number and type of ESD and latch-up (LUP) errors related to buffers and secondary ESD protections connected to pad frames significantly increased when pad layouts have been substituted to previously adopted black-box abstract views, resulting de facto into a clear inconsistency in the set of rules provided by the foundry. The layout of most of custom designed buffers have been therefore modified in order to prevent new DRC errors, demanding for example additional guard ring structures and more spaced transistors. Similarly, an additional guard ring was required to secondary ESD protections connected to *VDDA* and *GNDA* analogue power/ground rails.
A further interesting aspect resides in DRC errors introduced by the presence of the bump bonding pad in each pixel cell when full-chip design rules have been checked. As already discussed in fact, the bump pad has been designed according to design rules that apply to the flip-chip fabrication technology. These metal pads are considered by the verification tool as I/O cells to all effects when full-chip rules are checked. The input transistor of the analogue pixel Front-End chain exhibits therefore same ESD and LUP issues that usually affect I/O cells in the pad frame. Hence the layout of the basic replica 2×2 pixel region has been modified by adding specific non-physical exclude-layers in order to prevent bump pads to cause DRC errors.
Due to an initial usage of too-wide metal stripes for output signals $ANAOUT0$-$ANAOUT3$ and $DIGOUT0$-$DIGOUT3$, antenna-rule violations had to be fixed in the CHIPIX_VFE1/TO chip with layout modifications. Metal filling has been automatically performed instead by the foundry access service, hence metal density-rule violations did not represent an issue. Nevertheless, this task has been performed by hand for the layout of the basic analogue pixel cell to have a complete control with respect to fully-automated routines. Dummy floating metals have been therefore added in order to fully satisfy metal density rules in each analogue pixel cell. Design rules that specifically apply to MIM capacitors have been checked as well for CHIPIX_VFE1/PV and CHIPIX_VFE1/2x1 to properly take into account the usage of MIM capacitors in Front-End test structures and pixel array proposed by Pavia INFN.

The final layout of CHIPIX_VFE1 has been transferred to the foundry access service using a well-known industry-standard binary database adopted for the representation of mask layouts of integrated circuits. All layout data (layer and geometrical coordinates of each shape, cells, text labels, pins and other information) have been therefore hierarchically translated into a binary stream. All aspects and options of this procedure must be carefully understood in order to guarantee a perfect equivalence between the two designs and prevent any data corruption during the exchange process. Quite often in fact the design export can result into missing cells or wrong layer appearance. Before the actual foundry transfer, the binary database has been back-imported into the full-custom design environment for a final layout-versus-layout (LVL) check in order to ensure the equivalence between the foreseen submitted design and the exported layout. Since no LVL rule file was provided by the foundry, it was necessary to automatically generate a rule file from scratch with the aid of a verification tool and its LVS extraction rules.

3.8 Standard-cell based design of a column/row decoder for a current-steering DAC

The remaining part of this chapter is dedicated to personal contributions to IP block design and describes a fully-automated standard-cell based implementation of an 8-bit binary-to-thermometer column/row decoder. The layout of such a digital block was then employed in a 10-bit segmented current steering D/A converter for biasing designed by Bari INFN [Loddo 2014] as part of the RD53 collaboration IP block research program with INFN commitment. A dedicated DAC chip has been put on silicon in the first October 2014 CHIPIX65 submission. In order to investigate radiation damage effects on digital gates provided by the 65 nm logic library, the chip includes two data converters with different versions of column/row decoders. A first version uses full-custom and DFM-rules compliant cells, whereas the second version is implemented with STD cells. All aspects of the design flow are considered, covering the optimized synthesis of a behavioural Verilog HDL description of the block, the automated place-and-route (PNR) into the digital design environment and the final design import and verification back into a full-custom environment. The design of this block validated for the first time in Torino the fully-automated digital implementation flow attached to 65 nm CMOS technology.

DAC architecture and specifications

Current-steering D/A converters are fundamental building blocks in pixel ASIC design. They are commonly used wherever a precise, reliable and digitally-programmable bias current is required. Resolutions of 10 to 12-bit are usually demanded to generate reference currents for the analogue Front-End. Given the importance and widespread use of D/A converters in mixed-signal systems, extensive theoretical analyses and CMOS circuit topologies are discussed in dedicated reference literature [Razavi 1994, De Plassche 2003, Manganaro 2011, Radulov 2011]. In the following, the implemented DAC architecture and its specifications are briefly introduced. Most of the attention will be devoted instead to the practical implementation of the column/row DAC decoder with personal contributions.

In current-steering D/A converters a certain number of shunt current sources connected to a common output node by means of switches is summed to obtain the desired current I_{out}. As shown in Figure 3.42, in binary architectures current source values are binary-weighted. For a given maximum full-scale current I_{REF} and N-bit resolution, the least significant bit (LSB) current is $I_{LSB} = I_{REF}/2^N$. Thus N binary-weighted current sources I_{LSB}, $2I_{LSB}$, $4I_{LSB}$, ... $2^{N-1}I_{LSB}$ controlled by a binary word are required to achieve a N-bit resolution. As an alternative to binary weighting, a thermometer summing scheme can be adopted. In such architectures the output current is obtained by simply adding together a number of nominally identical unit current sources I_{LSB}, as sketched in Figure 3.43. In this case switches are controlled by a thermometer code, thereby a DAC with a N-bit resolution requires $2^N - 1$ unit current sources. As a matter of fact, each architecture offers both advantages and drawbacks. Given resolution and LSB current requirements, the choice of a binary solution leads to a more compact layout and much simpler circuitry with respect to a thermometer approach. In fact, since for a N-bit resolution the number of switches is N, they can be directly controlled by a bundle of N external digital signals without the need of any additional decoding logic. However, binary-weighted current sources are much more sensitive to mismatches and process variations. Monotonicity in the DAC characteristic is not a priori guaranteed and any bit variations in the middle of the transfer function can result into significant non-linearity errors. Proper transistor sizing and extensive MC simulations are therefore required. Moreover, a careful layout with common-centroid placement is mandatory, thus increasing the routing complexity [Razavi 2000, Hastings 2001, Saint 2002].

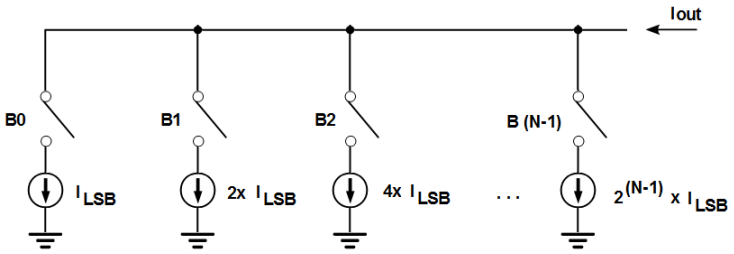


Figure 3.42: Principle of operation of a binary-weighted current-steering D/A converter.

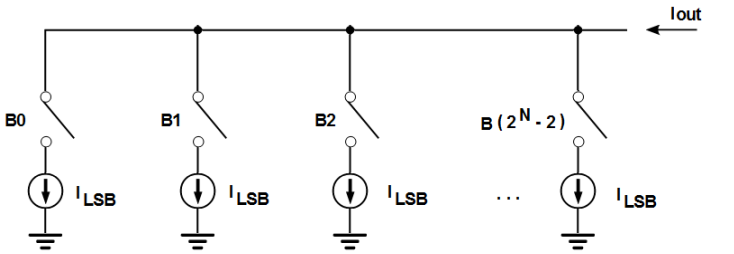


Figure 3.43: Principle of operation of a thermometer current-steering D/A converter.

Matching and layout requirements are more relaxed in thermometer architectures instead. When the digital input increases by one LSB in fact, one more unit current I_{LSB} is connected to the output node, without turning off current sources that are already switched to the output. Thus monotonicity is always guaranteed in the input/output transfer characteristic. The layout of the basic unit current source can be therefore simply replicated without the need of the common centroid technique. The main drawback of a thermometer encoding resides in the large number of switches and unit currents in the DAC array, resulting into a much larger layout area with respect to a binary architecture. For a N-bit resolution the number of switches and current sources increases to $2^N - 1$. Certainly it is totally unrealistic to directly control switches with a bundle of $2^N - 1$ digital signals if 10-12 bit resolutions are demanded. Hence the usage of an additional binary-to-thermometer decoder is required, such that a N-bit binary input word is translated into a $(2^N - 1)$-bit thermometer code. Unit currents must be arranged in form of a matrix of cells in order to minimize the overall area, hence the binary-to-thermometer decoder is partitioned in a column decoder and a row decoder. As a result, each unit cell must include a few additional local column/row decoding logic to properly generate a logic signal that connects or disconnects the current source to the output. An example 3-bit binary-to-thermometer conversion is presented in Table 3.6, whereas a schematic block diagram of a 6-bit thermometer DAC is depicted in Figure 3.44.

Decimal	Binary	Thermometer
0	000	0000000
1	001	0000001
2	010	0000011
3	011	0000111
4	100	0001111
5	101	0011111
6	110	0111111
7	111	1111111

Table 3.6: A 3-bit binary-to-thermometer conversion example. The number of logic 1s in the thermometer code is the decimal equivalent of the codified binary word.

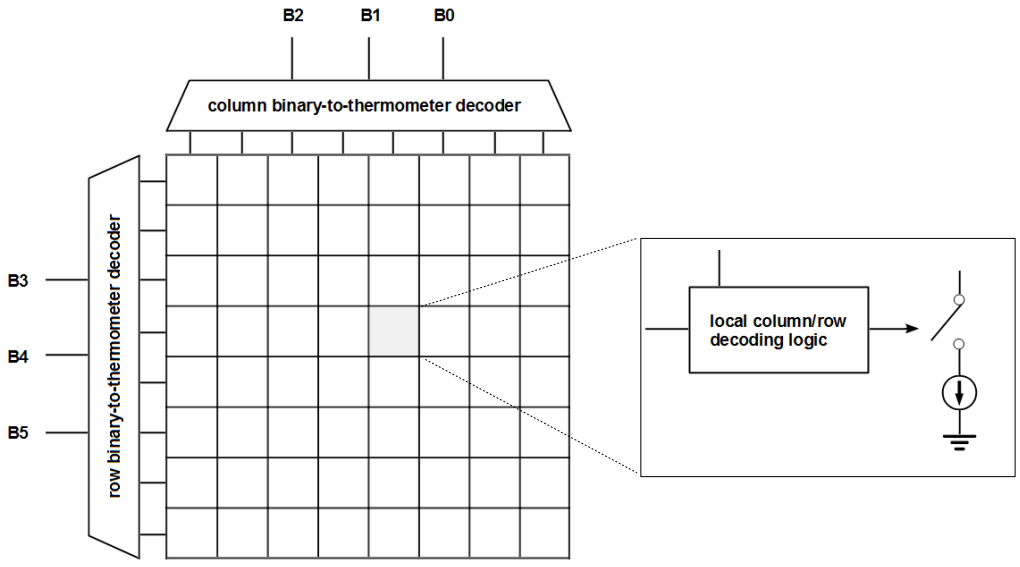


Figure 3.44: Block diagram of a 6-bit thermometer current-steering D/A converter [Razavi 1994].

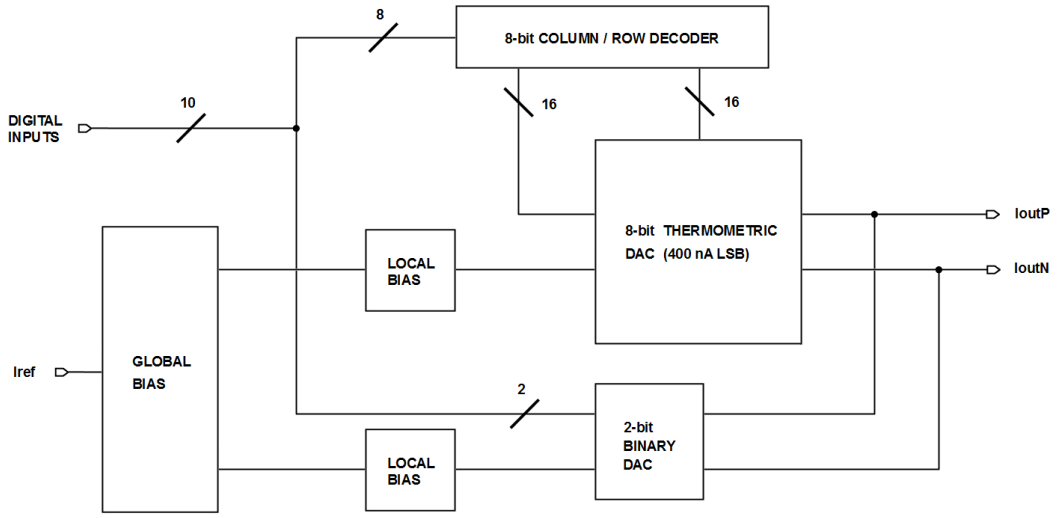


Figure 3.45: Block diagram of the implemented 10-bit segmented current-steering DAC architecture [Loddo 2014].

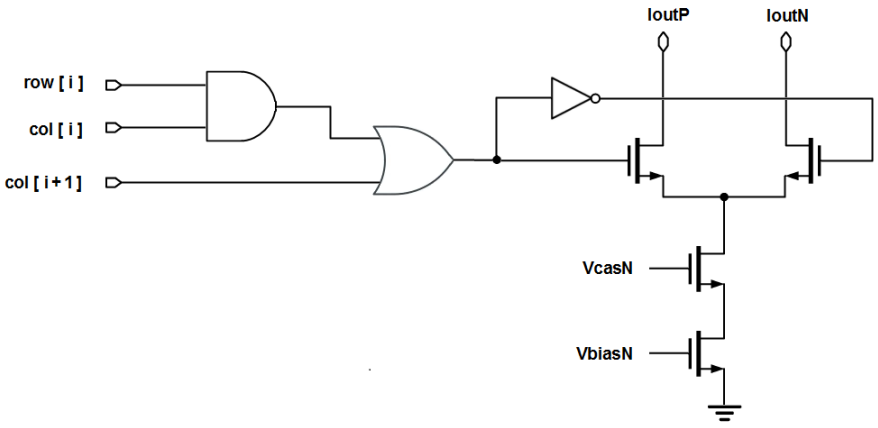


Figure 3.46: Replica unit-current cell for the implemented 8-bit thermometer sub-DAC. The current source uses a NMOS cascode. NMOS switches driven by full-custom logic gates are then required for local row/column decoding [Loddo 2014].

Parameter	Value
Architecture	segmented
Resolution	10-bit (2-bit binary + 8-bit thermometer)
Supply voltage	1.08 V - 1.32 V
Temperature range	-20 °C / +50 °C
LSB current I_{LSB}	25 nA - 100 nA depending on I_{REF}
Max. full scale current I_{REF}	102.3 uA
DNL	< 0.2 LSB
INL	< 1 LSB
Static power consumption	150 μW (I_{LSB} =100 nA), 40 μW (I_{LSB} = 25 nA)
Layout size	240 μm $\times$ 140 μm

Table 3.7: Final current-steering DAC specifications [Loddo 2014].

It turns out that the best compromise among resolution, area and linearity can be obtained by combining binary and thermometer architectures. This is performed in segmented current-steering DACs. In such architectures, a N-bit resolution DAC is partitioned into two independent sub-DACs whose output currents are simply added. A fine sub-DAC implements the first k least significant bits with a binary architecture, whereas an auxiliary coarse thermometer sub-DAC provides the remaining $m = N - k$ most significant bits. With such a solution, the resulting system has good linearity and monotonicity offered by the thermometer encoding with a reduction of the total area thanks to the usage of binary-weighted current sources for the fine current tuning. For a given resolution, a proper choice for k shows an optimum point in terms of differential non-linearity (DNL), integrated non-linearity (INL) and area.

A segmented architecture 8-bit/2-bit has been adopted for the design of a 10-bit biasing DAC in 65 nm CMOS technology [Loddo 2014]. A block diagram of the converter is shown in Figure 3.45. Most important design specifications and simulated performance are summarized in Table 3.7. The DAC receives a 10-bit binary input word which controls an 8-bit thermometer sub-DAC and a 2-bit binary sub-DAC. Two complementary output currents I_{outP} and I_{outN} are generated according to the value of the input word and the chosen reference current I_{REF}. An 8-bit column/row binary-to-thermometer decoder is used to address the unit current sources of the thermometer sub-DAC, arranged into a matrix of 15 $\times$ 15 cells. As shown in schematic of Figure 3.46, each unit current is a NMOS cascode current source. Row and column thermometer values are then locally combined by means of full-custom logic gates which drive a couple of NMOS switches and determine the status of the cell.

In the following, the implementation of the column/row decoder with a fully-automated digital implementation flow attached to the 65 nm logic library is discussed.

