Universität Bonn

Physikalisches Institut

Design and Development of Depleted Monolithic Active Pixel Sensors with Small Collection Electrode for High-Radiation Applications

Konstantinos Moustakas

Depleted monolithic active pixel sensors (DMAPS) have emerged as a low material and low cost alternative to the established hybrid technology. In this thesis, two large scale prototype chips that feature a fast column-drain readout architecture have been developed in order to demonstrate the feasibility of a small collection electrode DMAPS implementation for high rate and high radiation environments such as the HL-LHC ATLAS ITk upgrade. They have been fabricated using a novel modification of the TowerJazz 180 nm process that enhances the charge collection properties and allows for full depletion of the sensitive volume. High radiation tolerance is combined with high analog performance and low power consumption as a result of the low sensor capacitance. The first prototype, called TJ-Monopix1, has demonstrated the proof of concept yielding a low ENC of $10\,e^-$ and up to 97% detection efficiency after irradiation to $10^{15}\,n_{eq}/\mathrm{cm}^2$ NIEL. A full scale successor chip, called TJ-Monopix2, has been designed in order to further improve performance and enable easier system integration. Apart from LHC type applications, the TJ-Monopix DMAPS are ideally suited to other experiments such as lepton colliders and can also benefit the development of imaging devices.

Physikalisches Institut der
Universität Bonn
Nussallee 12
D-53115 Bonn

BONN-IR-2021-01
June 2021
ISSN-0172-8741

Design and Development of Depleted Monolithic Active Pixel Sensors with Small Collection Electrode for High-Radiation Applications

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Konstantinos Moustakas
aus
Larisa, Griechenland

Bonn, April 2021

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter:	Prof. Dr. Norbert Wermes
2. Gutachter:	Prof. Dr. Klaus Desch
Tag der Promotion:	09.09.2021
Erscheinungsjahr:	2021

Contents

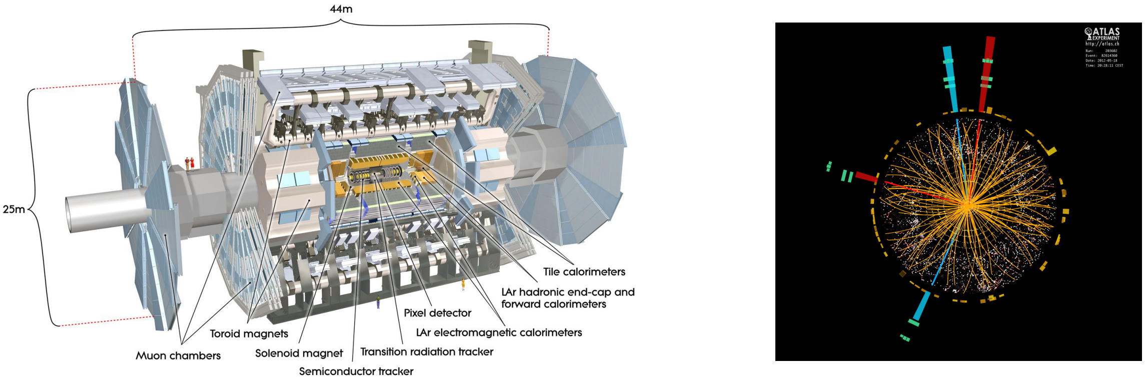
CHAPTER 1

Introduction

Understanding the nature of reality can be considered the ultimate goal of human intellect. The idea that the universe is composed of fundamental blocks was first proposed by Democritus. It remained only conceptual until starting from the early 20$^{\text{th}}$ century a sequence of fascinating discoveries coupled with leaps in technological capabilities led to the establishment of the standard model of particle physics which describes the elementary particles and their interactions with great precision. Progress in nuclear and particle physics is driven by the development of detectors that provide information about the properties of particles and radiation interacting with them. Following the discovery of radioactivity by H.Becquerel in 1896, the invention of the cloud chamber [1] and subsequently the bubble chamber [2] opened new horizons by allowing the visualization of particles and their interactions through the observation of their tracks. Even though early detectors have led to the discovery of many new particles, photographic exposure was the only way of recording complex interactions up until the 1980s. Therefore, their main limitation was the inability to cope with high interaction rates and perform fast analysis of the results.

Modern experiments using particle colliders require the simultaneous detection of hundreds of particle tracks with micrometer spacial and nanosecond timing resolution. The first step towards fully electronic readout was the development of the multi-wire proportional chamber [3] which allowed the detection of particle tracks with accuracy in the order or 1 mm, but the leap in detector technology that provided the means to address such requirements stems from the invention of the transistor and the evolution of silicon microelectronics technology. Miniaturization of electronic devices has progressed exponentially over the decades following the prediction of Moore's law stating that the number of transistors in integrated circuits doubles every $\approx$2 years, while manufacturing costs are reduced. Although semiconductor detectors have been used in the 1960s for gamma ray energy measurements, the key that revolutionized measurement quality is the ability to manufacture position-sensitive segmented detectors with micro-structured electrodes in the range of 50-100 µm. Silicon microstrip detectors [4], in which one electrode is segmented in thin parallel strips, were developed in the 1980s and are still in use today as part of tracking detectors further away from the interaction point. They feature channel density of the order of 100 channels/cm^2 and spatial resolution in the order of 10 µm.

Pixel detectors [5] are the next step in this evolution as they require more sophisticated technologies. While there is no large improvement in spatial resolution, their main advantage is that they return true 3D space points, which is a necessity for pattern recognition, tracking and reconstructing vertices of short-lived particles. Furthermore they combine low material budget with channel densities higher than 5000 channels/cm^2. Each pixel is an independent smart sensing element that incorporates advanced functionality to condition and process the signal and temporarily store data to be subsequently transmitted. This level of integration has been only achievable with modern sub µm Complementary Metal Oxide Semiconductor (CMOS) processes and renders pixel detectors capable of coping with high interaction rates and energies that are essential in probing of increasingly rare processes. Therefore, they are the instrument of choice for tracking and vertexing of particles close to the interaction point of three experiments at the Large Hadron Collider (LHC) [6] at the European Organization for Nuclear Research (CERN). Pixel detectors have been crucial in the experimental discovery of the Higgs boson [7, 8] by the ATLAS [9] (A Toroidal LHC Apparatus) and CMS [10] (Compact Muon Solenoid) detectors. The layout of the ATLAS detector is shown in Fig. 1.1(a), while an example candidate Higgs boson event as captured by the ATLAS detector is depicted in Fig. 1.1(b). No other detector instrument is capable of addressing as well the unprecedented track density, rate and fierce radiation environment originating from the current LHC figures of 14 TeV center of mass energy and instantaneous luminosity up to $2 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.