HDL design entry and synthesis for the column/row DAC decoder

According to modern top-down ASIC digital design methodologies, as the size and complexity of digital systems increase more design automation is introduced. The design entry is performed at different levels of abstraction with a suitable hardware description language (HDL). The attention is focused to the *functionality* that must be accomplished, without the need to care about of any actual hardware implementation. After extensive design validation using HDL simulations or more advanced functional verification techniques, the generation of an optimized schematic is delegated to a fully-automated hardware-generation tool (synthesizer). Synthesis algorithms translate the HDL description of the digital system into a gate-level netlist with instances of logic cells provided by the target fabrication technology. Only a subset of all available HDL constructs can be mapped to hardware, requiring the usage of well-known standard coding guidelines at the Register Transfer Level (RTL) [Smith 1998, Chang 1999, Chu 2006]. At the end, a place-and-route (PNR) tool is used to generate the physical layout of the digital block or of an entire digital or mixed-signal integrated system if a Digital-on-Top (DoT) approach is adopted. All steps of the automated block-level digital implementation flow has been validated up to the foundry transfer with the design of the column/row decoder using STD cells of 65 nm logic libraries.

Decoders are combinational digital circuits that converts a N-bit binary input word into a certain output code, e.g. thermometer, one-hot, Gray, 8-bit/10-bit etc. Due to the pure combinational nature of such blocks, timing does not represent a key issue as in the design of sequential digital circuits. The logic functionality provided by the required 8-bit binary-to-thermometer column/row decoder can be easily implemented in a compact way with a Verilog HDL[7] behavioural description of the block, using loop and conditional statements similar to those available in more traditional software languages. In practice, the number of logic 1s in the thermometer output word is the decimal equivalent of the codified binary input. The adopted HDL code is reported in the next page, along with a sample testbench simulation in Figure 3.47. The digital block has a bundle of 8-bit inputs and two 16-bit column/row outputs. Indeed, the local decoding logic implemented in each replica unit-current cell combines the logic value of one row with logic values of two adjacent columns, as already shown in Figure 3.46. An additional bit `col[16]` always tied-low is therefore used to properly match the local decoding in the last column. The first 4 least significant bits of the binary input word determine a 16-bit thermometer code for the DAC rows. The remaining 4 most significant bits determine instead a 16-bit thermometer code for the columns. Note that it is expected that `row[15]` is always low, whereas `col[0]` is always high. As discussed shortly thereafter, the presence of logic constants must be carefully taken into account.
The synthesis process is fully-automated and uses scripts with specific commands to properly configure the environment, import the HDL code, constrain the design and constantly instruct the tool at each step of the design flow. These scripts have been written from scratch with necessary references and settings to work with 65 nm digital libraries.
As a first step, the HDL code is loaded into the synthesizer, along with all available technology data, preferences and other settings. At least, STD cells delay, noise and power information must be specified. They are described using an industry-standard format in dedicated timing libraries provided by the manufacturer for many different PVT corners. Since only worst-case delays are important in digital design, worst-case timing libraries have been adopted. Furthermore, as CMOS technologies scale down the parasitic effects due to interconnections become more relevant. As a result, modern synthesis tools improve the synthesis optimization by including also STD cells physical layout information and RC-delay data for the metal interconnections, that are used instead in the PNR flow. A preliminary elaboration of the imported design is then performed in order to create an internal representation of the design hierarchy.

[7] IEEE std. 1364-2001, *IEEE Standard Verilog Language Reference Manual*, 2001

```
1   module dac_decoder(
2      input  [7:0]   Bin,       // binary input word
3      output [15:0]  row,       // and thermometer
4      output [16:0]  col  );    // row/column outputs
5
6      reg [15:0] row;
7      reg [16:0] col;
8
9      integer i;
10
11     // procedural statements
12     always @(Bin)
13     begin
14        for(i = 0; i <= 15; i = i+1)
15        begin
16           row[i] = (Bin[3:0] > i);    // conditional
17           col[i] = (Bin[7:4] >= i);   // assignments
18        end
19
20        col[16] = 0;   // tied-low
21     end
22
23  endmodule
```

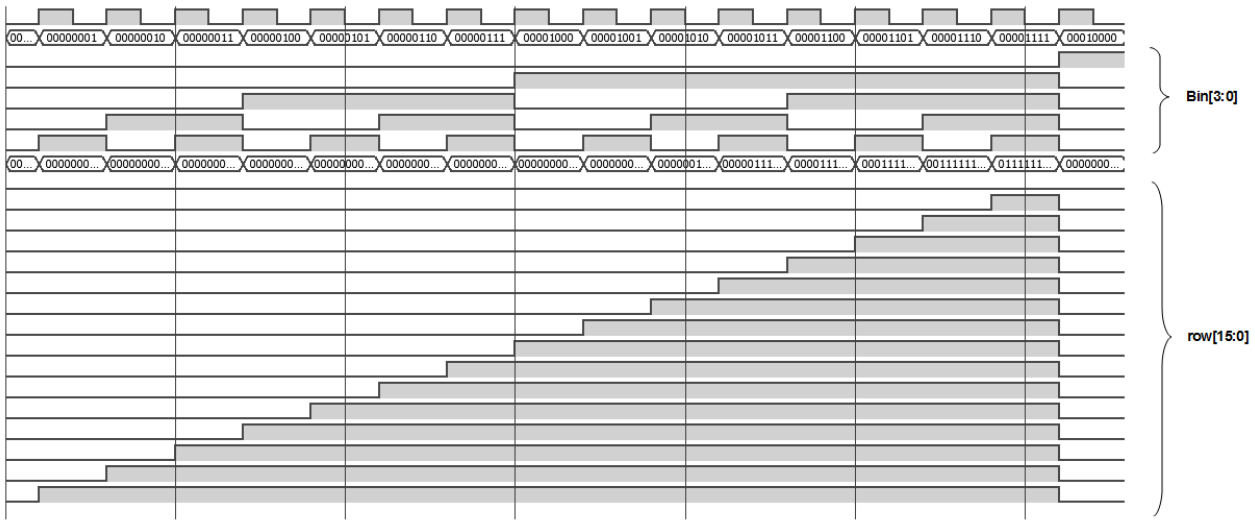


Figure 3.47: Example HDL behavioural simulation of rows binary-to-thermometer conversions. The Verilog decoder module has been coupled to a synchronous counter to generate binary inputs. Arbitrary time scale.

Certainly the most important task in the automated synthesis flow is to properly constrain the design, defining all necessary timing, electrical, power and area requirements for the digital block. Such information is then used by the tool to generate optimized hardware in order to satisfy constrained parameters. Thanks to the absence of clock signals and sequential structures in the DAC decoder, this simply reduces to define the expected drive resistance and load capacitance for all I/O ports. On the one hand, decoder binary inputs connect to analogue input pads through secondary ESD protections in the final chip. Hence each binary input was supposed to see a total 270 Ω input resistance, taking into account the about 220 Ω resistance of the secondary ESD protection and an additional 50 Ω termination which is a typical value for laboratory instrumentation. On the other hand, each thermometer output has to drive the parasitic capacitance of an entire column/row line. The worst-case value for the load capacitance has been estimated to be 500 fF. Since timing closure was not a limiting factor, a relaxed maximum input-to-output propagation delay of 10 ns has been demanded instead. All design constraints adopted for the synthesis of the decoder are summarized in Table 3.8.
The actual digital synthesis is a complex and iterative process. A first technology-independent synthesis is performed to identify main logic functionalities in the design, generating a structural description of the design hierarchy in terms of generic digital components such as basic logic gates, arithmetic blocks or other common sequential digital circuits. Then, a gate-level synthesis maps the design into actual logic primitives provided by the chosen target fabrication technology. Further iterations are then required in order to optimize timing, area and power according to input design constraints. Furthermore, design for testability (DFT) features (e.g. scan-chain insertion) or low power features (e.g. clock gating) can be automatically added to the design if required. At each step of the flow the synthesis tool can generate detailed textual reports for various quantities and design objects (timing paths, area, power, inferred cells, ports, nets etc.) in order to validate and debug synthesis results. The synthesized design can be then exported as a gate-level Verilog netlist.

The automatically generated schematic of the column/row decoder is shown in Figure 3.48. Most important results for the synthesized digital block are presented in Table 3.9 instead. As expected, only basic CMOS logic gates AND, OR and inverters, as well as complex gates AND-OR-INVERT (AOI) and OR-AND-INVERT (OAI) from the 65 nm STD cells library have been used by the synthesis tool to implement the pure combinational functionality of the decoder, whereas no latches or FlipsFlops have been inferred. Note that the value of the total area reported by the synthesizer only represents a significant underestimate of the actual total area of the final decoder layout, without considering floorplan, power distribution, extra buffering, tie/filler cells insertion and routing performed in the subsequent place-and-route flow.
One notable aspect of the generated gate-level netlist is the presence of three residual Verilog `assign` statements for thermometer outputs `row[15]`, `col[0]` and `col[16]` tied to constant logic values 0/1, as expected. From a simple hardware point of view, it means that these outputs are not driven by any logic gates, resulting de facto into three unconstrained outputs in terms of drive strength. As a matter of fact, residual `assign` statements in a synthesized netlist are always undesirable and can lead to issues when the design is parsed by a PNR tool for physical layout implementation. These statements must be therefore fixed by hand with the insertion of additional buffers. Alternatively, the synthesizer or the PNR tool can be instructed to automatically perform such a substitution. This demonstrates how despite the simplicity of the implemented digital block, the designer has to carefully understand all steps and results of the automated flow.
The logic functionality of the synthesized design has been then validated with a post-synthesis HDL simulation. Using the same testbench module, the initial behavioural description of the DAC decoder has been replaced with the gate-level netlist, attaching the simulation to Verilog models provided by the foundry for STD cells in 65 nm.

Parameter	Value
Input pins drive resistance	270 Ω
Output pins load capacitance	0.5 pF
Maximum transition time	500 ps
Maximum input-to-output delay	10 ns

Table 3.8: Design constraints specified for the synthesis of the DAC decoder.

Parameter	Value
Total number of inferred gates	48
Sequential instances	0
Combinational instances	48
Cell area	90.720 μm^2
Net area	142.402 μm^2
Total area	233.122 μm^2
Inverters	16 (19% cell area)
Logic	32 (81% cell area)
Maximum fanout	9 (col[8])
Minimum fanout	1 (col[6])
Average fanout	2.2
Leakage power	32.816 nW
Dynamic power	5992.067 nW
Total power consumption	6024.883 nW

Table 3.9: Synthesis results.

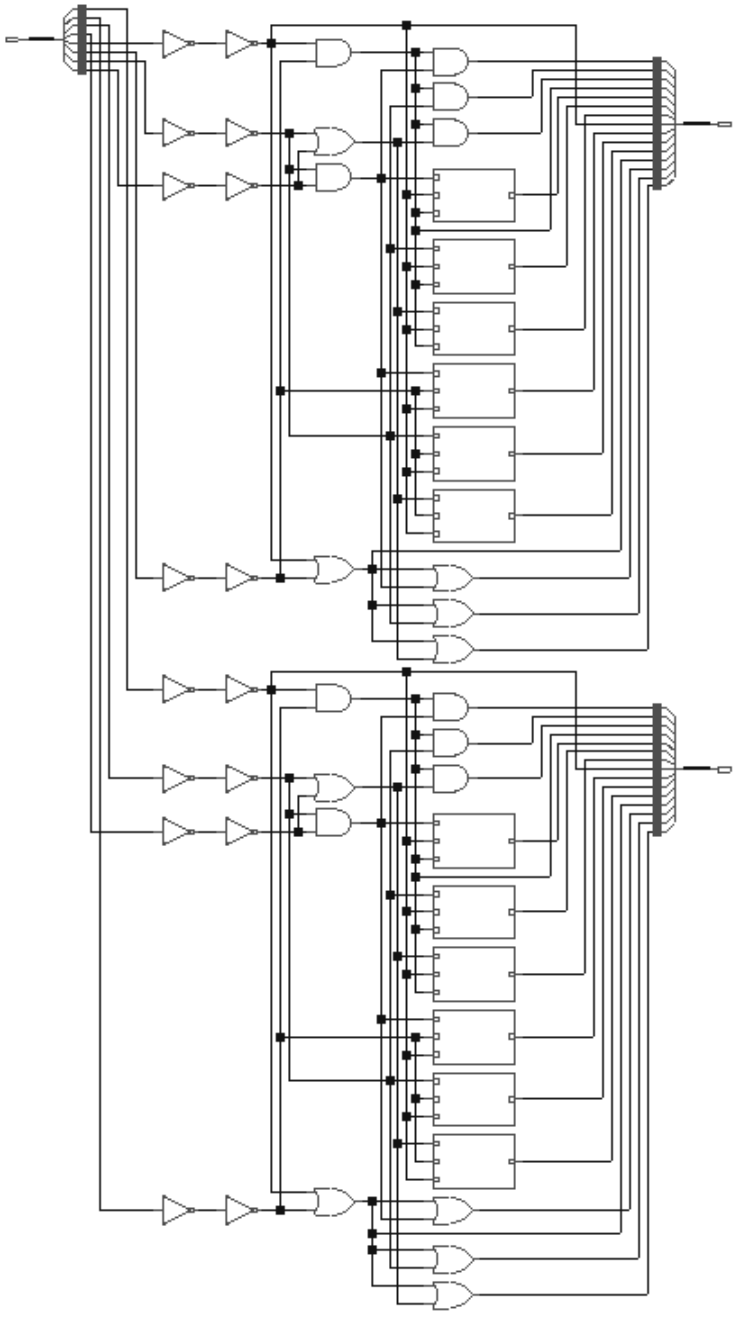


Figure 3.48: Automatically generated schematic for the binary-to-thermometer DAC decoder. Box cells are combinational AOI/OAI gates. Note how buffers have been inserted for all binary inputs.

Automated place-and-route (PNR)

At the end of synthesis process, a Verilog HDL gate-level netlist with instances of digital gates provided by the 65 nm logic library is generated. Moreover, post-synthesis timing and electrical constraints are provided by the synthesizer. As already performed for the synthesis flow, all scripts to automatize the PNR steps have been written from scratch with necessary references to 65 nm STD cell libraries and proper settings suggested for such a technology node. Most significant intermediate results in the generation of the DAC decoder layout are presented in Figure 3.49.
The automated physical implementation flow begins by importing the synthesized netlist and all available technology-related information into the PNR tool. On the one hand, a description of the physical layout and of I/O pins and power/ground rails of STD cells is mandatory. These information is provided by the foundry in form of plain-text physical libraries using an industry-standard layout description language. A detailed description of the actual content of each cell is unnecessary for an automated place-and-route flow. Hence full layout views are not required and an abstract representation of STD cells is sufficient. Physical libraries also contain a limited subset of all process design rules necessary to ensure a DRC-clean design at the end of the flow. On the other hand, STD cells timing information, RC-delay data for the metal interconnections as well as synthesis output constraints are used by the tool for optimized timing-driven placement and routing. In order to take into account PVT variations, worst-case models have been adopted. Other preferences can be specified at this stage. As a notable example, the tool has been instructed to drive all residual Verilog `assign` statements in the synthesized netlist with buffers.
Once the design is loaded into the PNR tool, a top-level placement floorplan and power distribution floorplan must be specified. The maximum available area for the decoder has been constrained to be 58 μm $\times$ 18 μm [Loddo 2014]. Hence the external layout boundary has been fixed to satisfy such a requirement. STD cells have been arranged in two abutted rows with vertical mirroring, defining a core area of 46 μm $\times$ 3.6 μm. Power and ground have been distributed using two concentric metal rings around the core area, with metal stripes of 2.5 μm width and spaced 0.5 μm. The first routing metal has been employed to generate horizontal stripes, the second one for vertical stripes. Power routing has been then automatically performed such that STD cell power/ground rails have been connected to the power distribution network after proper logical mapping between STD cell power/ground pin names and top-level power/ground net names.
The exact position and metal layer of digital binary inputs and column/row thermometer outputs have been specified using an I/O assignment script according to block integration requirements. After that, STD cells have been automatically placed with a timing-driven placement. Furthermore, tie-cells were used to provide 0/1 logic constants, avoiding direct gate connections to power/ground nets. Empty spaces have been filled using dedicated filler-cells provided by the 65 nm digital library. Finally, metal interconnections have been generated with a fully-automated routing engine using the first 3 metal layers offered by the 65 nm CMOS process. Moreover, via optimization has been performed by automatically replacing all single-vias with double-vias in a second routing iteration, as usually suggested by DFM rules to improve the manufacturing yield.
The correctness of the resulting layout has been validated with DRC capabilities included in the PNR tool, verifying basic geometric rules (minimum width, spacing etc.), antenna rules, metal density rules, signals connectivity (presence of opens, shorts or floating nets) and power density. As discussed, due to its pure combinational nature, timing does not represent a tight constraint in the decoder design. Nevertheless, basic timing verification has been performed as well, and interconnection delays have been back-annotated for a gate-level post-layout HDL simulation using the same initial testbench module. The final layout of the digital block has been then exported to the full-custom design environment, as described in the next section.