(a) The ATLAS experiment at CERN. The pixel detector is located close to the interaction point [9]

(b) Event display of a $H \rightarrow 4e$ candidate event captured by the ATLAS detector in 2012 [11]

Figure 1.1: The ATLAS detector at the LHC

Hybrid pixel detectors [5, 12] in which the pixellated sensor-diode and the readout (R/O) chip are separate entities, mated by employing bumping and flip-chipping technology, have been invented for the LHC and have matured over the years. They are so far the only viable concept to achieve the required radiation tolerance and hit rate capability. Their disadvantages are most notably material thickness, production complexity and cost. An alternative approach that mitigates these drawbacks are the so-called Monolithic Active Pixel Sensors (MAPS) [13–16] that integrate the sensor and the readout electronics on the same silicon crystal and can be manufactured using commercial CMOS technologies. Their use has so far been restricted in low radiation environments due to slow and incomplete charge collection mainly by diffusion. Recently, advancements and freedom in CMOS technologies allowing the use of high-voltage (HV) and high-resistivity (HR) add-ons and modifications have come in the R&D focus of monolithic pixels for high rate/radiation levels. New MAPS developments, called Depleted Monolithic Active Pixel Sensors (DMAPS) [17–21] are full CMOS pixel structures with fast charge collection by drift in a depleted active volume.

DMAPS development has been mainly driven by the High-Luminosity upgrade of the LHC (HL-LHC) [22, 23] in 2026. To be able to cope with an order of magnitude increase in luminosity, tracking detectors will be completely replaced and new generations of pixel detectors are being devised. Although hybrid pixels have been selected as the baseline for the CMS and ATLAS inner tracker (ITk) [24, 25] upgrade and materialized through the RD53 collaboration [26], CMOS DMAPS have been proposed as a cost and production efficient solution to cover the large area of the ATLAS ITk pixel detector outer layer, where the requirements in terms of occupancy and radiation are comparable to the current ATLAS pixel detector inner layers [9, 27]. Different prototyping designs in various technologies have been developed in the framework of an ATLAS CMOS demonstrator program [24]. After the successful characterization of stand-alone sensors, as well as pixel arrays bonded to a dedicated readout chip (usually FE-I4 [28]), the community engaged on the design of large fully monolithic CMOS pixel matrices with on-chip digital readout. Currently, there is high R&D swing on DMAPS, which is expected to grow as they are considered the preferred technology for future ultra-light, highly-granular, poly-layer vertex detectors. Furthermore, as was the case for hybrid pixels, imaging devices using MAPS are also spinning off targeting astrophysics, biomedical and crystallography applications using X-ray and synchrotron radiation.

This thesis is dedicated to the design and development of large scale DMAPS prototypes with small collection electrode and fully integrated standalone readout architecture for high rate and high radiation applications such as the ATLAS ITk outer pixel layers. A small collection electrode leads to high granularity and low power consumption but its charge collection properties are not inherently suited for high radiation environments. In order to enhance radiation tolerance, a novel modification of the TowerJazz 180 nm CMOS imaging process [29] has been employed. The first, half scale ($1 \times 2\,\mathrm{cm}^2$), prototype chip that was developed on the modified process is called TJ-Monopix1 [30–32] and has been successfully characterized and proven to be fully functional [33, 34]. TJ-Monopix2 is a full scale ($2 \times 2\,\mathrm{cm}^2$) successor that incorporates several improvements and addresses the shortcomings of TJ-Monopix1 towards a fully efficient DMAPS implementation with small collection electrode and is currently in the initial testing phase.

CHAPTER 2

Challenges and Requirements for the New Generation of Pixel Detectors

2.1 Pixel detectors for particle tracking and vertex identification

Pixel detectors are an essential part of experimental high-energy physics instrumentation. In this context, they constitute tracking detectors that image the trajectories of charged particles and are highly segmented and fast i.e. able to capture millions of events (images) per second. When specifically optimized to topologically reconstruct the displaced vertex structure of interactions, indicating particle creation or annihilation they are called vertex detectors.

The development of high-granularity solid-state pixel detectors has been triggered mainly by the possibility of studying short-lived particles and the capability of coping with high interaction rates and energies (number of particles) in modern particle accelerators. High-energy colliders can generate elementary particle collisions at a high rate (>10 MHz), each producing 10-100 particles. Some rare, short-lived particles such as heavy quarks and leptons, that are of particular interest, quickly decay into a few secondary particles. The primary vertex indicates the location of the initial hard interaction, while secondary vertices often reflect the presence of a heavy particle such as a charm quark or a tau lepton. The topology of such a decay is depicted in Fig. 2.1.

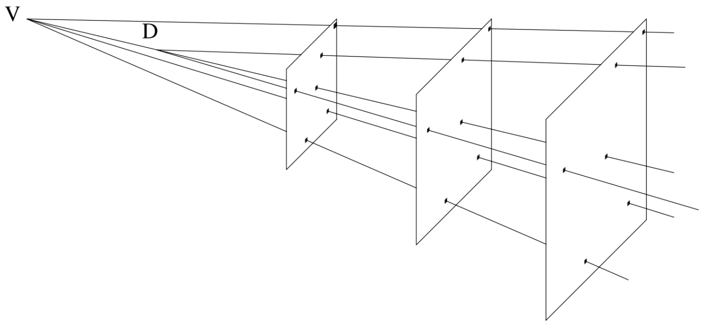


Figure 2.1: Topology of a short-lived particle decay. V indicates the collision vertex while D indicates the decay vertex. Three pixel detector planes provide true three-dimensional (3D) space points are used for track reconstruction [5].

The tracker is therefore fundamental in the identification of jets (tagging) such as b-jets and τ-leptons that signify important physics processes. The decay of B-hadrons and τ-leptons is mediated by the weak interaction and usually takes place inside the beam pipe with a decay length that is increased by the Lorentz boost to $l_{\mathrm{dec}} = \beta\gamma c\tau_{\mathrm{dec}}$ where τ_{dec} is the particle life-time. An example of an event at a collider experiment which contains the decay of a B hadron is shown in Fig. 2.2. Extrapolated tracks that miss the primary interaction point suggest secondary decay vertices. This deviation is quantified by the impact parameter d_0 which is defined as the closest approach to the interaction point at the momentum deflection plane. The resolution of the impact parameter σ_{d_0} is the primary figure of merit for vertex tracking accuracy and as will be shown depends on the detector single point resolution, geometry and material budget. With a single point (position) resolution of ≊ 10 µm, a vertex separation in the order of 100 µm can be typically resolved [35].

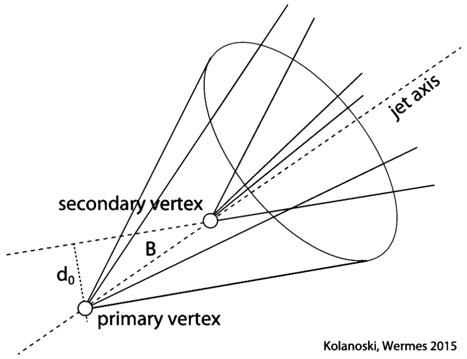


Figure 2.2: Sketch of a b-quark jet. A B hadron decays at a secondary vertex, which is quantified by the distance of the closest approach from the primary vertex, called the impact parameter d_0 [35].

In modern high-energy collider experiments, while seeking new physics, increasingly rare processes must be probed. Therefore, a high interaction rate is essential. Combined with the high primary particle energy, it results in a large amount of particles and dense tracks being generated. Thus, the tracker does not only need to be accurate but must also be highly granular meaning that enough sensing elements are available to reduce pile-up and distinguish tracks that pass close to the decay point, a feature unique to pixel detectors.

Particle detectors at accelerators are usually immersed in a magnetic field and act as a magnetic spectrometer that allows the determination of the momentum of charged particles. Therefore, another role of tracking detectors is the precise reconstruction of the particle track curvature that infers momentum and charge. In collider experiments, typically an arrangement of a solenoid spectrometer is used where the beams cross the center of the solenoid. To cover as high percentage as possible of the full solid angle, but also be mechanically feasible, the sensitive layers are usually arranged as coaxial barrel layers supplemented by several end-cap disks.

A number of competing conditions such as the physics needs, the interaction rate and intensity, technical capabilities and the available funds influence the design of a pixel detector and the solution to this optimization problem has led to a variety of designs in use today and envisioned in the future. In this chapter, the basic design parameters that influence the tracking performance of pixel detectors will be presented. State-of-the art, performance drivers and the requirements for the new generation of pixel sensors will be also discussed with emphasis on the HL-LHC ATLAS inner tracker (ITk) upgrade as it forms the basis for the specifications of the pixel sensors developed and presented in this work.

2.1.1 Position resolution

Pixel detectors employ structured electrodes to take a precise measurement of the space coordinates of a signal in a reference plane. The resolution is determined by the segmentation width (pixel pitch), the signal distribution function to neighboring pixels (charge sharing) and the signal to noise ratio (SNR), therefore the minimum achievable threshold. In some applications only binary information is given while in others the signal is measured in proportion to the integrated charge (analog readout). In order to understand the basic principles of vertex measurement, a simple approximation of a noiseless binary readout assuming uniform particle occupancy and full efficiency is sufficient. To simplify the analysis, the spatial resolution across one dimension (x coordinate) will be considered. The extension towards two orthogonal coordinates is straightforward. For equal pitch and resolution across the x and y dimensions, the two dimensional space resolution is $\sigma_r = \sigma_x \cdot \sqrt{2}$ [35].

A binary response is given for example when the charge density distribution width is small compared to the electrode width and no interpolation is done. The spatial resolution is defined by the standard deviation of the distribution of the measurement error, that is calculated by the root mean square function. For the presumed uniform distribution $f(x) = 1$ of particle illumination across the electrode width d from $-d/2$ to $d/2$ the standard deviation of the measurement error is given by:

$$\sigma_x^2 = \frac{\int_{-d/2}^{d/2} x^2 f(x)\,\mathrm{d}x}{\int_{-d/2}^{d/2} f(x)\,\mathrm{d}x} \Rightarrow \sigma_x = \frac{d}{\sqrt{12}} \tag{2.1}$$

The position resolution can be improved with the expense of a larger total data volume by recording and transmitting analog hit charge information per pixel. In this case the reconstructed hit position can be obtained by the center of gravity method:

$$x_{\text{rec}} = \frac{\sum(S_i + n_i)x_i}{\sum(S_i + n_i)} \tag{2.2}$$

where the S_i and n_i are the signal and noise fractional weights respectively and x_i is the center of each individual pixel in the cluster [20, 35]. The optimal achievable resolution depends on the proper matching of the electrode width with the signal charge distribution for a given SNR. While silicon trackers have typically had granularities greater than or equal to the charge cloud deposit, in current detector development increased spatial and temporal granularity is essential to make use of the structure of charge deposits.

2.1.2 Multiple scattering

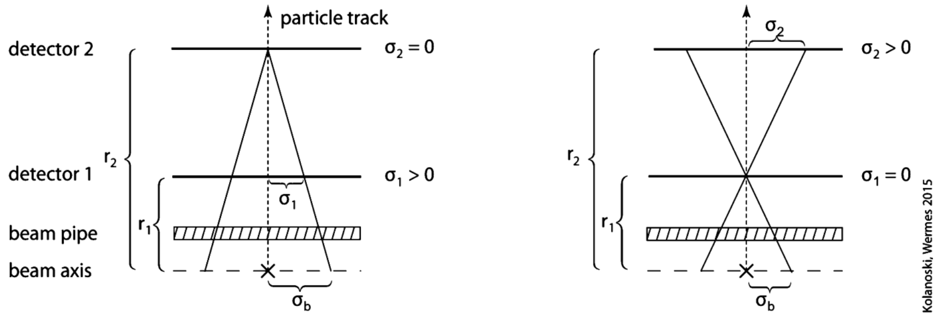
Charged particles passing though matter are scattered in the Coulomb field of the nuclei. This process is described by the Rutherford cross section which yields high probability for small angles. Therefore, for a not too thin scattering medium, a large number of independent scatters occur that can result in a significant deviation of the particle direction. This process is called multiple or Molière scattering and can be in most practical cases approximated by a Gaussian distribution. The characteristic quantity of multiple scattering is the standard deviation of the distribution of the projected scattering angle which can be approximated by [35]:

$$\theta_{\mathrm{ms}} = \frac{0.0136\,\mathrm{GeV/c}}{p\beta}|z|\sqrt{\frac{x}{X_0}} \tag{2.3}$$

where p, β and z are momentum, normalized velocity and charge of the particle respectively and x/X_0 is the path length in the scattering medium in units of the radiation length X_0. From (2.3) it is obvious that the uncertainty due to multiple scattering depends on p and becomes important for low momentum particles.

2.1.3 Impact parameter and vertex resolution

It is important to assess which parameters are crucial for a microvertex detector with regards to its accuracy especially in terms of pixel pitch, detector geometry and area (number of layers, level arm, distance of the first layer to the primary vertex) and material thickness. Doing so will allow us to understand the significance of high granularity, low mass and low production effort that monolithic pixels (DMAPS) can offer. We can start the analysis by considering a simple, but indicative, case of a two layer vertex detector installed outside of the beam pipe near the interaction point as shown in Fig. 2.3. For simplicity, we assume planar modules and straight tracks traversing the layers perpendicularly.



(a) Detector 2 assumed to be perfect.

(b) Detector 1 assumed to be perfect.

Figure 2.3: Simplified two-layer vertex detector. The layers are cylindrically arranged at distances r_1,r_2 and have position resolutions σ_1,σ_2 in the plane perpendicular to the beam [35].