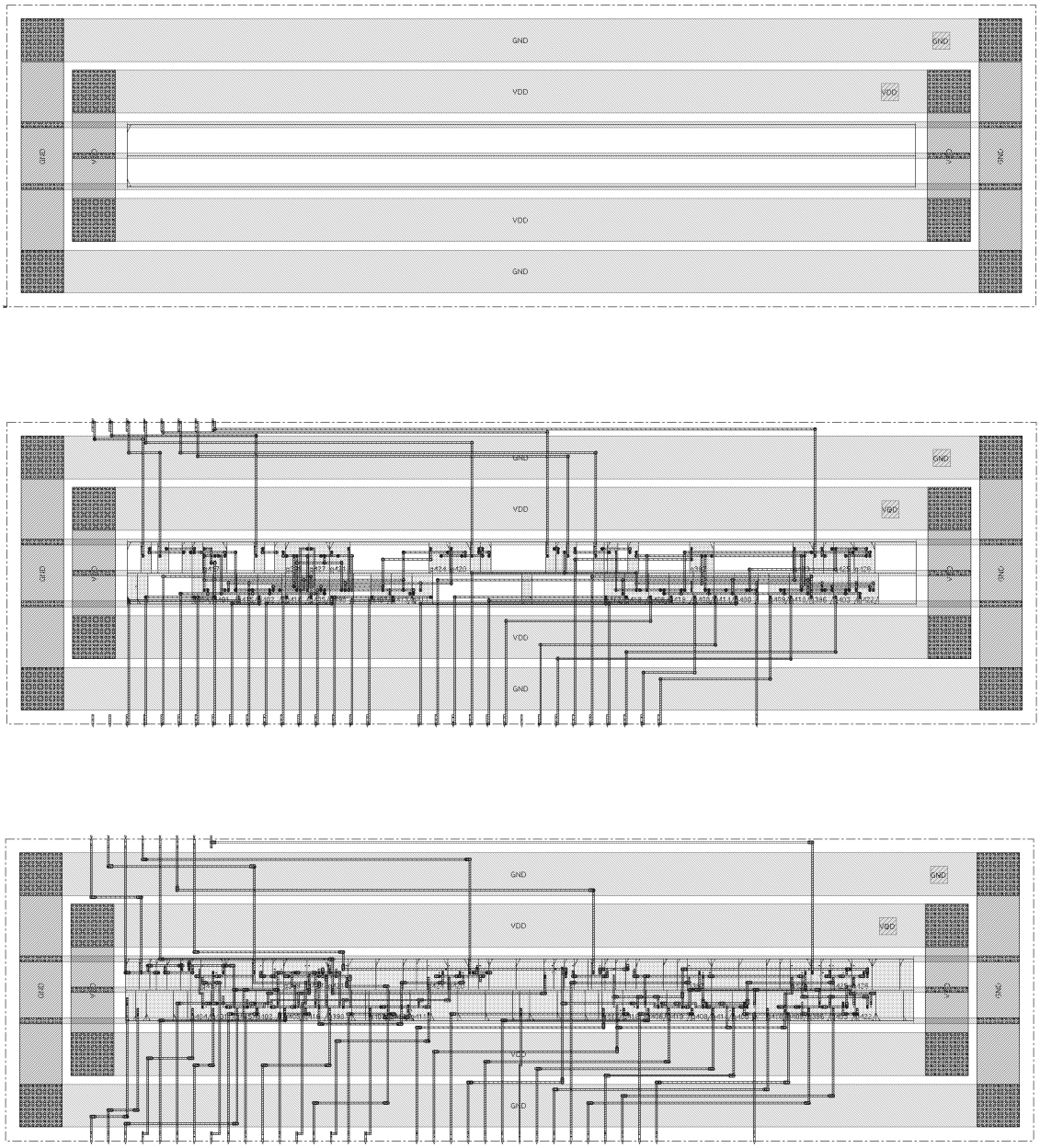


Figure 3.49: Fully automated place-and-route flow into the digital implementation environment. From top to bottom: design floorplan and power distribution; STD cells placement, I/O pins placement and preliminary trial routing; tie-cells and filler-cells placement, routing and final routing optimization.

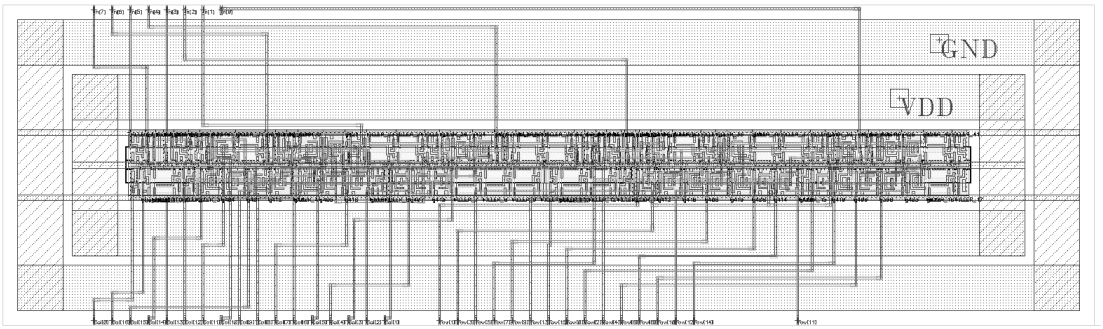


Figure 3.50: Final decoder layout exported to the full-custom design environment for next chip assembling, 58 μm $\times$ 15.6 μm. Binary inputs are placed at the top-left of the block, whereas column/row thermometer outputs are on the bottom side.

Design export/import and final verification

According to a traditional Analogue-on-Top (AoT) approach, at the end of the automated PNR flow the physical layout of the digital block must be exported to the full-custom design framework for subsequent integration with other analogue and mixed-signal blocks. This is accomplished by translating layout information into one of the available industry-standard binary databases used for the representation of the mask layouts of integrated circuits. Alternatively, a plain-text description of the block can be automatically generated by exporting the design into an ASCII file that uses the same description language employed to represent STD cell abstracts. As already mentioned, all aspects and options of the design export/import procedure must be carefully understood in order to guarantee perfect equivalence between the two designs and prevent any data corruption (missing cells or pins, different layers appearance etc.) during the exchange process.
The layout of the column/row decoder ready for chip integration is presented in Figure 3.50. Beside the exported layout, a physical Verilog gate-level netlist is provided by the PNR tool in order to store both signals and power/ground connectivity. Then a schematic view can be automatically generated into the full-custom design environment in order to validate the design with a LVS cross check and transistor-level simulations if required. As stressed, due to an initial unavailability of schematic and layout views of STD cells in the chosen 65 nm CMOS process, this task has been accomplished before the submission only when a tapeout design kit has been available.
The final layout of the current DAC including the implemented decoder is presented in Figure 3.51. Proper functionality of the digital block coupled to the 8-bit thermometer sub-DAC has been also validated with transistor-level simulations. Figure 3.52 shows instead the complete chip submitted as part of the CHIPIX_BIAS prototypes. As mentioned, the chip contains two independent data converters with different versions of column/row decoders. The top DAC uses a decoder designed with full-custom logic cells and has been implemented without the support of automated placement and routing engines. As discussed in Chapter 2, these cells are much larger with respect to those provided by the 65 nm digital library in order to satisfy DFM rules and transistor sizing guidelines proposed by the RD53 collaboration upon preliminary radiation-hardness qualification results. The bottom DAC uses instead the decoder implemented using STD cells in order to explore radiation damage effects on logic gates provided by 65 nm technology. Since these cells target to maximize the integration density, they do not comply with DFM rules.

Figure 3.51: Complete 10-bit segmented current DAC layout, 240 μm $\times$ 140 μm [Loddo 2014]. Local bias and 8-bit thermometer sub-DAC (1). Global bias (2). 2-bit binary sub-DAC (3). 8-bit column/row binary-to-thermometer decoder (4).

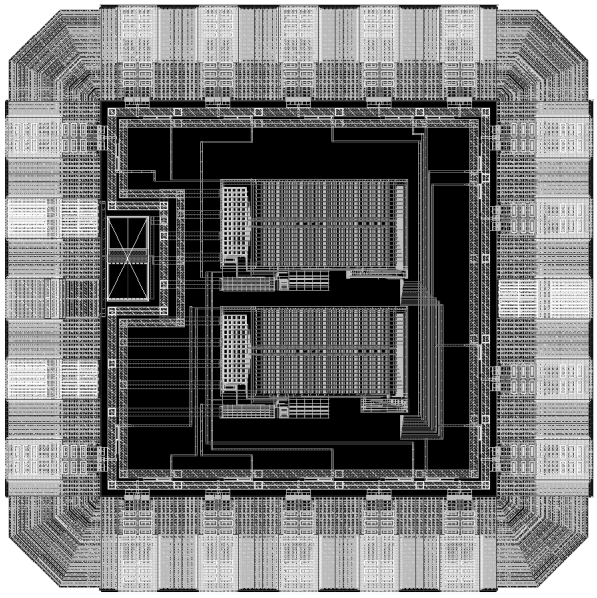


Figure 3.52: Final DAC chip layout, 730 μm $\times$ 730 μm [Loddo 2014]. The chip includes two different versions of column/row decoders: full-custom, DFM rules compliant (top) and STD-cell based, non DFM rules compliant (bottom).

3.9 Summary

First prototypes of pixel Front-End electronics and IP blocks in 65 nm CMOS technology have been submitted by the CHIPIX65 collaboration in October 2014. The submission included three different silicon dies of area 1.96 mm $\times$ 1.96 mm. CHIPIX_BIAS and CHIPIX_SRAM have been dedicated to IP blocks. CHIPIX_VFE1 contains three independent ASICs equipped with different solutions of pixel Front-End electronics. Most important aspects in the assembling of CHIPIX_VFE1/TO and CHIPIX_VFE1/2x1 chips have been discussed.
Additional personal contributions to the submission were in IP block design, with a standard-cell based implementation of a binary-to-thermometer decoder for a current-steering DAC. The design of this block validated for the first time in Torino the automated digital implementation flow attached to the 65 nm CMOS technology up to the foundry transfer.

Chapter 4

Experimental setup and measurements

First prototypes submitted in 65 nm CMOS technology by the CHIPIX65 collaboration have been received back from the foundry at the beginning of 2015. The last chapter of this thesis with personal contributions is dedicated to a description of the the experimental setup commissioned in Torino and to preliminary results derived from bench characterizations at the oscilloscope of the CHIPIX_VFE1/TO chip.

Keywords: Test board, footprint, wire-bonding, PCB, instrumentation, oscilloscope, LSA, post-layout simulation, threshold scan, s-curve, threshold dispersion, ENC

4.1 Introduction

After fabrication, detailed experimental characterizations at the bench completes the ASIC design flow. Test measurements are essential in order to validate circuit performance and to identify most critical points in a first prototyping phase. Furthermore, comparisons between test results and post-layout simulations are fundamental in determining reliability and accuracy of CAD simulation models provided by the manufacturer with respect to the actual physical implementation on silicon. The chapter gives a brief description of the custom test board designed to support first bench characterizations of the CHIPIX_VFE1/TO chip and of the available instrumentation. Preliminary results are then presented. As already discussed, the chip has been equipped with a full-analogue readout of a small pixel array of 8×8 pixels implementing the proposed synchronous Front-End architecture. Hence a complete characterization of the prototype can be accomplished by means of simple measurements at the oscilloscope, whereas proper digital signals to trigger synchronous operations have been generated with a Logic State Analyzer (LSA). Available results include both performance measurements for the analogue Front-End in terms of gain, TOT linearity and noise as well as a validation of the overall synchronous approach. Comparisons with post-layout simulation results are discussed as well. At the time of writing, no measurements are available instead for the larger pixel matrix equipped with digital serial readout and configuration. The chip requires in fact a more complex data acquisition (DAQ) infrastructure currently under development.

4.2 CHIPIX_VFE1/TO test board design

A custom printed circuit board (PCB) was designed to support all CHIPIX_VFE1/TO bench characterizations. It provides mechanical sustain, wire-bonding landing pads, power, bias circuits, configuration bits and necessary input/output access points. The overall concept and design of the test board has been performed with personal contributions in collaboration with the Torino INFN Electronics Laboratory staff [Rotondo 2014]. Detailed schematics are reported on following pages. The final layout of the board is shown in Figure 4.1.

The test board has been implemented as a 4-layers PCB, with a copper layer entirely dedicated to a common low-impedance ground plane. After wire-bonding, *GNDA*, *GNDD*, *VSSA* and *VSSD* are therefore shorted together on a common ground plane. The chip requires a supply voltage of 1.2 V with independent analogue and digital power domains. Two external power connectors are used to power the board at a nominal supply voltage of 5 V. Discrete commercial voltage regulators with additional protecting diodes are then used to obtain regulated 1.2 V supply voltages *VDDA* and *VDDD* for the chip. A few commercial operational amplifiers are powered with unregulated 5 V analogue/digital supply voltages from power connectors. Extensive usage of bypass capacitors guarantee necessary decoupling and filtering for all power supply lines. Analogue and digital core voltages can be also directly provided to the chip by means of coaxial LEMO connectors, selectable with jumper headers. This is required to perform a direct measurement of the total current absorbed by the chip, hence of the power dissipation.
As shown in Figure 4.2, a custom footprint has been designed for the chip according to the floorplan of the wire-bonding. The overall CHIPIX_VFE1 silicon die is hosted by a central 79 mils × 79 mils[1] square surface mount copper pad connected to the ground plane. Then, rectangular surface mount pads of 6 mils × 20 mils with 100 mils pitch provide landing pads for bonding wires. The first CHIPIX_VFE1/TO chip wire-bonded on the test board is presented in Figure 4.3.
As already discussed in Chapter 3, necessary bias currents have been implemented with simple external voltage dividers and on-chip diode-connected MOS devices. All bias currents flowing into dedicated pads are 4 times larger than nominal values and can be regulated by means of discrete trimmers. Test points allow a measurement of the voltage drop across a bias resistor with a digital multimeter (DMM) for the compute of the total bias current. LEMO connectors have been used instead to provide DC steady voltages *VREF_KRUM*, *CAL_LEVEL*, *VTH_DISC* and *VBL_DISC*, requiring therefore suitable instrumentation. Nevertheless, in perspective of tests under irradiation, same voltages can be also generated on the test board with trimmers and voltage dividers connected to *VDAA* through selectable jumper headers.
Global configuration bits for the synchronous Front-End *ANAOUT_EN*, *DIGOUT_EN*, *TOT_EN*, *SUBSTRATE_NOISE*, *SEL_C2F*, *SEL_C4F*, *SEL_CIN50F* and *SEL_CIN100F* are provided by an 8-contacts slide switch with Dual In-line Package (DIP). Logic zeros are tied to ground, whereas pull-up resistors connected to *VDDD* are used to provide logic ones.
Digital signals *RESET_DFF_ADDRESS*, *CLK_ADDRESS_2x2* and *CLK_SEL_PIXEL* for the pixel addressing are generated with normally-open push-buttons. In order to avoid mechanical chatter, switch debouncing is provided by a low-pass RC filter coupled to a commercial low-voltage single Schmitt-trigger inverter.

[1] In PCB design, manufacturing dimensions and tolerances are usually specified in a inch-based system, where 1 mils = 0.001 in = 25.40 μm.

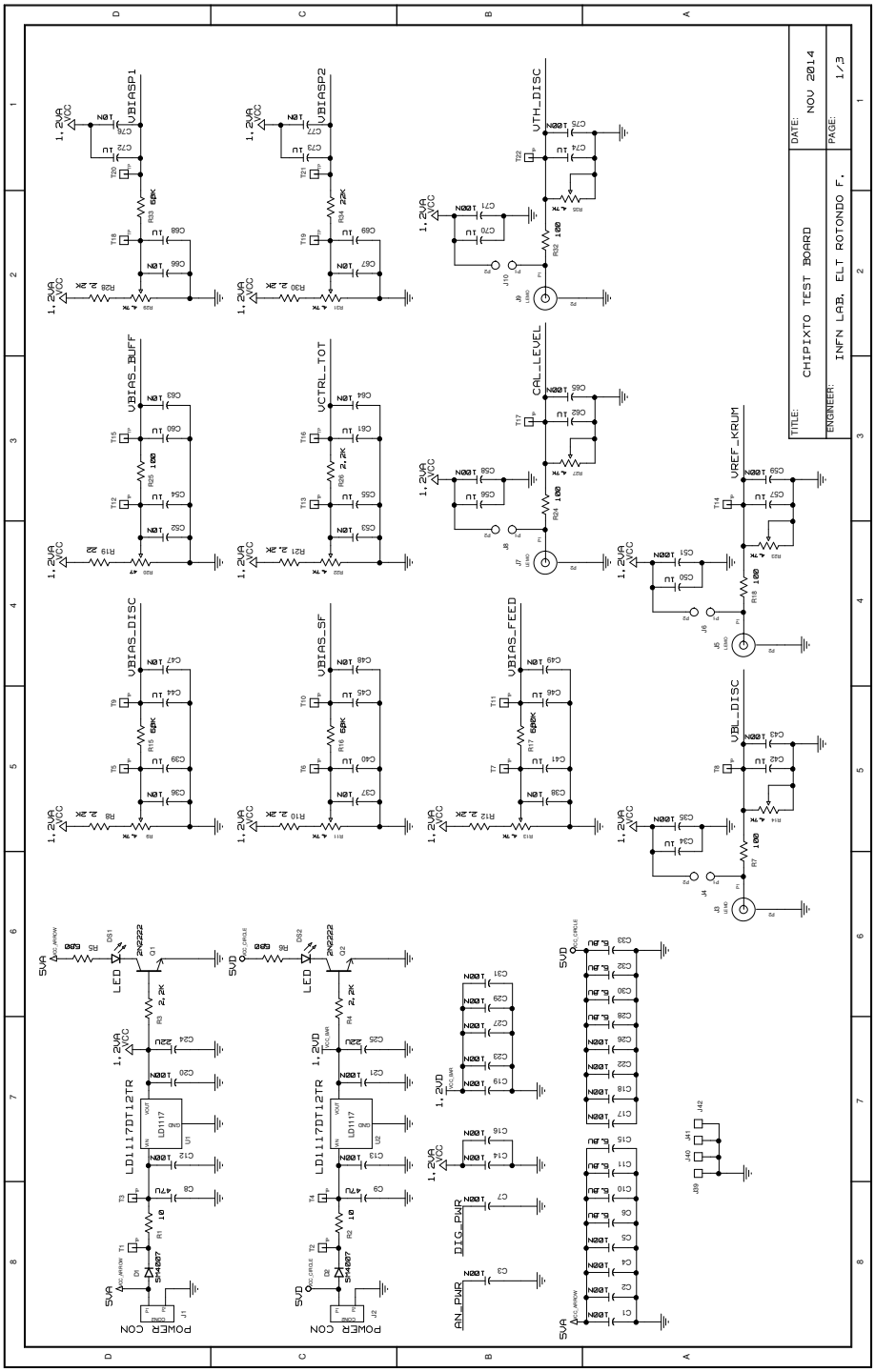
The remaining digital signals *CLK*40, *PHI_CAL*, *RESET_DFF_TOT* and *TESTP* were supposed to be generated by a Logic State Analyzer (LSA). Hence they can be fed to the test board using two dedicated pin headers that perfectly match with the output connector of the available instrument. Alternatively, female Sub-Miniature type A (SMA) RF connectors placed at the edge of the board allow these signals to be provided by simple waveform generators, avoiding the need of a LSA. The selection of the input source is accomplished with jumper headers.
Finally, 4 analogue outputs *ANAOUT* and digital counterparts *DIGOUT* are directly accessible at the oscilloscope using through-hole (TH) access points, pin headers or SMA connectors. High-speed commercial operational amplifiers in non-inverting configuration were therefore used to provide necessary extra buffering and proper impedance matching for RF connectors.
Several test points and pin headers for CHIPIX_VFE1/TO signals and power/ground rails complete the board. For a compact design, all discrete components have been chosen with a surface-mount (SMD) package. A final assembled test PCB is presented in Figure 4.4. The overall dimensions are 11.5 cm × 10.5 cm.

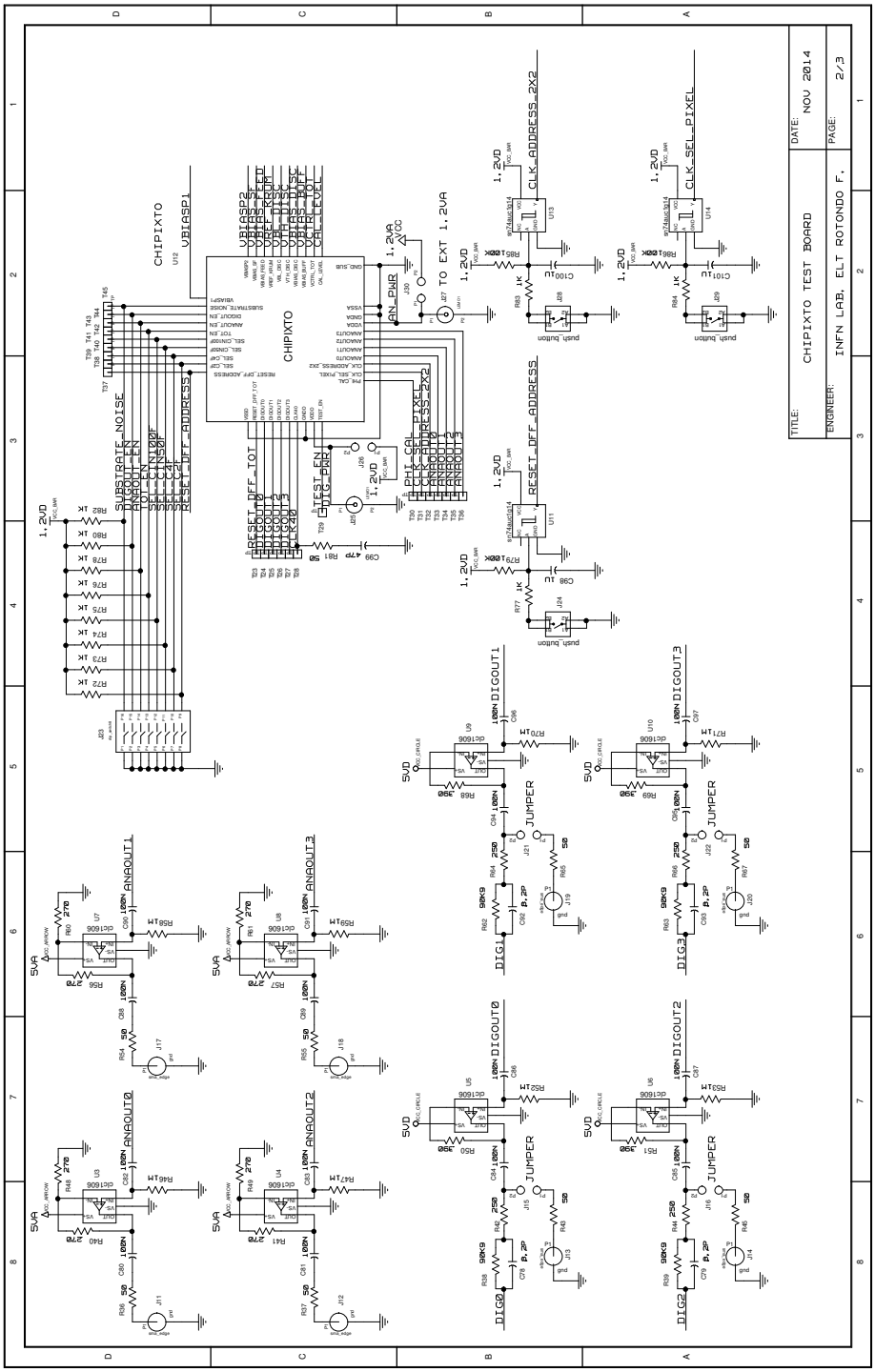
4.3 Test setup and instrumentation

At the time of writing, two CHIPIX_VFE1/TO chips have been wire-bonded onto different test boards equipped with all necessary electronic components. A semi-automated wedge wire-bonding has been performed by the Torino INFN bonding laboratory staff [Pini 2015].

The experimental setup installed since February 2015 is shown in Figure 4.5. The test board is powered at nominal 5 V with standard power supply modules. Additional DC voltages *VTH_DISC*, *VBL_DISC* and *CAL_LEVEL* are provided by dedicated power units and fed to the board using coaxial LEMO cables. The basic equipment is completed by a digital multimeter (DMM) and a 2.5 GHz, 40 MS/s 4-channels digital storage oscilloscope (DSO). Beside more traditional passive probes, a high-performance active probe[2] has been adopted to improve the quality of the results thanks to its low input capacitance. CMOS 1.2 V digital control signals *RESET_DFF_TOT*, *CLK*40, *PHI_CAL*, and *TESTP* are generated by a LSA connected to the test board.
The available instrument offers the possibility of defining arbitrary periodic logic sequences with 500 ps time resolution. Custom test-patterns must be defined in ASCII form using a specific command language. Hence the instrument is also connected to an auxiliary VGA display and to a PS/2 keyboard for user interaction.
The test charge injection is triggered in a selected pixel by *TESTP*, synchronized with a 40 MHz master clock. This requires a proper delay with respect to the main reset signal *RESET_DFF_TOT* for the latch control logic, which in turn defines the overall timing. The timing of the calibration signal *PHI_CAL* can be varied instead in order to study autozeroing performance.
All bench characterizations have been derived from oscilloscope measurements for different Front-End configurations in terms of feedback capacitance, total input capacitance, feedback current, threshold voltage, injected charge and signals timing.

[2] Active probes differ from much simpler passive counterparts because they include a high-impedance, high-frequency amplifier mounted in the probe head. Such an amplifier is used in buffer configuration in order to increase the isolation between the the circuit under test, the cable and the oscilloscope. As a result, the circuit is loaded with a low capacitance (as low as 0.5 pF) and a high DC resistance.


TITLE: CHIPIXTO TEST BOARD
DATE: NOV 2014
ENGINEER: INFN LAB. ELT ROTONDO F.
PAGE: 1/3
VBIASP1
VBIASP2
VTH_DISC
CAL_LEVEL
VBIAS_BUFF
VCTRL_TOT
VREF_KRUM
VBIAS_DISC
VBIAS_SF
VBIAS_FEED
VBL_DISC
LED
POWER CON
LD1117DT12TR
DIG_PWR
AN_PWR
1.2VA
1.2VD
5VA
5VD


TITLE: CHIPIXTO TEST BOARD
DATE: NOV 2014
ENGINEER: INFN LAB, ELT ROTONDO F.
PAGE: 2/3
CHIPIXTO
VBIASP1
VBIASP2
VDRAIN_SF
VDRAIN_KRUM
VRESET_DISC
VDRAIN_DISC
VDRAIN_BUF
CAL_LEVEL
ON_PWR
TO EXT 1.2VA
CLK_ADDRESS_2X2
CLK_SEL_PIXEL
RESET_OFF_ADDRESS
SUBSTRATE_NOISE
ANAOUT_EN
TOT_LOCK_EN
TOT_LOCK_CONF
SEL_CONF
RESET_OFF_ADDRESS
RESET_OFF_TOT
ANAOUT0
ANAOUT1
ANAOUT2
ANAOUT3
CLK_SEL_PIXEL
CLK_ADDRESS_2X2
ANAOUT4
TEST_EN
DIG_EN
DIGOUT0
DIGOUT1
DIGOUT2
DIGOUT3
JUMPER
1.2VD
1.2VA
3.3VD
5VA
push_button

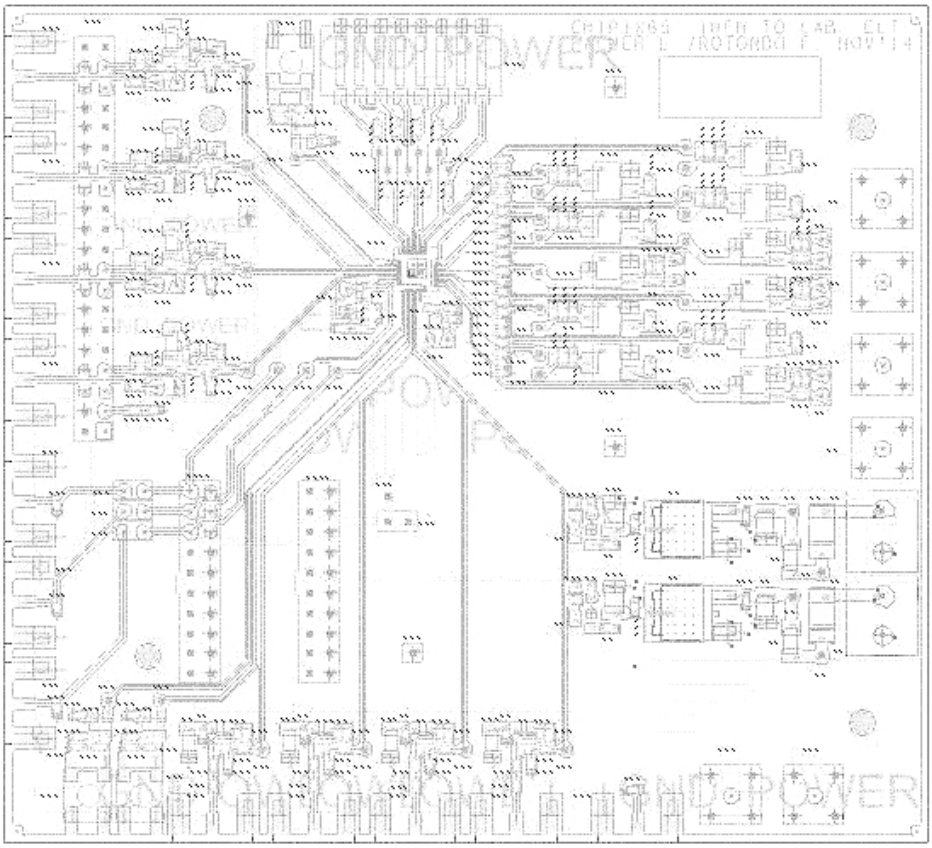


Figure 4.1: CHIPIX_VFE1/TO test board final layout.

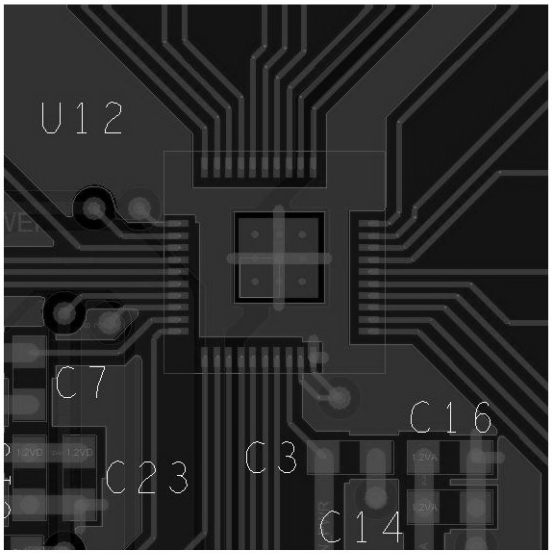


Figure 4.2: CHIPIX_VFE1/TO chip footprint.

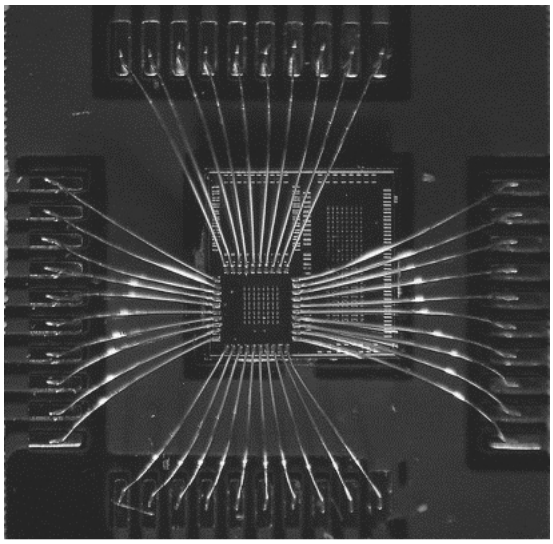


Figure 4.3: The first CHIPIX_VFE1/TO chip wire-bonded to the test board.