The impact parameter error of the track, σ_b can be calculated by assuming that in each case one of the two layers is perfectly accurate. It can be geometrically determined by the resolution of each layer and the aspect ratio of the two layers and is equal to [12, 35]:

$$\sigma_b^2 = \left(\frac{r_1}{r_2 - r_1}\sigma_1\right)^2 + \left(\frac{r_2}{r_2 - r_1}\sigma_2\right)^2 \tag{2.4}$$

This result can be expanded for a linear track fit in a general case of N equally distributed layers with spacing $L_p/(N-1)$ measured with spatial resolution σ_{meas}, as given by [35]:

$$\sigma_{d_0} = \frac{\sigma_{\mathrm{meas}}}{\sqrt{N}}\sqrt{1 + \frac{12(N-1)}{N+1}\left(\frac{x_0}{L_p}\right)^2} \tag{2.5}$$

where x_0/L_p is the ratio of the extrapolation lever arm to the length over which the measurement points are distributed. If the origin of the reference frame is chosen in the center of the track, x_0 is the distance from the center of measurements to the extrapolation point (primary vertex).

The impact parameter resolution is further reduced by the effects of multiple scattering discussed in section 2.1.2. For the aforementioned case of equally spaced detector layers, equal thickness d and single layer scattering angles $\theta_{\mathrm{ms,sl}}$, an estimator of the slope error σ_b can be calculated by summing in quadrature the contribution of each layer [35]:

$$\sigma_{b,\mathrm{ms}} = \frac{0.0136\,\mathrm{GeV/c}}{p\beta\sin\theta}|z|\sqrt{\frac{d/\sin\theta}{X_0}}\sqrt{\frac{N(2N-1)}{6(N-1)}} \tag{2.6}$$

where d is the thickness of the detector plane and θ is the angle between the trajectory and the direction of the magnetic field. Considering only the error due to multiple scattering, the primary vertex can be extrapolated from the first layer at r_1. Therefore, the impact parameter error due to multiple scattering is given by:

$$\sigma_{d_0,\mathrm{ms}} = \sigma_{b,\mathrm{ms}} r_1 \tag{2.7}$$

with r_1 being the distance from the first layer to the primary vertex.

To achieve high vertex resolution, high intrinsic position resolution (segmentation) and a large lever arm (L_p) is required. The innermost layer of the vertex detector should be as close as possible to the interaction point to take advantage of the short extrapolation length and its resolution is the most important. In general the minimal distance is limited by the beam pipe, the radius of which should not be too small because of the radiation background near the beam. Low detector material (x/X_0) is needed in order to minimize multiple scattering, especially for low momentum tracks.

2.1.4 Momentum resolution

The transverse momentum p_T is calculated by the curvature κ of the particle track inside a magnetic field B:

$$p_T = \frac{0.3|z|B}{|\kappa|}, \quad [p_T] = \mathrm{GeV}/c \tag{2.8}$$

It is typically determined by the measurement of the sagitta of the curved path between the points of the entrance into and exit from the detector volume. For a detector with N equally spaced layers and in the limit of large N, the resolution of the transverse momentum is given by the 'Gluckstern' formula, which is valid for high momenta where multiple scattering can be neglected [20, 35]:

$$\left(\frac{\sigma_{p_T}}{p_T}\right) = \frac{p_T}{0.3|z|}\frac{\sigma_{\mathrm{meas}}}{L_p^2 B}\sqrt{\frac{720}{N+4}} \tag{2.9}$$

where σ_{meas} is the position resolution of each layer and L_p is the radial length of the detector.

The contribution of multiple scattering to momentum resolution can be written as [20, 35]:

$$\left(\frac{\sigma_{p_T}}{p_T}\right)_{\mathrm{ms}} = \frac{0.0136}{0.3\beta L_p B}\sqrt{\frac{(N-1)\ d/\sin\theta}{X_0}}\sqrt{C_N} \tag{2.10}$$

where L_p is the detector (tracker) length projected onto the plane perpendicular to the magnetic field and $(d/\sin\theta)/X_0$ is the total material thickness traversed by a particle incident with polar angle θ with respect to the beam in units of radiation length. C_N is a factor that describes different detector layouts and depends on the number of layers. For the minimum of three layers ($N = 3$) one obtains $C_N = 2.5$, while for $N \to \infty$ the continuous scattering case is approached with $C_N \to 1.33$.

The dependence of momentum resolution to the lever arm L_p is strong and scales with L_p^2. To achieve a large lever arm while optimizing performance and cost, in most cases the tracker is split in layers with different technologies. High granularity pixel detectors are used for the innermost layers (vertex detector) while the outer layers of the tracker consist of less expensive concepts such as strip detectors. The error of momentum resolution due to position measurement increases with the transverse momentum P_T while the error due to multiple scattering is independent of P_T. Therefore, for small momenta the total error reaches a plateau given by the multiple scattering term while for high momenta the measurement term dominates.

2.2 Challenges and requirements for state-of-the-art and future trackers

Experiments at the energy frontier are dominated by large accelerators and multipurpose collider detectors. The current state-of-the-art is represented by the detectors in operation or under construction at the LHC at CERN, prominently by the ATLAS and CMS experiments. The challenges that drive ongoing development of detectors for hardon colliders can be broadly categorized into scale, intensity and performance [20]. Planned upgrades such as the High-Luminosity LHC (HL-LHC) [23] and the subsequent High-Energy LHC (HE-LHC) [36] and future hadron colliders such as the conceived Future Circular Collider (FCC) [37] focus on two main parameters that allow to refine the standard model and improve the physics reach: increasing the energy and intensity of the beam. By doing so, the number of interactions per bunch crossing drastically increases. From the pixel detector perspective two main demands arise: the ability store and process hits at a high rate per unit area and high radiation tolerance to survive the harsh environment close to the interaction point.

In order to cope with high hit rates, the pixel size has to be small and the outer radius of the detector has to be increased. Since every "hit" on the pixel detector must be time-stamped and stored for the trigger latency interval, more memory per unit area is required to be included in the readout integrated circuit (ROIC). Therefore, silicon CMOS technologies in smaller process nodes are needed to be exploited in order to achieve the required logic density. Higher intensity also means that in order to distinguish events of interest from the background, the trigger rate must be increased. Therefore, pixel detectors output much more data that requires integrated, fast electrical links that can maintain high signal quality while using low mass cables.

Challenges for lepton colliders such as the Belle II [38] experiment at superKEKB, the planned compact linear collider (CLIC) [39] at CERN and planned the international linear collider (ILC) [40] differ from those of hadron colliders. Exploring the physics potential requires precision measurements, especially at low momenta and mandates for higher granularity and lower mass by one to two orders of magnitude compared to the current LHC detectors. On the other hand, the requirements imposed by the beam structure, data rate and radiation dose are much more modest compared to the LHC. An exception is the time structure of the beam at CLIC that requires fast timing response due to the small bunch separation (0.5 ns).