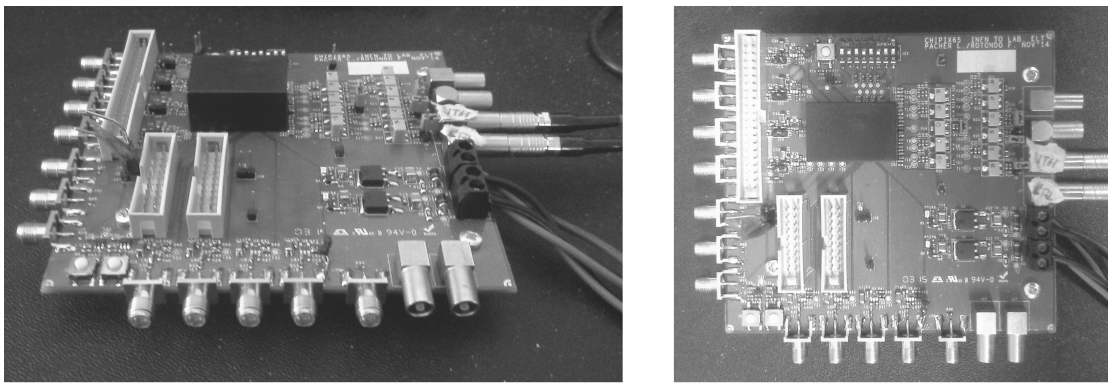


Figure 4.4: CHIPIX_VFE1/TO test board. The chip lies under the protecting box

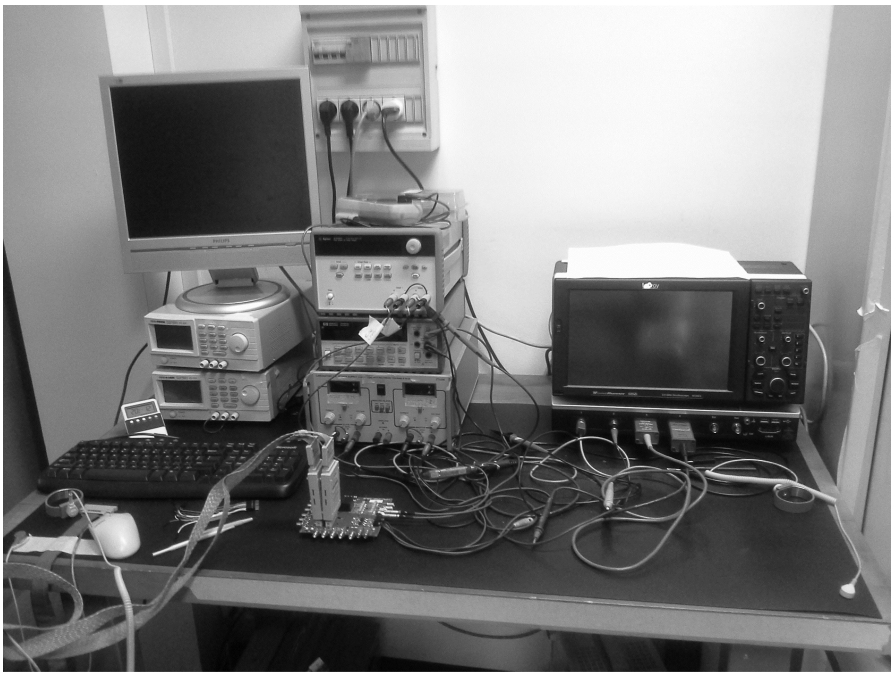


Figure 4.5: CHIPIX_VFE1/TO experimental setup.

4.4 Preliminary results from CHIPIX_VFE1/TO

In the following, preliminary test results derived for the first CHIPIX_VFE1/TO chip wire-bondend on a test board are discussed. They include sample analogue waveforms at the oscilloscope and performance characteristics for the analogue Front-End chain.

Electrical functionality

Proper electrical functionality of synchronous Front-End amplifiers was validated by addressing the test charge injection in different pixels and inspecting the waveforms at the oscilloscope.
As a first step, no 40 MHz clock has been distributed to the pixel matrix. Sample oscilloscope waveforms for 2 ke^- and 10 ke^- injected charges are presented in Figure 4.6. The instrument trigger is provided by the *TESTP* digital signal, generated via LSA. All 8×8 pixels have been electrically stimulated, validating electrical functionality of the serial addressing scheme and of all Front-End amplifiers. All stimulated pixels generated a pulse. Despite the usage of an active probe, the peaking time of the signal is about 30 ns. Indeed, this is in agreement with post-layout transient simulations, that revealed the presence of underestimated parasitic contributions from both internal and top routing metal layers. It must be pointed out that the analogue waveforms seen at the instrument undergo extra-filtering from buffers, metal interconnections towards the chip periphery, pads and cables. These are only test features. As shown in Figure 4.7, post-layout transient simulations for the analogue signal at the CSA output node and assuming a minimum input charge of 1 ke^- reveal that layout parasitics increase the peaking time to about 20 ns. A time response below 25 ns for the minimum signal of interest is therefore guaranteed. At the time of writing, an improved version of the layout of the analogue pixel cell is already under development. Analogue waveforms for different values of input capacitance are presented in Figure 4.8.

Peaking time variations with respect to the nominal value of 12.5 ns do not compromise proper functionality of the synchronous comparator, as discussed in Chapter 2. Provided that a signal is found above the nominal threshold when positive feedback is enabled in the latch, a digital hit pulse is always generated at the output of the discriminator. This can be appreciated in oscilloscope waveforms presented in Figure 4.9 for the same selected pixel, but providing a 40 MHz clock to the test board. A digital pulse is generated by the synchronous comparator in correspondence of the analogue signal. Since the board was configured with $TOT\_EN = 0$, the pulse duration is an integer number of 25 ns clock cycles, as expected. Autozeroing capacitors must be periodically initialized in order to guarantee proper DC operating conditions for the discriminator. The LSA test pattern adopted for *PHI_CAL* foresee an initial pre-calibration cycle of 1 μs followed by a periodic calibration cycle of 25 ns pulse-width every 1 ms. The injection of the test charge is triggered a few μs before a 25 ns calibration pulse.
Proper functionality of the discriminator as a local oscillator can be appreciated in waveforms shown in Figure 4.10. Configuring the board with $TOT\_EN = 1$, the latch in the selected pixel turns into a local clock generator when a signal if found above the threshold. The current in the delay line has been trimmed such that the frequency of the resulting clock is of the order of 100 MHz. This is the maximum frequency observed at the oscilloscope without a large degradation in the quality of the signal. As discussed in Chapter 3, the usage of simple CMOS output pads in CHIPIX_VFE1/TO limits the maximum frequency that can be effectively observed. Furthermore, no special routing has been performed on the test board for *DIGOUT* traces. A mentioned, the usage of differential operations and LVDS or SLVDS pads will be addressed in a second iteration.

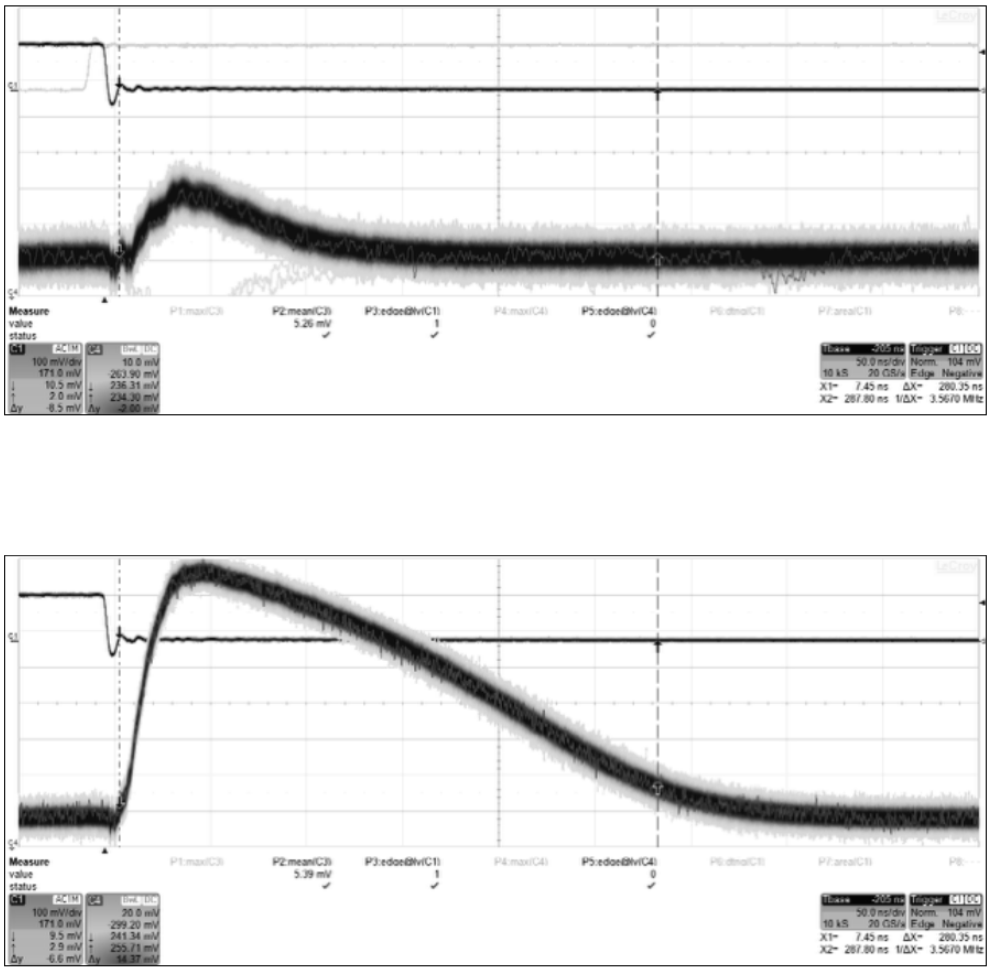


Figure 4.6: Sample analogue waveforms at the oscilloscope for 2 ke^- (top) and 10 ke^- (bottom) injected charges in a selected pixel. The time base is 50 ns/div. The peaking time is about 30 ns. No external 40 MHz clock has been fed to the chip. A 10 nA feedback current has been adopted in order to minimize the ballistic deficit in the Krummenacher feedback, thus maximizing the charge-to-voltage gain (slower discharge of the nominal 4 fF feedback capacitance, $\approx$ 250 ns for 10 ke^- injected charge). The acquisition has been performed without averaging in order to show the noise contribution.

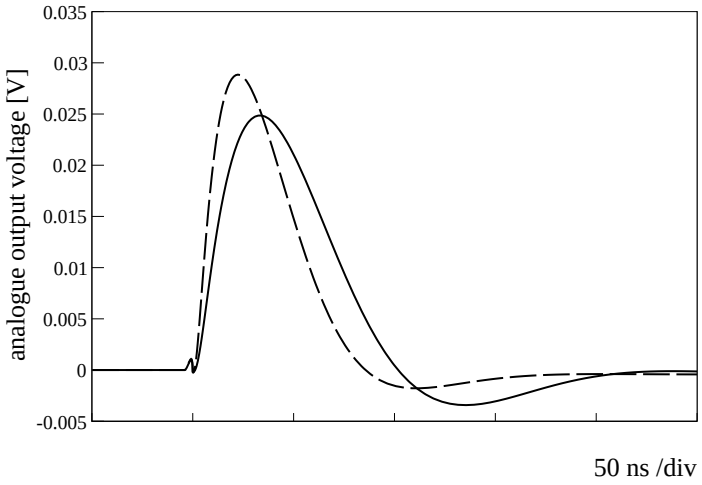


Figure 4.7: Pre-layout (dashed) and post-layout (solid) transient simulations for the analogue pulse at the CSA output node assuming 1 ke^- injected charge and nominal Front-End parameters (4 fF feedback capacitance, 100 fF input capacitance and 40 nA feedback current). The simulation includes extracted parasitics from on-pixel internal routing layers only. After layout the peaking time of the Front-End amplifier increases to about 20 ns. The peaking time further increases up to $\approx$ 30 ns by including the extra filtering of the on-pixel $ANAOUT$ analogue output buffer and parasitic contributions from top metal layers and routing towards the chip periphery. After layout the charge-to-voltage gain decreases from nominal 30 mV/ke^- to 25 mV/ke^-. Only for a better visualization, baseline variations between pre- and post-layout conditions have been cancelled with a series capacitor.

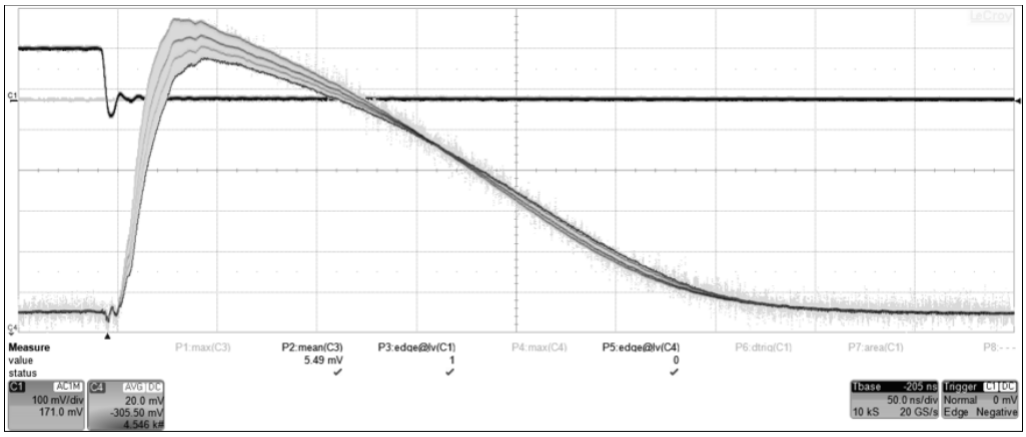


Figure 4.8: Analogue waveforms for different values of selectable input capacitance. Averaged acquisitions with infinite persistence.

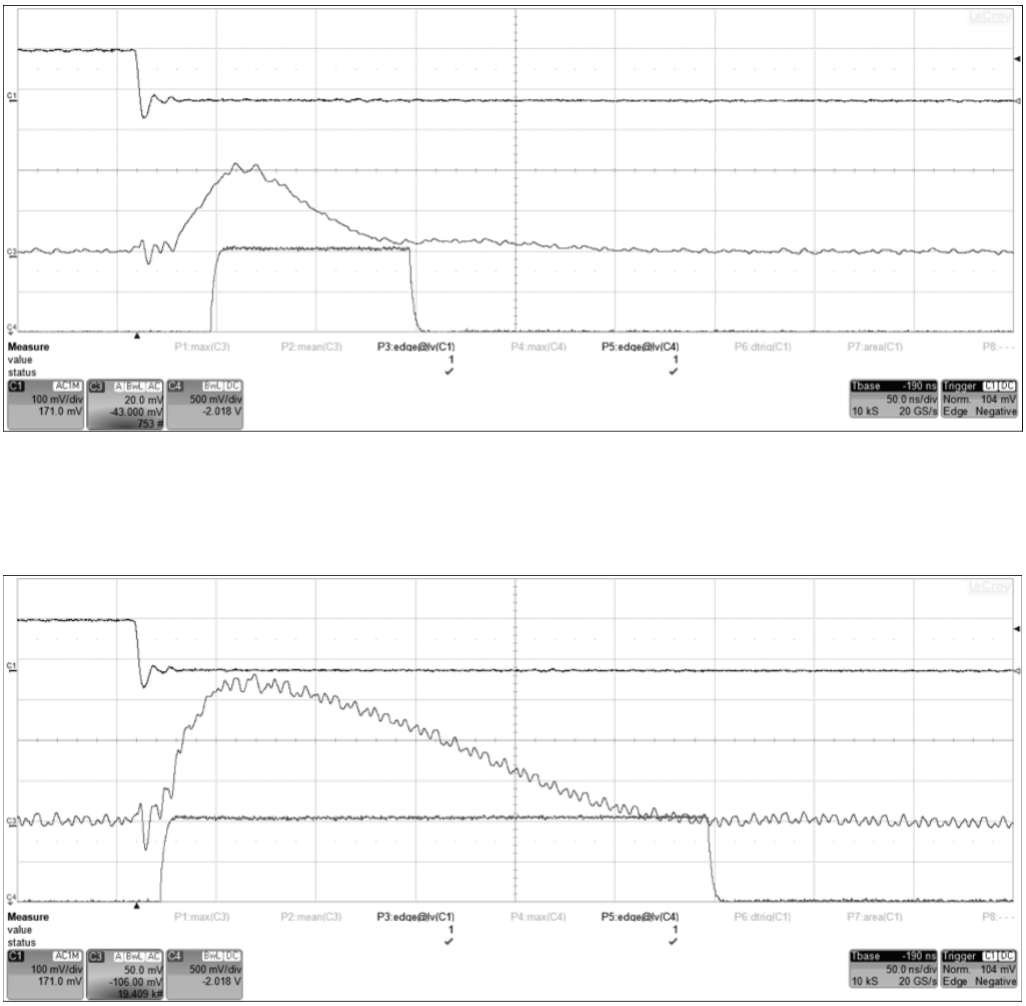


Figure 4.9: Sample analogue waveforms at the oscilloscope for 2 ke^- (top) and 10 ke^- (bottom) injected charges and providing a 40 MHz clock to the test board (not displayed). The threshold voltage *VTH_DISC* is about 12 mV above the baseline. CMOS digital pulses are generated in correspondence of analogue signals. The board has been configured with $TOT\_EN = 0$, hence the hit duration is an integer number of clock cycles. The time base is 50 ns/div.

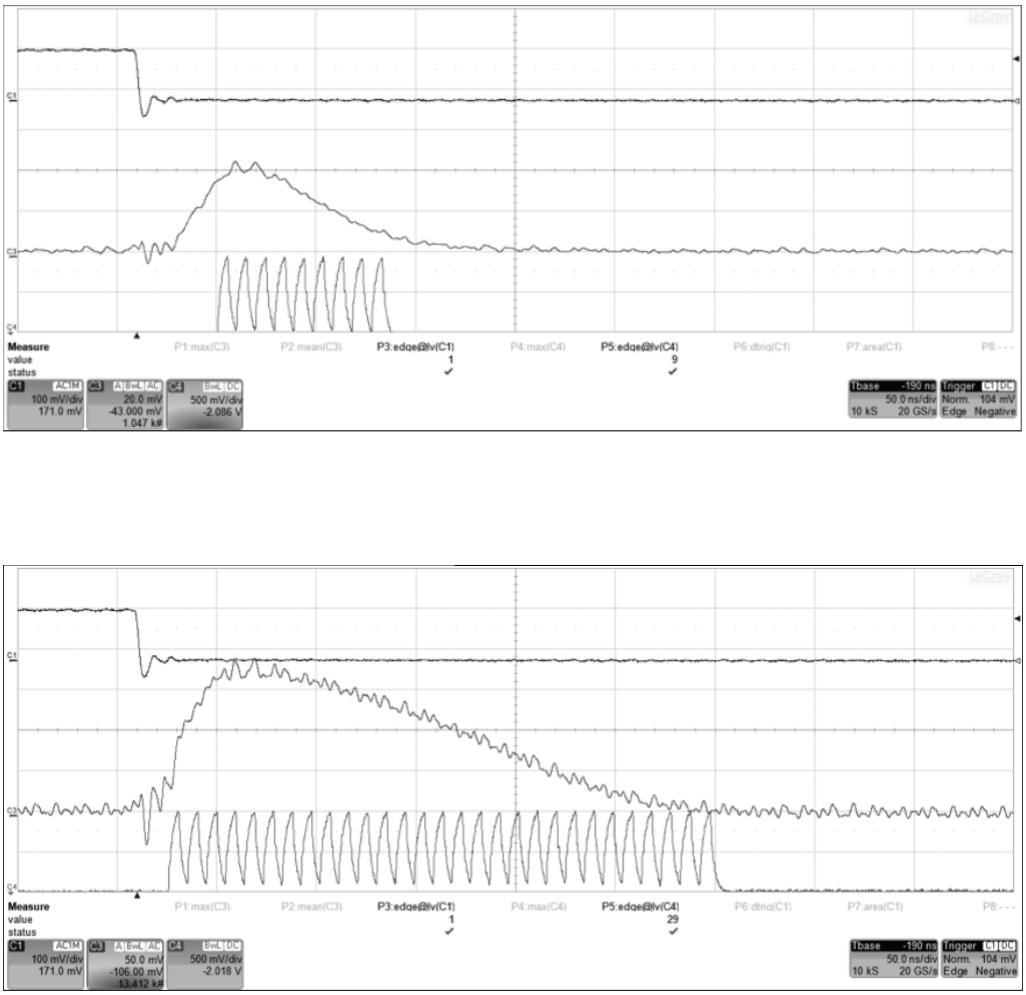


Figure 4.10: Sample analogue waveforms at the oscilloscope for 2 ke^- (top) and 10 ke^- (bottom) injected charges, configuring the board with $TOT\_EN = 1$ in order to turn the latch into a local oscillator. The maximum frequency observed at the oscilloscope without a too large degradation in the quality of the signal is 100 MHz. The time base is 50 ns/div.

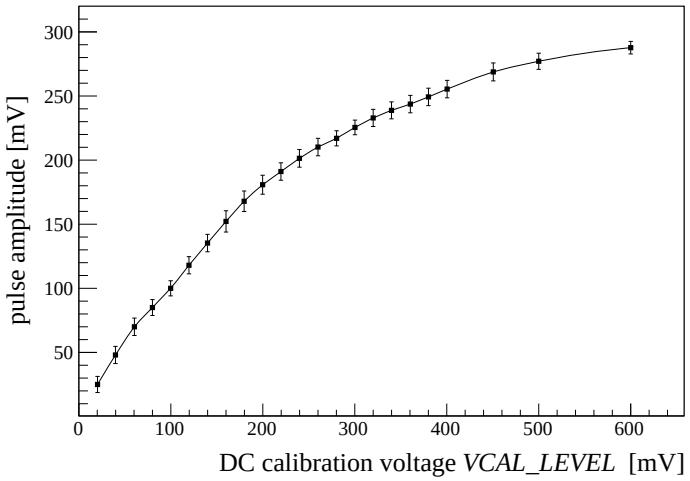


Figure 4.11: Measured pulse amplitude as a function of the DC calibration voltage for a selected pixel (20 mV $\approx$ 1ke^- injected charge). As expected, charge-to-voltage linearity is maintained only up to about 8-10 ke^-. Slow mode operations with 10 nA feedback current and 4 fF feedback capacitance are assumed.

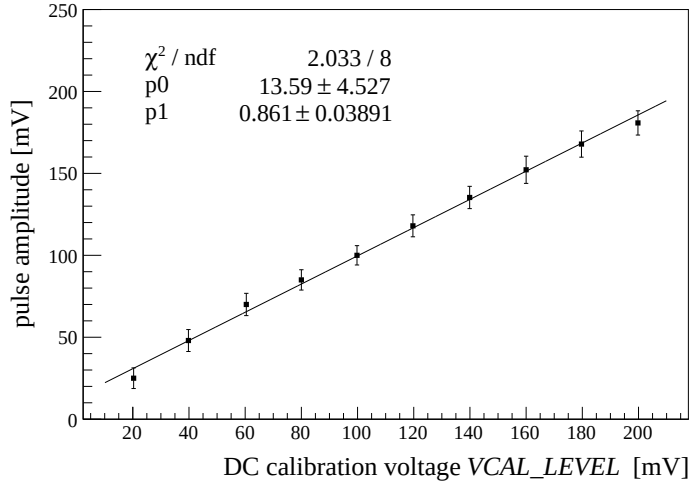


Figure 4.12: Linear fit on measured amplitude data up to 200 mV DC calibration level ($\approx$ 10 ke^-). The charge-to-voltage gain derived from fit parameters is about 17 mV/ke^-.

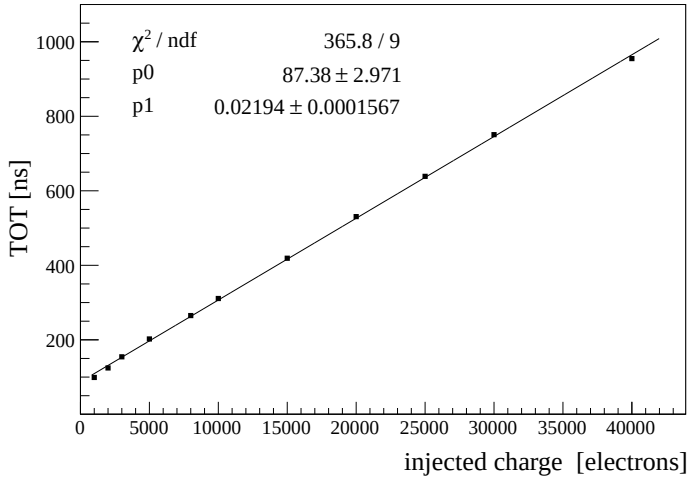


Figure 4.13: Time-over-threshold values measured at the oscilloscope using cursors as a function of the injected charge. As expected, despite saturation in the charge-to-voltage characteristic the TOT linearity is guaranteed up to 40 ke^-.

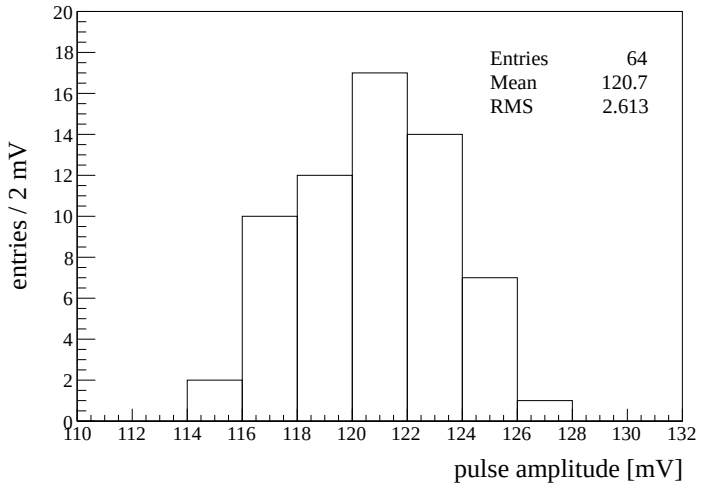


Figure 4.14: Distribution of measured amplitude values for all 64 pixels and 6 ke^- injected charge. The amplitude dispersion is below 2.2% RMS. Pixels exhibit an excellent gain uniformity across the matrix.

Front-End linearity

Simple charge-to-voltage and TOT characteristics have been obtained by varying the amount of charge injected at the Front-End input node and performing amplitude and time measurements with cursors at the oscilloscope. The magnitude of the input charge is determined by the DC voltage *VCAL_LEVEL* provided by a power supply module and fed to the board through a coaxial LEMO cable.

The measured pulse amplitude as a function of the DC calibration voltage *VCAL_LEVEL* for a selected pixel is presented in Figure 4.12, assuming 4 fF feedback capacitance and 10 nA feedback current. As derived from calibration curves and simulations, a voltage step of 20 mV at the CSA input node corresponds to about 1 ke^- injected charge. As expected, charge-to-voltage linearity is guaranteed only for a limited range of input charges, up to about 10 ke^- ($\approx$ 200 mV DC calibration voltage). A linear fit on measured data is shown in Figure 4.12. The charge-to-voltage gain derived from fit parameters is about 17 mV/ke^-. As already mentioned, displayed waveforms undergo extra-filtering and attenuation with respect to actual signals generated at the pixel level, hence the gain is reduced with respect to the simulated post-layout value of 25 mV/ke^-.
For the same pixel and same configuration, Figure 4.13 shows the time-over-threshold measured at the oscilloscope with cursors as a function of the injected charge[3] As expected, despite saturation in the charge-to-voltage characteristic, the TOT is linear with the input charge. Linearity is guaranteed up to 40 ke^-, as already derived in simulations.
Pixel-to-pixel amplitude variations have been measured for all 64 pixels in the matrix and 6 ke^- injected charge. The distribution of measured amplitude values is presented in Figure 4.14. As one can see, pixels exhibit an excellent gain uniformity across the matrix, with amplitude variations below 2.2% RMS. This is in agreement with simulation models provided by the foundry.

Threshold scan and noise

A set of threshold and noise measurements have been performed in order to derive first estimates for the uncompensated pixel-to-pixel threshold dispersion and for the ENC as a function of the selectable input capacitance. At the time of writing, a characterization of the threshold dispersion after autozeroing is still under investigation.

As usually performed in Front-End systems, the effective threshold and the noise level can be determined by means of s-curves. Measurements can be performed either with a threshold scan and a fixed input charge or with a signal scan and a fixed threshold value. Although both procedures are supposed to give same results, a threshold scan is usually preferred [Spieler 2005].
All s-curves presented in the following have been extracted using a fixed charge of 2 ke^- injected at the input node of the pixel under consideration. For a given threshold, an acquisition is performed over 1000 injection cycles generated by the LSA and the number of hit occurrences is recorded at the oscilloscope using the instrument sampling capabilities. The acquisition is then repeated for different threshold values. The plot for the fraction of hit occurrences (hit efficiency) as a function of the threshold represents the s-curve.

[3] In the remaining of the chapter, charge values are reported instead of *VCAL_LEVEL* voltage equivalents.

As shown in Figure 4.15, as the threshold is scanned from low to high, initially all signal pulses are recorded because the signal is always found above the threshold. Conversely, no hit is generated if the threshold significantly exceeds the signal, hence the efficiency drops to zero. As expected, the hit efficiency decreases by increasing the threshold value, but due to electronic noise the transition is broadened. Assuming Gaussian-distributed noise contributions, measured points can be fitted using an error function. The effective threshold is determined as the threshold value (or equivalently, input charge) for which the hit efficiency is 0.5, whereas the noise RMS is determined by the variance. Both values are provided by fit parameters.
A first set of threshold scans has been recorded for 16 different pixels[4], assuming 2 ke^- fixed injected charge, 100 fF input capacitance, 4 fF feedback capacitance and 100 nA feedback current. With such a feedback current a 10 ke^- signal returns to the baseline in about 90 ns (fast mode). Indeed, due to ballistic deficit in the Krummenacher scheme this reduces the charge-to-voltage gain, hence it is expected a larger ENC value. The resulting s-curves are presented in Figure 4.16. The distribution of threshold values extracted from fit parameters is presented in Figure 4.17.
As already mentioned, all measurements have been performed by triggering the test charge injection about 1 ms after a 25 ns calibration cycle for the offset compensation. After a so long amount of time the offset stored on autozeroing capacitors de facto can be assumed lost. Hence the distribution gives a preliminary estimate of pixel-to-pixel threshold variations without offset compensation. According to measured data the threshold distribution is about 272 e^- RMS. Ongoing test activities are focused on determining autozeroing performance .
The histogram of variances extracted from s-curves is presented in Figure 4.18. The mean value of the distribution corresponds to the ENC. Measurements indicate that ENC $\approx$ 170 e^- RMS at 100 fF input capacitance and 100 nA feedback current. Both threshold and noise distributions refer to 16 pixels only, hence are affected by limited statistics. At the time of writing, no post-layout simulated values are available for data comparison.
A preliminary study of the ENC as a function of the input capacitance for a single pixel is shown in Figure 4.19. ENC values have been derived from 3 threshold scans performed at different values of input capacitance. In order to minimize the ballistic deficit in the Krummenacher feedback a low current of 10 nA has been adopted, thus maximizing the charge-to-voltage gain. Each point represents the variance extracted from a fit with an error function on the corresponding s-curve. Measured data show that ENC $\approx$ 106 e^- RMS at 100 fF input capacitance.

4.5 Summary

Tests results obtained from CHIPIX_VFE1/TO synchronous pixels are encouraging. All basic electrical functionalities have been validated at the oscilloscope. Each pixel in the matrix generates a pulse if stimulated. Synchronous discriminator operations have been observed in both normal and fast TOT modes. In particular, latch operations as a local oscillator have been proved to work. Performance characteristics in terms of linearity and pixel-to-pixel gain variations are in good agreement with simulations. Preliminary results for the untrimmed threshold dispersion and for the system noise have been presented. Ongoing test activities are focused on determining actual autozeroing performance and collecting data for all 64 pixels in the matrix in order to increase the statistic significance of results.

[4] One pixel for each 2 $\times$ 2 pixel region

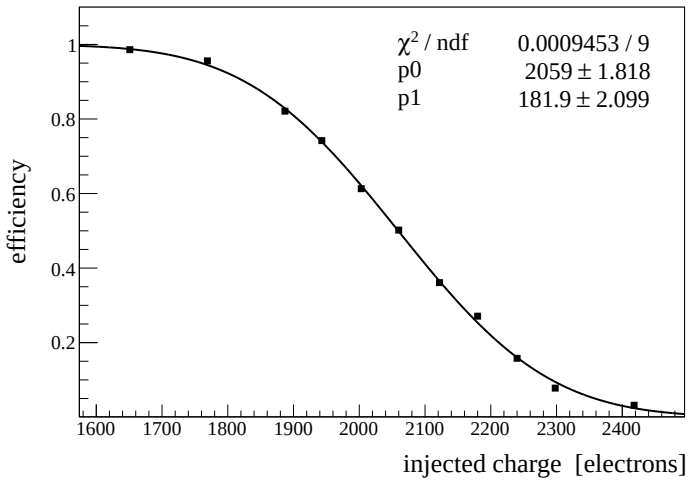


Figure 4.15: Determination of the effective threshold and of the noise level by means of threshold scan and s-curve. The input charge has been fixed to 2 ke^-. Measured data are fitted with an error function. Fit parameters provide the effective threshold (mean) and the electronic noise contribution (variance).

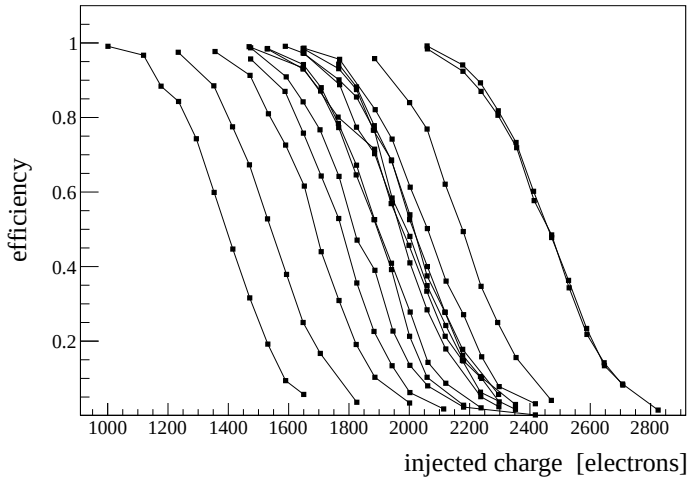


Figure 4.16: Measured s-curves for 16 different pixels assuming 2 ke^- injected charge, 100 fF input capacitance, 4 fF feedback capacitance and 100 nA feedback current.