The aforementioned challenges should be compounded by the basic performance of the sensor and readout electronics. The most important factors are the signal amplitude, the sensor capacitance and leakage, the device specific transconductance and the noise of the ROIC. A set of these parameters and a maximum power budget defines an upper limit in the analog performance that is well understood and the ROIC should be designed to perform close to this limit.

Future trackers must function in extreme environments while maintaining or improving their performance. As previously discussed, high resolution demands a high-granularity, low-mass, poly-layer detector close to the interaction point. Furthermore this should be achieved with low power consumption and low cost and production effort. The performance drivers and design aspects that emerge in detector R&D can be summarized as follows:

- *Pixel size, hit rate and position resolution:* The two main drivers for the sensor pixel pitch P are the position (single point) resolution and two track separation, both scaling as P and local occupancy scaling as the pixel area (P^2 for square pixels) times the readout time. In high rate environments, such as the HL-LHC, a small pixel size is necessary to maintain full efficiency and avoid pileup and becomes more important when the first layer is placed close to the interaction point in order to improve the resolution. Furthermore, smaller pixels help to keep the leakage current per pixel small and improve noise after irradiation. In order to further improve the resolution and two track separation, the analog charge information can be utilized given that the SNR is high enough. Recent developments [20] address the processing of multi-hit pixel clusters as complex objects to exploit directional information. A high channel count, especially in the case of analog readout, requires higher logic density and an efficient, fast, zero-suppressed readout.

  When aiming for high resolution, the following limitations should be considered regarding the pixel size:

  - The smallest pixel size is determined by the amount of in-pixel electronics that are necessary to amplify, discriminate and store the hit information.
  - The spread of the charge cloud due to diffusion, which is given by:

    $$\sigma_x = \sqrt{2D_x/u_D} \tag{2.11}$$

    where D is the diffusion constant, x the drift distance to the electrode and u_D the drift velocity. The spread due to diffusion sets a constraint for the minimum pixel size since a pixel pitch below σ_x would lead to excessive charge sharing.
  - The technological limits of detector fabrication and assembly. An example is the fine pitch bump bonding process in the case hybrid pixel detectors (see section 3.1).

- *Material budget, radiation length and collection thickness:* Multiple scattering dominates the resolution for precision track fitting, particularly at low momenta. It becomes extremely important for heavy-ion, nuclear physics and rare muon decay experiments that operate in a multiple scattering dominated regime. Therefore, the material inside the detector has to be as low as possible. Traditionally, the pixel sensor and ROIC were the most significant contributors to the total material budget. In current developments, thinner sensors are being explored in order to reduce the material, provide faster charge collection and improve the charge spread due to diffusion. Monolithic pixels (DMAPS) allow for very low material by combining the sensor and ROIC in the same silicon crystal. Thicknesses down to 50 µm can be achieved beyond which mechanical supports, wiring and cooling services typically dominate.

- *Detector area, cost and power:* Cost and power are two important constraints that scale with the detector total area. A large detector area is a consequence of the following requirements:
  - Greater length along the beam direction that improves acceptance.
  - A large number of measurements (N) improves the impact parameter accuracy due to the resulting higher position resolution σ_{meas}. However, the resolution error due to multiple scattering also increases with N. Since having many measurement points is beneficial for pattern recognition and specific ionization measurements. in practice the number of layers must be optimized according the requirements and specific parameters of the detector.
  - Large radii that reduce the local pixel occupancy, at least for the outer layers, and increase the lever arm L_p which improves track fitting and vertex resolution.

  These requirements drive a strong focus on cost reduction which is also vital for valorization of the technologies developed for particle physics in other applications e.g. in the medical field. Since large areas have to be covered ($> 10\,\text{m}^2$), large volume manufacturing with less effort is essential. In the case of hybrid pixels, sensor manufacturing in commercial CMOS processes (passive sensors) is currently considered to reduce cost. DMAPS are not only manufactured in commercial processes, but also do not require the laborious bump-bonding process and therefore further reduce the cost and complexity (more details in section 3.2).

  A large detector area and small pixel size result in millions of pixels that have to operate with tolerable power consumption. For a given ROIC CMOS process and power budget the analog performance and power consumption are essentially determined by the detector capacitance, which is therefore important to be minimized by optimizing the sensor technology. Furthermore, given the huge number of channels, powering the detector using a conventional parallel wire scheme becomes impractical. In the case of the HL-LHC, a serial powering scheme with on-chip regulators will be implemented in order to reduce the required wiring. Low power is mandatory for compact arrays that require very low multiple scattering if cooling does not dominate the overall material budget.

- *Radiation tolerance:* Radiation effects mainly the include the sensor bulk damage due to non Ionizing Energy Loss (NIEL) and the damage on the readout electronics due to the accumulated total Ionizing Dose (TID). The innermost layer of the vertex detector, which should be close to the beam pipe to improve resolution, is the most severely affected by radiation damage. Radiation tolerance has been an crucial performance aspect even for the current state-of-the art detectors at the LHC and has been extensively studied with the help of a strong R&D program led through the RD50 [41] collaboration. For high energy and high intensity future accelerators, radiation damage is expected to increase by more that an order of magnitude and will become one of the most critical challenges.

- *Time resolution:* The charge collection time is typically in the order of $3 - 10\,\text{ns}$ and depends on the sensor thickness and electric field. It usually goes hand-in-hand with radiation tolerance as the later often relies on a fast signal collection by a strong drift field. For applications with particle bunches colliding at a specific frequency such as the LHC, the pixel detector has to be at least fast enough to assign a correct time stamp to each event. Very high timing resolution ($< 100\,\text{ps}$), required to identify the common origin of particles, has not yet been achieved by pixel detectors and therefore different detector types such as scintillators are used to provide this information. Recently, amplification structures in silicon such as the Low Gain Avalanche Diode (LGAD) [20] have been brought forward in an attempt to cope with these demands.