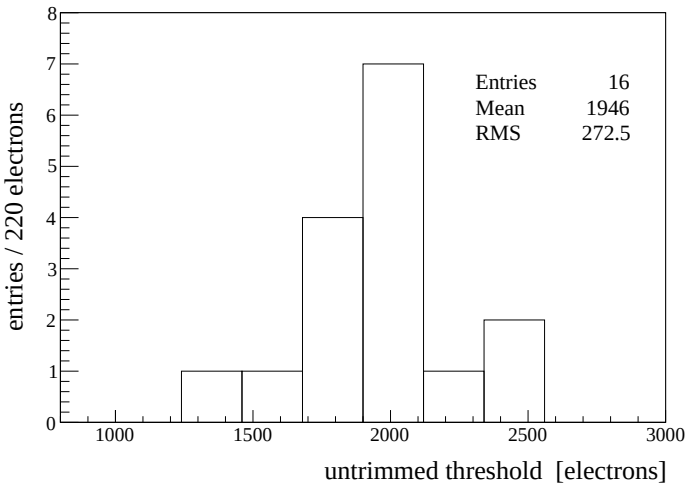


Figure 4.17: Distribution of effective threshold values extracted from s-curves. The charge injection is triggered about 1 ms after an autozeroing calibration pulse. The offset stored on autozeroing capacitors can be therefore assumed lost. Hence the distribution gives a preliminary estimate for pixel-to-pixel threshold variations without offset compensation. The distribution indicates that uncompensated threshold variations are of about 220 e^- RMS. Measurements refer to 16 pixels only, hence are affected by limited statistics.

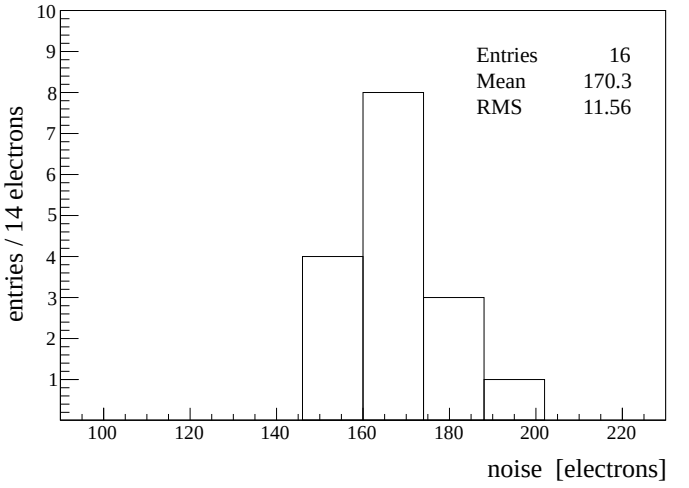


Figure 4.18: Distribution of noise values (variances) extracted from s-curves. The histogram suggests that ENC $\approx$ 170 e^- RMS at 100 fF input capacitance and 100 nA feedback current.

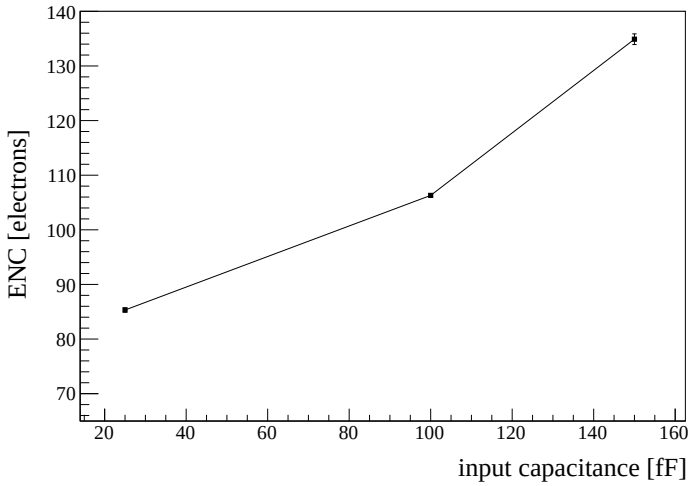


Figure 4.19: Measured ENC as a function of the input capacitance for a selected pixel assuming 4 fF feedback capacitance and 10 nA feedback current (slow mode). Each point represents the variance extracted from a fit with an error function on the corresponding s-curve. ENC $\approx$ 106 e^- RMS at 100 fF input capacitance.

Conclusions

The foreseen High-Luminosity LHC upgrade will require the installation of a new silicon pixel detector for the CMS experiment. A development plan devoted to the design of a new pixel ASIC has started. The usage of a modern CMOS fabrication technology represents the only means to address the challenges introduced by reduced pixel size and unprecedented data rates and radiation levels expected at HL-LHC, 10-times higher than nominal values. A commercial 65 nm CMOS technology has been identified by the pixel ASIC community as a promising candidate for the implementation of the new readout chip. Research activities on 65 nm CMOS are now part of the cross-experiment ATLAS/CMS/LCD international collaboration RD53 and of the Italian INFN/GR5 CHIPIX65 project.

Within these frameworks, the overall Ph.D. research activity has been focused on the development of analogue and digital pixel Front-End electronics in 65 nm CMOS technology suitable for the CMS pixel upgrade at HL-LHC. The choice of the Front-End architecture was essentially driven by the aim of explore innovative solutions for both the hit discrimination and a charge encoding performed at the pixel level with the TOT technique, taking advantage of increased speeds offered by a 65 nm technology node.

A discrete-time solution for the hit discriminator has been proposed in this work. The usage of a track-and-latch voltage comparator introduces fundamental advantages. The hit generation is synchronized with a 40 MHz master clock, sampling the CSA analogue output around its peaking time. Thanks to the usage of positive feedback very low signals above the nominal threshold can be efficiently discriminated. A time response of 12.5 ns is always guaranteed, thus providing a reliable solution that naturally overcomes time-walk issues in the time-stamp assignment. Threshold variations have been compensated by means of capacitors using an autozeroed scheme, without the need of a local on-pixel D/A converter for digital trimming. Flexible and high-speed TOT charge digitizations can be retrieved by turning the latch into a local oscillator using asynchronous logic. Such a technique is extensively adopted in modern commercial SAR A/D converters.

The basic layout of the chosen pixel cell architecture was used to assemble two small pixel matrices put on silicon as part of the first CHIPIX65 submission. Additional personal contributions to the submission have been in IP block design, validating with a column/row binary-to-thermometer decoder the fully-automated digital implementation flow attached to the 65 nm CMOS technology.

Preliminary experimental results derived for synchronous pixels in the CHIPIX_VFE1/TO chip are encouraging. Proper functionality of the implemented synchronous architecture has been validated. In particular, latch operations as a local oscillator have been proved to work. Measured performance characteristics are in good agreement with circuit simulation models provided by the foundry, assessing the reliability of the chosen fabrication technology.

Design and characterization of first Front-End prototypes in 65 nm CMOS provided the necessary foundation and know-how towards the design of a future complete ASIC demonstrator suitable for the long-term CMS pixel upgrade.

Glossary

ADC	Analogue-to-Digital Converter
ASIC	Application-Specific Integrated Circuit
ATLAS	A Toroidal LHC ApparatuS
BGA	Ball Grid Array
BW	Bandwidth
CAD	Computer-Aided Design
CFD	Constant Fraction Discriminator
CMOS	Complementary Metal-Oxide-Semiconductor
CMS	Compact Muon Solenoid
CSA	Charge Sensitive Amplifier
DAC	Digital-to-Analogue Converter
DAQ	Data Acquisition
dB	Decibel
DFM	Design For Manufacturability
DFT	Design For Testability
DIP	Dual In-line Package
DMM	Digital Multi-Meter
DRC	Design Rule Check
DSM	Deep Sub-Micron
DSO	Digital Storage Oscilloscope
EDA	Electronic Design Automation
ELT	Enclosed Layout Transistor
ENC	Equivalent Noise Charge
EOC	End-Of-Column
ESD	Electrostatic Discharge
FEE	Front-End Electronics
FIFO	First-In First-Out
FOM	Figure Of Merit
FPGA	Field Programmable Gate Array
FSM	Finite State Machine

GBW	Gain-Bandwidth product
HDL	Hardware Description Language
HEP	High Energy Physics
HV	High Voltage
IC	Integrated Circuit / Inversion Coefficient
IOS	Input Offset Storage
IP	Intellectual Property
KCL	Kirchoff Current Law
LDO	Low Drop-Out voltage regulator
LET	Linear Energy Transfer
LHC	Large Hadron Collider
LHCC	Large Hadron Collider Committe
LP	Low-Power
LSA	Logic State Analyzer
LSB	Least Significant Bit
LVDS	Low-Voltage Differential Signaling
LVL	Layout Versus Layout
LVS	Layout Versus Schematic
MC	Monte-Carlo
MI	Moderate Inversion
MIM	Metal-Insulator-Metal capacitor
MIP	Minimum Ionizing Particle
MOM	Metal-Oxide-Metal capacitor
MOS	Metal-Oxide-Semiconductor transistor
MPV	Most Probable Value
MPW	Multi-Project Wafer
MS	Mixed-Signal
MSB	Most Significant Bit
NM	Noise Margin
OOS	Output Offset Storage
OTA	Operational Transconductance Amplifier

PCB	Printed Circuit Board
PDF	Probability Density Function
PLL	Phase-Locked Loop
PM	Phase Margin
PMT	Photo-Multiplier Tube
PNR	Place-and-Route
PR	Pixel Region
PSD	Power Spectrum Density
PSRR	Power Supply Rejection Ratio
PU	Pile-Up
PUC	Pixel Unit Cell
PVT	Process, Voltage and Temperature
RAM	Random Access Memory
RD	Research and Development
RF	Radio Frequency
RHBD	Radiation-Hardened By Design
RMS	Root Mean Square
ROC	Read-Out Chip
RTL	Register Transfer Level
SAR	Successive Approximation Register
SCE	Short Channel Effects
SEE	Single-Event Effect
SEU	Single-Event Upset
SI	Strong Inversion
SLVDS	Scalable Low-Voltage Differential Signaling
SMA	Sub-Miniature type A
SMD	Surface Mount Device
SNR	Signal-to-Noise Ratio
SPICE	Simulation Program with Integrated Circuit Emphasis
SR	Slew Rate
SRAM	Static Random Access Memory

TDC	Time-to-Digital Converter
TOT	Time-Over-Threshold
TH	Through-Hole
TID	Total Ionizing Dose
UDSM	Ultra-Deep Sub-Micron
VCO	Voltage-Controlled Oscillator
VLSI	Very Large Scale Integration
VTC	Voltage Transfer Characteristics
WI	Weak Inversion

References

[Allen 2002] P. E. Allen and D. R. Holberg. *CMOS Analog Circuit Design.* Oxford Press, 2002

[Amsler 2008] C. Amsler et al. (Particle Data Group) *Plots of cross sections and related quantities.* Physics Letters B, 2008

[Anelli 1999] G. Anelli et al. *Radiation Tolerant VLSI Circuits in Standard Deep Submicron CMOS Technologies for the LHC Experiments: Practical Design Aspects.* IEEE Transactions on Nuclear Science, 1999

[Athavale 2005] A. Athavale and C. Christensen. *High-Speed Serial I/O Made Simple. A Designers' Guide with FPGA Applications.* Xilinx Press, 2005

[Baker 2012] R. J. Baker. *CMOS Circuit Design, Layout and Simulation.* Wiley/IEEE Press, 2010

[Ballabriga 2006] R. Ballabriga et al. *The Medipix3 Prototype, a Pixel Readout Chip Working in Single Photon Counting Mode with Improved Spectrometric Performance.* IEEE Nuclear Science Symposium Conference, 2006

[Barbero 2004] M. Barbero et al. *Design and test of the CMS pixel readout chip.* Nuclear Instruments and Methods A, 2004

[Bartz 2005] E. Bartz. *The 0.25 μm token bit manager chip for the CMS pixel readout.* Proceedings of the 11th Workshop on Electronics for LHC and Future Experiments, 2005

[Bichsel 1988] H. Bichsel. *Straggling in thin silicon sensors.* Reviews of Modern Physics, 1988

[Bonacini 2011] S. Bonacini et al. *Characterization of a commercial 65 nm CMOS technology for SLHC applications.* JINST, 2011

[Bonacini 2014] S. Bonacini. CERN PH/ESE, Private communication, 2014

[Brianti 1997] F. Brianti, A. Manstretta and G. Torelli. *High-speed autozeroed CMOS comparator for multistep A/D conversion.* Microelectronics Journal, 1997

[Calin 1996] T. Calin, M. Nicolaidis and R. Velazco. *Upset Hardened memory design for submicron CMOS technologies.* IEEE Transactions on Nuclear Science, 1996

[Campbell 2001] M. Campbell. *Electronics for pixel detectors.* CERN/LHCC-2001-005, Proceeding of the LEB2001 Workshop on Electronics for LHC Experiments, 2001

[Ceresa 2014] D. Ceresa et al. *Macro Pixel ASIC (MPA): the readout ASIC for the pixel-strip (PS) module of the CMS outer tracker at the HL-LHC.* JINST, Proceedings of the Workshop on Intelligent Trackers, 2014

[Chatrchyan 2008] S. Chatrchyan et al. (The CMS Collaboration) *The CMS experiment at the CERN LHC.* JINST, 2008

[Chatrchyan 2012] S. Chatrchyan et al. (The CMS Collaboration) *CMS technical design report for the pixel detector upgrade.* CERN/LHCC-2012-016, CMS-TDR-011

[Chatrchyan 2014] S. Chatrchyan et al. (The CMS Collaboration) *Description and performance of track and primary-vertex reconstruction with the CMS tracker.* CSM/TRK-11-001

[Cucciarelli 2006] S. Cucciarelli, M. Konecki, D. Kotlinski and T. Todorov. *Track reconstruction, primary vertex finding and seed generation with the CMS Pixel Detector.* CMS/NOTE-2006-026

[Christiansen 2013] J. Christiansen and M. Garcia-Sciveres. *RD Collaboration Proposal: Development of Pixel Readout Integrated Circuits for Extreme Rate and Radiation.* CERN/LHCC-2013-008, 2013

[Calvo 2008] D. Calvo et al. *A silicon pixel readout ASIC in CMOS 0.13 μm for the PANDA microvertex detector.* IEEE Nuclear Science Symposium Conference Record, 2008

[Chang 1999] K. C. Chang. *Digital Systems Design with VHDL and Synthesis.* IEEE Computer Society Press, 1999

[Chu 2006] P. P. Chu. *RTL Hardware Design Using VHDL.* Wiley, 2006

[Conti 2015] E. Conti. *Design of Dedicated Electronic Systems for the Readout of Pixel Radiation Sensors.* PhD Thesis, University of Perugia, 2015

[De Gaspari 2014] M. De Gaspari et al. *Design of the analog front-end for the Timepix3 and Smallpix hybrid pixel detectors in 130nm CMOS technology.* JINST Proceedings of the Topical Workshop on Electronics for Particle Physics, 2014

[Degerli 2003] Y. Degerli et al. *Low-Power Autozeroed High-Speed Comparator for the Readout Chain of a CMOS Monolithic Active Pixel Sensor Based Vertex Detector.* IEEE Transactions on Nuclear Science, 2003

[De Geronimo 2005] G. De Geronimo and P. O'Connor . *Mosfet optimization in deep submicron technologies for charge amplifiers.* IEEE Transactions on Nuclear Science, 2005

[Delagnes 2000] E. Delagnes et al. *SFE16, a low noise front-end integrated circuit dedicated to the read-out of large Micromegas detectors.* IEEE Transactions on Nuclear Science, 2000

[Delaurenti 2006] P. Delaurenti. *Analysis and Design of a Fast Binary Front-End Chip for the COMPASS Experiment at CERN.* Master thesis, University of Torino, 2006

[Demaria 2013] N. Demaria, INFN Torino. *CHIPIX65: CALL Project Proposal, INFN/CSN5*, 2013

[Demaria 2014] N. Demaria, INFN Torino. CHIPIX65 project internal design collaboration, 2014

[De Plassche 2003] R. van De Plassche. *CMOS Integrated Analog-to-Digital and Digital-to-Analog Converters.* Kluwer, 2003

[Di Pietro 2012] V. Di Pietro. *A Low Power Front-End Amplifier for the Microstrip Sensors of the PANDA Microvertex Detector.* Master thesis, University of Torino, 2012

[Enz 1995] C. C. Enz, F. Krummenacher and E. A. Vittoz. *An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications.* Analog Integrated Circuit Signal Processing, 1995

[Enz 2006] C. C. Enz and E. A. Vittoz. *Charge-based MOS Transistor Modeling. The EKV model for low-power and RF IC design.* Wiley, 2006

[Erdmann 2005] W. Erdmann *The 0.25 μm front-end for the CMS pixel detector.* Nuclear Instruments and Methods A, 2005