2.2.1 The high luminosity LHC (HL-LHC)

The Large Hadron Collider (LHC) [6] at the European Organization for Nuclear Research (CERN) is currently the most advanced particle accelerator. While it is mainly a proton collider, it can also operate with heavy ions. Protons in two counter-rotating beams are accelerated through a complex of accelerators, finally reaching a center of mass energy of 13 TeV in the 27 km long ring. The beams collide at four crossing points where the four largest experiments are located called ATLAS (A Toroidal LHC Apparatus) [9], CMS (Compact Muon Solenoid) [10], ALICE (A Large Ion Collider Experiment) [42] and LHCb [43]. The interactions between the counter-rotating beams are not continuous since the protons are bunched together and collide at discrete intervals, 25 ns apart, yielding a bunch crossing frequency of 40 MHz. The quantity that measures the ability of a particle accelerator to produce the required number of interactions is called luminosity $\mathcal{L}$ and is defined as the ratio of the number of events per second $\mathrm{d}R\,/\mathrm{d}t$ to the production cross-section σ_p:

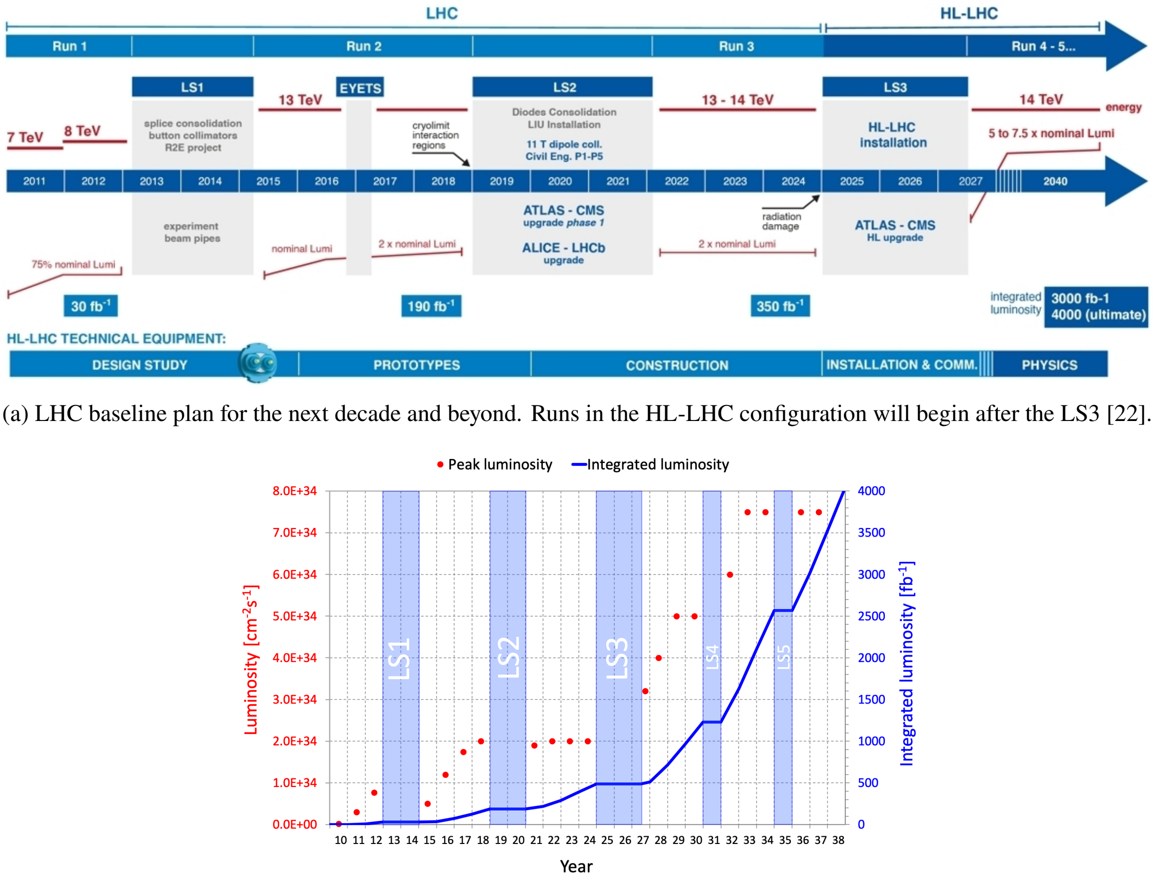
$$\frac{\mathrm{d}R}{\mathrm{d}t} = \sigma_p \cdot \mathcal{L} \tag{2.12}$$

The luminosity can be semi-qualitatively obtained from:

$$\mathcal{L} = \frac{N^2}{t \cdot \sigma_{\mathrm{eff}}} \tag{2.13}$$

where N is the number of protons in each beam (assuming they are equal), t is the time between bunches and σ_{eff} is the effective cross section of the collision that depends on the beam profile. The two largest experiments, ATLAS and CMS, are currently designed to operate at a nominal luminosity equal to $1 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, however a peak luminosity up to approximately $2 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ has been achieved before the Long Shutdown 2 (LS2) in 2019, while the integrated luminosity approached $190\,\mathrm{fb}^{-1}$.

The LHC has been successful in the detection of the Higgs boson and has provided valuable precision measurements to refine the standard model. However, since rare processes are vital to increase the physics reach, the statistical gain of operating the accelerator beyond 2020 without significantly increasing the luminosity will become marginal. To improve statistics of potential interesting interactions and utilize the full potential of the LHC, a major upgrade is planned to take place during the long shutdown 3 (LS3). The timeline of the upgrade is shown in Fig. 2.4(a) and the expected luminosity is depicted in Fig. 2.4(b). After the HL-LHC installation is complete, the instantaneous luminosity will gradually increase to its nominal value of $5 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, which is five times higher than the current LHC configuration, and can reach a maximum of $7.5 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The HL-LHC integrated luminosity is expected to become approximately 10 times higher than the integrated luminosity accumulated in the current LHC lifetime. Such an increase in intensity sets stringent requirements for the detector systems in terms of hit rate and radiation tolerance, especially in the case of the pixel detector inner layers. To cope with these requirements, the inner trackers of the CMS and ATLAS experiments will be upgraded.



(a) LHC baseline plan for the next decade and beyond. Runs in the HL-LHC configuration will begin after the LS3 [22].

(b) Forecast for peak luminosity (red dots) and integrated luminosity (blue line) in the HL-LHC era according to the ultimate HL-LHC parameters [23].

Figure 2.4: The HL-LHC upgrade schedule and luminosity forecast

2.2.2 Pixel detectors for the HL-LHC: the ATLAS ITk upgrade

The current ATLAS Inner Detector (ID) [44] is depicted in Fig. 2.5 (a). It consists of the pixel detector, the semiconductor tracker (SCT) that is composed of silicon strips and the transition radiation tracker (TRT). The pixel detector, shown in Fig. 2.5 (b) was originally designed as a system with three barrel layers around the beam axis and three disks in each direction along it. It covers a total area of approximately 1.7 m^2 and is composed of 80 million silicon hybrid pixels of $50 \times 400\,\mu\mathrm{m}^2$ size. During the long shutdown 2 (LS2), it was upgraded to improve the vertex resolution and efficiency in the pretense of high pile up. A fourth pixel layer, called the Insertable B-Layer (IBL) [27] was installed at $r = 3.3$ cm between the old innermost pixel layer and a new smaller radii beam pipe. Due to the higher track density during Run-2, the IBL operates with smaller pixels of $50 \times 250\,\mu\mathrm{m}^2$ size in order to reduce occupancy.

The LHC ATLAS inner detector will be replaced by the new ATLAS inner tracker (ITk) [24] for the HL-LHC operation. The proposed layout of the ATLAS ITk is shown in Fig. 2.6 (a).The total size is about the same as the ATLAS ID since the systems that surround it (e.g. calorimeter) will not be

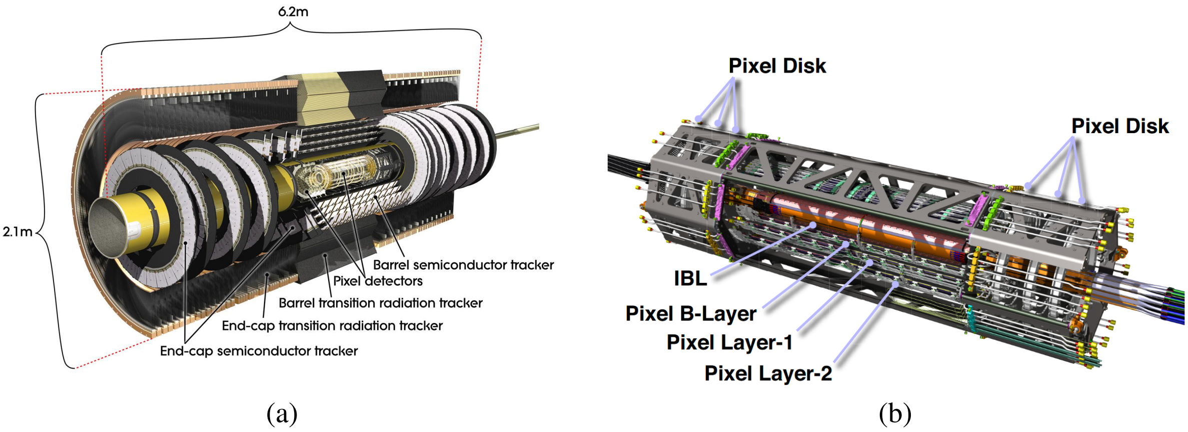

6.2m
2.1m
Barrel semiconductor tracker
Pixel detectors
Barrel transition radiation tracker
End-cap transition radiation tracker
End-cap semiconductor tracker
(a)
Pixel Disk
Pixel Disk
IBL
Pixel B-Layer
Pixel Layer-1
Pixel Layer-2
(b)

Figure 2.5: a) Cut-out view of the ATLAS inner detector [44], b) Layout of the ATLAS pixel detector after LS1 [27].

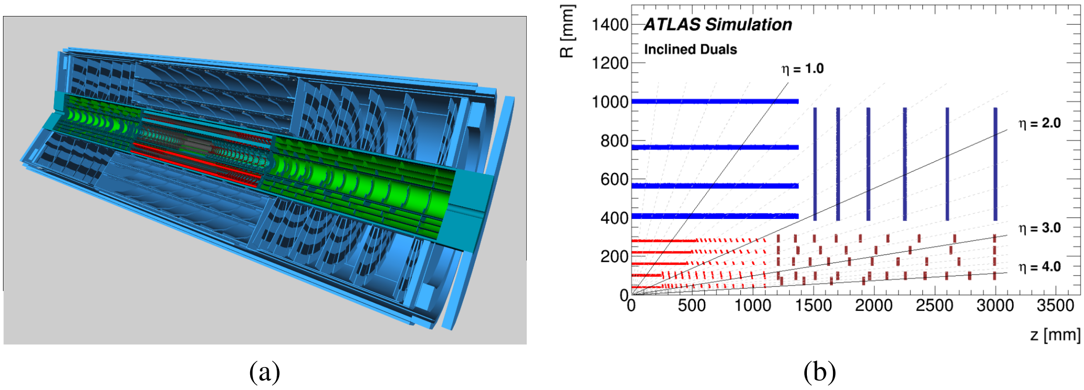


Figure 2.6: a) Display of the ATLAS Phase-II Inner Tracker ITk with the inclined duals detector layout, b) Detailed sketch of the pixel detector (red) and strip detector (blue) layers [24].

modified. It is comprised of two subsystems: a pixel detector and a strip detector that surrounds it. The most crucial part of the ITk, the pixel detector, is designed with the following set of goals and requirements:

- *Area and geometry:* The acceptance of a detector is usually expressed by the geometrical parameter of pseudorapidity $\eta = -\ln\tan\theta/2$ that describes the angle θ of a particle relative to the beam axis. The ITk pixel detector extends the tracking coverage to a pseudorapidity of $|\eta < 4|$ compared to $|\eta < 2.5|$ of the ATLAS ID. The pixel detector is composed of cylindrical barrels and several end-cap disks covering small θ angles. The number of barrel layers have been increased to five in order to cope with the higher rate and improve coverage and tracking performance. The inner layer is placed at a radius of $r = 3.6\,\mathrm{cm}$ and the outer layer (L4) has a radius of $r = 27.3\,\mathrm{cm}$, that is approximately 2 times higher than the outer layer of the ATLAS ID pixel detector. Thus, the total area is significantly higher compared to the ATLAS ID and is estimated to be as large as $12 - 14\,\mathrm{m}^2$.

- *Hit rate:* The HL-LHC luminosity is translated to an increase in the number of interactions per crossing from 20 to about 200 every 25 ns. Therefore, this leads to a high number of pixels being activated (hit) per unit of area and time (hit rate). The average hit rate for the inner layer of the ITk pixel detector is expected to be as high as 3 GHz/cm^2 while for the outer layer it drops to approximately 100 MHz/cm^2 which is similar to the hit rate of the inner layers of the current ATLAS pixel detector. To cope with the high hit rate, the pixel size of the ITk has been reduced to $50 \times 50\,\mu\text{m}^2$ and a new $19.2 \times 20\,\text{mm}^2$ ROIC in 65 nm CMOS technology with high logic density and fast data transmission has been designed by the RD53 collaboration [26].

- *Radiation hardness:* The radiation tolerance of the sensor and the ROIC is a critical requirement due to the high expected particle flux. As a result of the extreme radiation, the two innermost pixel layers are designed to be replaceable. In the case of the inner pixel layer, the total ionizing dose (TID) over 5 years of operation is expected to be about 500 Mrad and the NIEL fluence is expected to be approximately $2 \cdot 10^{16}\ n_{eq}/\text{cm}^2$. In the case of the outer layer (L4), the requirements are more relaxed and equal to 50 – 80 Mrad TID and $1 \cdot 10^{15} - 2 \cdot 10^{15}\ n_{eq}/\text{cm}^2$ NIEL.
- *Efficiency and noise occupancy:* The hit detection efficiency must be higher than 97% for the whole life-time of the detector, even after the expected radiation dose has been accumulated, and should be achieved while keeping the noise hit occupancy per pixel lower than $1 \cdot 10^{-6}$ /25ns.
- *Time resolution:* The bunch crossing frequency of the HL-LHC will remain equal to 40 MHz. Therefore, hits should be detected and recorded within 25 ns, otherwise they will be assigned a wrong time stamp and will be regarded as noise.
- *Material budget and power consumption:* The total material budget should be below 2% in radiation length (x/X_0) units. Furthermore, the power consumption should remain smaller than 500 mW/cm^2 in order to keep the cooling and power delivery requirements reasonable and avoid increasing the material budget.