[Evans 2008] L. Evans and P. Bryant. *LHC machine.* JINST, 2008

[Figueiredo 2004] P. M. Figueiredo and J C. Vital. *Low kickback noise techniques for CMOS latched comparators.* IEEE Proceedings of the International Symposium on Circuits and Systems, 2004

[Figueiredo 2006] P. M. Figueiredo and J C. Vital. *Kickback noise reduction techniques for CMOS latched comparators.* IEEE Transactions on Circuits and Systems II, 2006

[Figueiredo 2009] P. M. Figuerido and J. C. Vital. *Offset Reduction Techniques in High-Speed Analog-to-Digital Converters.* Springer, 2009

[Fisher 2001] P. Fisher. *Pixel Electronics for the ATLAS Experiment.* Nuclear Instruments and Methods A, 2001

[Fu 2014] Y. Fu et al. *The charge pump PLL clock generator designed for the 1.56 ns bin size time-to-digital converter pixel array of the Timepix3 readout ASIC.* JINST Proceeding of the TWEPP-13 Topical Workshop on Electronic for Particle Physics, 2014

[Gabathuler 2005] K. Gabathuler. *PSI46 Pixel Chip External Specifications.* Paul Scherrer Institute official documentation, 2005

[Gaioni 2011] L. Gaioni, M. manghisoni, L. Ratti, V. Re and G. Traversi. *Front-end electronics in a 65 nm CMOS process for high density readout of pixel sensors.* Nuclear Instruments and Methods A, 2011

[Gaioni 2014] L. Gaioni, INFN Pavia. CHIPIX65 project internal design collaboration, 2014

[Gaioni 2015] L. Gaioni et al. *Low-power clock distribution circuits for the Macro Pixel ASIC.* JINST Proceeding of the TWEPP-14 Topical Workshop on Electronics for Particle Physics, 2015

[Garcia-Sciveres 2011] M. Garcia-Sciveres et al. *The FE-I4 pixel readout integrated circuit.* Nuclear Instruments and Methods A, 2011

[Garcia-Sciveres 2013] M. Garcia-Sciveres et al. *Towards third generation pixel readout chips.* Nuclear Instruments and Methods A, 2013

[Goll 2015] B. Goll and H. Zimmermann. *Comparators in Nanometer CMOS Technology.* Springer, 2015

[Gonella 2007] L. Gonella et al. *Total Ionizing Dose effects in 130-nm commercial CMOS technologies for HEP experiments.* Nuclear Instruments and Methods A, 2007

[Graupner 2006] A. Graupner. *A Methodology for the Offset-Simulation of Comparators.* The Designer's Guide Community, application note, 2006

[Gray 2001] P. R. Gray, P. J. Hurst, S. H. Lewis and R. G. Meyer. *Analysis and Design of Analogue Integrated Circuits.* Wiley, 2001

[Gray 2013] J. A. Gray. *The CMS phase-1 pixel detector.* JINST Proceeding of the 13th Topical Seminar on Innovative Particle and Radiation Detectors, 2013

[Gregorian 1999] R. Gregorian. *Introduction to CMOS OP-AMPs and Comparators.* Wiley, 1999

[Gromov 2010] V. Gromov et al. *Gossipo-3: A prototype of a Front-End Pixel Chip for Read-Out of Micro-Pattern Gas Detectors.* JINST Proceeding of the TWEPP-09 Topical Workshop on Electronics for Particle Physics, 2010

[Hartmann 2009] F. Hartmann. *Evolution of Silicon Sensor Technology in Particle Physics.* Springer, 2009

[Hastings 2001] A. Hastings. *The Art of Analog Layout.* Prentice Hall, 2001

[Havranek 2014] M. Havranek et al. *Pixel front-end development in 65 nm CMOS technology.* JINST Proceeding of the TWEPP-13 Topical Workshop on Electronics for Particle Physics, 2014

[Huang 2013] Y. Huang, H. Schleifer, and D. Killat. *Design and analysis of novel dynamic latched comparator with reduced kickback noise for high-speed ADCs.* IEEE European Conference on Circuit Theory and Design (ECCTD), 2013

[He 2001] Z. He. *Review of the ShockleyRamo theorem and its application in semiconductor gamma-ray detectors.* Nuclear Instruments and Methods A, 2001

[Hemperek 2009] T. Hemperek et al. *Digital Architecture of the New ATLAS Pixel Chip FE-I4.* IEEE Nuclear Science Symposium Conference Record, 2009

[Horisberger 2001] R. Horisberger. *Readout architectures for Pixel detectors.* Nuclear Instruments and methods A, 2001

[Iadarola 2014] G. Iadarola et al. *Analysis of the electron cloud observations with 25 ns bunch spacing at the LHC.* Proceedings of the IPAC2014 International Particle Accelerator Conference, 2014

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

[Khandpur 2005] R. Khandpur. *Printed Circuits Boards: Design, Fabrication and Assembling.* McGraw-Hill, 2005

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

[Karagounis 2011] M. A. Karagounis. *Analog Integrated CMOS Circuits for the Readout and Powering of Highly Segmented Detectors in Particle Physics Applications.* PhD thesis, University of Bonn, 2011

[Kasinski 2012] K. Kasinski. *Multichannel Integrated Circuits for Silicon Strip Detectors Readout with Timestamping and Amplitude Pulse Measurement.* PhD thesis, AGH University of Krakow , 2012

[Kastli 2006] H. C. Kastli et al. *Design and performance of the CMS pixel detector readout chip.* Nuclear Instruments and Methods A, 2006

[Kastli 2013] H. C. Kastli. *Frontend electronics development for the CMS pixel detector upgrade.* Nuclear Instruments and Methods A, 2013

[Kipnis 1997] I. Kipnis et al. *A time-over-threshold machine: the readout integrated circuit for the BABAR Silicon Vertex Tracker.* IEEE Transactions on Nuclear Science, 1997

[Kirkpatrick 1989] J. M. Kirkpatrick. *Electronic Drafting and Printed Circuit Board Design.* Cengage Learning, 1989

[Kobayashi 1993] T. Kobayashi, K. Nogami, T. Shirotori, and Y. Fujimoto. *A current-controlled latch sense amplifier and a static power-saving input buffer for low-power architecture.* IEEE JSSC, 1993.

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

[Kugathasan 2011] T. Kugathasan. *Low-Power High Dynamic Range Front-End Electronics for the Hybrid Pixel Detectors of the PANDA MVD.* PhD thesis, University of Torino, 2011

[Laker 1994] K. R. and W. M. C. Sansen. *Design of Analog Integrated Circuits and Systems.* Mc-Graw Hill, 1994

[Lee 2011] H. H. K. Lee. *Circuit and Layout Techniques for Soft-Error-Resilient Digital CMOS Circuts.* PhD Thesis, Stanford University, 2011

[Leo 1994] W. R. Leo. *Techniques for Nuclear and Particle Physics Experiments.* Springer, 1994

[Liu 2010] C. C. Liu, S. J. Chang, G. Y. Huang and Y. Z. Lin. *A 10-bit 50 MS/s SAR ADC with a Monotonic Capacitor Switching Procedure.* IEEE JSSC, 2010

[Llopart 2002] X. Llopart et al. *Medipix2,a 64k pixel readout chip with 55 μm square elements working in single photon counting mode.* IEEE Transactions on Nuclear Science, 2002

[Llopart 2007] X. Llopart et al. *TimePix, a 65k programmable pixel readout chip for arrival time, energy and/or photon counting measurements.* Nuclear Instruments and Methods A, 2007

[Loddo 2014] F. Loddo, INFN Bari. CHIPIX65 project internal design collaboration, 2014

[Mahapatra 2002] N. R. Mahapatra, A. Tareen and S. V. Garimella. *Comparison and analysis of delay elements.* Proceedings of the 45th Midwest Symposium on Circuits and Systems, 2002

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

[Manghisoni 2007] M. Manghisoni, L. Ratti, V. Re, V. Speziali, and G. Traversi. *Resolution limits in 130 nm and 90 nm CMOS technologies for analog front-end applications.* IEEE Transactions on Nuclear Science, 2007

[Manghisoni 2009] M. Manghisoni, L. Ratti, V. Re, and G. Traversi. *Design optimization of charge preamplifiers with CMOS processes in the 100 nm gate length regime.* IEEE Transactions on Nuclear Science, 2009

[Manghisoni 2011] M. Manghisoni, L. Gaioni, L. Ratti, V. Re and G. Traversi. *Introducing 65 nm CMOS technology in low-noise read-out of semiconductor detectors.* Nuclear Instruments and Methods A, 2011.

[Marconi 2014] S. Marconi, E. Conti, P. Placidi, J. Christiansen and T. Hemperek. *The RD53 Collaboration SystemVerilog-UVM Simulation Framework and its General Applicability to Design of Advanced Pixel Readout Chip.* JINST 9, P10005, 2014

[Martoiu 2006] V. S. Martoiu. *Design, Test and System Integration of Front-End Electronics for Particle Detection in High-Energy Nuclear and Subnuclear Experiments.* PhD thesis, University of Torino, 2006

[Martoiu 2009] S. Martoiu et al. *A pixel front-end ASIC in 0.13 μm CMOS for the NA62 experiment with on pixel 100 ps Time-to-Digital Converter.* IEEE Nuclear Science Symposium Conference Record, 2009

[Mazza 2012] G. Mazza et al. *A CMOS 0.13 μm Silicon Pixel Detector Readout ASIC for the PANDA experiment.* JINST, Proceedings of the TWEPP-11 Topical Workshop on Electronics for Particle Physics, 2012

[Mekkaoui 2001] A. Mekkaoui and J. Hoff. *30Mrad(SiO_2) radiation tolerant pixel front end for the BTEV experiment.* Nuclear Instruments and Methods A, 2001

[Mekkaoui 2013] A. Mekkaoui, M. Garcia-Sciveres and D. Gnani. *Results of 65 nm pixel readout chip demonstrator array.* JINST Proceeding of the TWEPP-12 Topical Workshop on Electronics for particle Physics, 2013

[Mekkaoui 2014] A. Mekkaoui and M. Garcia-Sciveres. RD53 Top-Level Working Group meeting, 2014

[Monteil 2013] E. Monteil. *Front-End amplifiers in 65nm CMOS technology for the upgrade of the pixel detector of the CMS experiment.* Master thesis, University of Torino, 2013

[Monteil 2014] E. Monteil, University of Torino and INFN. CHIPIX65 project internal design collaboration, 2014

[Montrose 2000] M. I. Montrose. *Printed Circuit Board Design Techniques for EMC Compliance.* Wiley-IEEE Press, 1998

[Morsani 2014] F. Morsani, INFN Pisa. CHIPIX65 project internal design collaboration, 2014

[O'Connor 2002] P. O'Connor and G. De Geronimo. *Prospects for charge sensitive amplifiers in scaled CMOS.* Nuclear Instruments and Methods A, 2002

[Parker 1997] S. Parcker, C. Kenney, and J.Sengal. *3D - A proposed new architecture for solid-state radiation detectors.* Nuclear Instruments and methods A, 1997

[Parkes 2003] C. Parkes. *Silicon detectors at the LHC.* Nuclear Instruments and methods A, 2003

[Peric 2004] I. Peric. *Design and Realisation of Integrated Circuits for the Readout of Pixel Sensors in High-Energy Physics and Biomedical Imaging.* PhD thesis, University of Bonn, 2004

[Peric 2006] I. Peric et al. *The FEI3 readout chip for the ATLAS pixel detector.* Nuclear Instruments and Methods A, 2006

[Pini 2014] B. Pini, INFN Torino. Private internal communication, 2014

[Poikela 2014] T. Poikela et al. *Timepix3: A 65k channel hybrid pixel readout chip with simultaneous ToA/ToT and sparse readout.* JINST Proceeding of the 15th International Workshop on Radiation Imaging Detectors, 2014

[Rabaey 2003] J. M. Rabaey, A. Chandrakasan and B. Nikolioc. *Digital Integrated Circuits. A Design Perspective.* Prentice Hall, 2003

[Radulov 2011] G. Radulov, P. Quinn, H. Hegt and A. van Roermund. *Smart and Flexible Digital-to-Analog Converters.* Springer, 2011

[Ratti 2014] L. Ratti, INFN Pavia. CHIPIX65 project internal design collaboration, 2014

[Ravera 2013] F. Ravera. *Characterization and performance of 3D silicon pixel detectors for CMS.* Master thesis, University of Torino, 2013

[Razavi 1992] B. Razavi and B. A. Wooley. *Design Techniques for High-Speed High-Resolution Comparators.* IEEE JSSC, 1992

[Razavi 1995] B. Razavi. *Principles of Data Conversion System Design.* IEEE Press, 1995

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

[Re 2005] V. Re, M. Manghisoni, L. Ratti, V. Speziali, and G. Traversi. *Total ionizing dose effects on the analog performance of a 0.13 μm CMOS technology.* IEEE Radiation Effects Data Workshop, 2005

[Re 2014] V. Re. *Performance Requirements for the Analog Section in the Pixel Readout Chip.* RD53 Analog Working Group internal note, CERN/RD53-NOTE-2014-002, 2014

[Rivetti 2001] A. Rivetti et al. *A Low-Power 10-bit ADC in a 0.25-μm CMOS: Design Considerations and Test Results.* IEEE Transactions on Nuclear Science, 2001

[Rolo 2013] M. D. Rolo et al. *TOFPET ASIC for PET applications.* JINST Proceeding of the IWORID2012 International Workshop on Radiation Imaging Detectors, 2013

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

[Rossi 2012] L. Rossi and O. Bruning. *High Luminosity Large Hadron Collider A description for the European Strategy Preparatory Group.* CERN/ATS-2012-236

[Rotondo 2014] F. Rotondo, INFN Torino. Private design collaboration, 2014

[Rumolo 2012] G. Rumolo et al. *LHC experience with different bunch spacings in 2011 (25, 50 and 75ns).* Proceedings of the 2012 Chamonix Workshop on LHC Performance, 2012

[Sansen 1990] W. M. C. Sansen and Z. Y. Chang. *Limits of low noise performance of detector readout front-ends in CMOS technology.* IEEE Transactions on Circuits and Systems, 1990

[Sansen 2006] W. M. C. Sansen. *Analog Design Essentials.* Springer, 2006

[Saint 2002] C. Saint and J. Saint. *IC Mask Design Essential Layout Techniques.* McGraw-Hill, 2000

[Schneider 2010] M. C. Schneider and C. G. Montoro. *CMOS Analog Design Using All-Region MOSFET Modelling.* Cambridge University Press, 2010

[Sicard 2007] E. Sicard and S. D. Bendhia. *Basics of CMOS Cell Design.* McGraw-Hill Professional, 2007

[Smith 1998] D. J. Smith. *HDL Chip Design. A Practical Guide for Designing, Synthesizing and Simulating ASICs and FPGAs Using VHDL or Verilog.* Doone Pubns, 1998

[Snoeys 2000] W. Snoeys et al. *Layout techniques to enhance the radiation tolerance of standard CMOS technologies demonstrated on a pixel detector readout chip.* Nuclear Instruments and Methods A, 2000

[Snoeys 2001] W. Snoeys et al. *Pixel readout electronics development for the ALICE pixel vertex and LHCb RICH detector.* Nuclear Instruments and Methods A, 2001

[Spieler 2005] H. Spieler. *Semiconductor Detector Systems.* Oxford University Press, 2005

[Stabile 2014] A. Stabile, INFN Milano. CHIPIX65 project internal design collaboration, 2014

[Szczygiel 2010] R. Szczygiel. *Krummenacher feedback analysis for high-count-rate semiconductor pixel detector readout.* IEEE Proceedings of the 17th International Conference on Mixed Design of Integrated Circuits and Systems, 2010

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

[Uyemura 2002] J. P. Uyemura. *CMOS Logic Circuit Design.* Kluwer, 2002

[Valerio 2012] P. Valerio, R. Ballabriga and M. Campbell. *Design of the 65nm CLICpix demonstrator chip.* CERN/LCD-Note-2012-018, 2012

[Valerio 2014] P. Valerio et al. *A prototype hybrid pixel detector ASIC for the CLIC experiment.* JINST, Proceedings of the 2013 Topical Workshop On Electronics Particle Physics, 2014

[Vignali 2015] M. Centis Vignali. *Characterization of thin irradiated epitaxial silicon sensors for the CMS phase II pixel upgrade.* JINST Proceeding of the 7th International Workshop on Semiconductor Pixel Detectors for Particles and Imaging, 2015

[Weste 2011] N. H. Weste and D. M. Harris. *CMOS VLSI Design. A Circuits and Systems Perspective.* Addison-Wesley, 2001

[Weilhammer 2000] P. Weilhammer. *Overview: Silicon Vertex Detectors and Trackers.* Nuclear Instruments and Methods A, 2000

[Yukawa 1985] A. Yukawa. *A CMOS 8-bit High-Speed A/D Converter IC.* IEEE JSSC, 1985