Hybrid pixels that are being developed based on the RD53 ROIC have been selected for the ITk pixel detector in order to cope with the extreme conditions close to the interaction point. While an implementation of monolithic pixels (DMAPS) for the innermost layers would benefit performance (for example in the case of b and τ-tagging) due to their small pixel size and thickness, at present the logic density required for the foreseen hit rate has not been achieved as a result of the feature sizes offered by current CMOS technologies suitable for DMAPS. Furthermore, the radiation hardness that is required by these layers is a serious challenge for current DMAPS designs and has not yet been systematically checked. On the contrary, DMAPS constitute a promising alternative for the outer layer of the ITk pixel detector since it covers the largest area and its radiation and rate requirements are reduced by roughly an order of magnitude compared to the inner layers and are comparable to the present LHC ATLAS pixel detector. It is estimated that the module production cost of the outer layer (L4) can be reduced by about 2.5 MCHF if DMAPS are adopted [23]. Table 2.1 gives a summary of the most important requirements of pixel detectors for the HL-LHC and compares them with representative experiments that can favor different types of pixel technologies.

The DMAPS detector chips developed in the context of this thesis have been designed to target the requirements of ATLAS ITk outer layer (L4), but are also ideally suited for other experiments, such as future upgrades of the Belle II detector where the pixel size, material and power consumption is of utmost importance while other requirements such as radiation hardness and hit rate are more relaxed.

Table 2.1: Comparison of the HL-LHC ATLAS characteristic requirements with representative HEP experiments

	STAR	**ILC**	**ALICE LHC** (Heavy-Ion)	**ATLAS LHC**	**ATLAS HL-LHC** **Outer**	 **Inner**
Time resolution (ns)	110	350	20000	25	25	
Particle Rate (MHz/cm^2)	0.4	25	1	100	100	up to 3000
NIEL Fluence (n_{eq}/cm^2)	$> 10^{12}$	$> 10^{12}$	$> 10^{13}$	$2 \cdot 10^{15}$	$1 - 2 \cdot 10^{15}$	$2 \cdot 10^{16}$
TID (Mrad)	0.2	0.4	0.7	80	up to 80	> 500
Pixel pitch (µm^2)	21x21	—	28x28	50x250	50x50	50x50
Pixel type	MAPS	—	MAPS	Hybrid	Hybrid or DMAPS	Hybrid

CHAPTER 3

Monolithic Pixel Detector Fundamentals

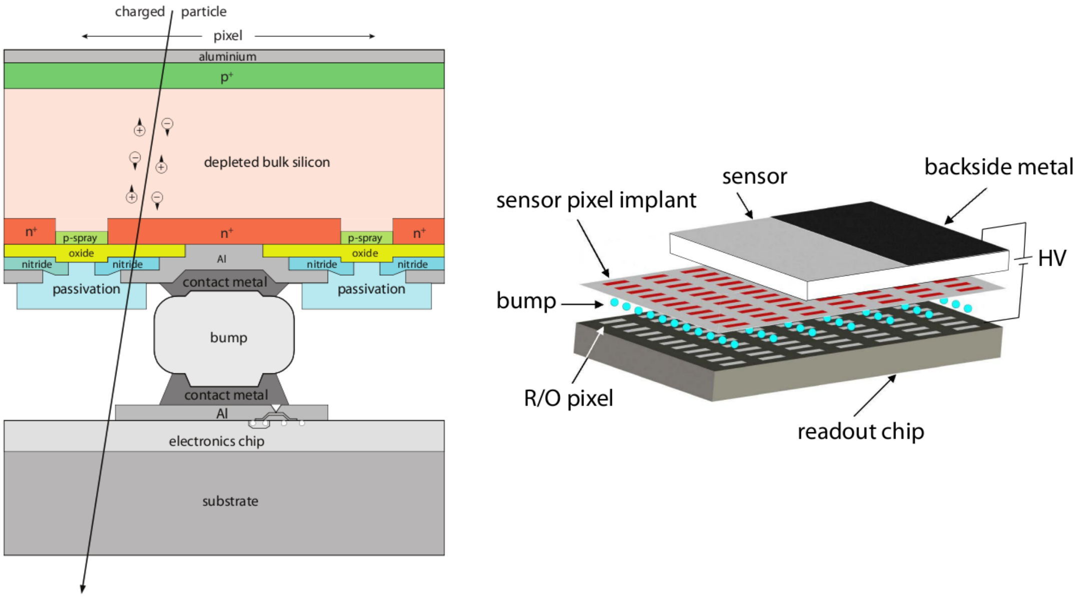
Pixel detectors are composed of highly segmented solid-state semiconductor elements. Their basic structure consists of two parts, the sensor where a signal of electron hole pairs (e/h) is generated by ionization from a traversing particle and the readout electronics that process, digitize, store and transmit the hit data. Silicon is by far the most widely used material for semiconductor detectors because it is abundant, stable, has a relatively low bandgap that allows for sufficient charge generation and can also form high quality insulators (SiO_2). For the detection of charged particles in accelerators such as the LHC, silicon is an ideal material, while for other applications, such as X-ray detection, other semiconductors (e.g. CdTe) provide better absorption characteristics. Readout chips are nearly exclusively produced from silicon, following the microelectronics CMOS technology revolution.

The sensor and the ROIC can be separate entities or integrated in the same silicon crystal leading to two different approaches in the way a pixel can be structured, called hybrid and monolithic respectively. In this chapter, the concept of Depleted Monolithic Active Pixel Sensors (DMAPS) along with its unique characteristics will be described and compared to the state-of-the-art hybrid pixels. To provide a better understanding of the concepts that will be discussed, fundamental principles of silicon detector operation and key performance criteria will be introduced.

3.1 Silicon detector fundamentals and the state-of-the-art: Hybrid pixels

Hybrid pixels have been established as the technology of choice for high occupancy and high radiation environments by the LHC ATLAS and CMS experiments. They are made up of two parts, as shown in Fig. 3.1, which exactly match: The sensor part that is produced in specialized sensor grade silicon material and one or several readout chips with the same cell pattern that are manufactured in standard CMOS processes. Sensor and chip are connected in every pixel using flip-chipping/bump-bonding and 3D integration techniques. As mentioned in Chapter 2, hybrid pixels with sizes down to $50 \times 250\,\mu m^2$ have been successfully manufactured with high yield for the LHC, while $50 \times 50\,\mu m^2$ size pixels will be installed during the HL-LHC upgrade. The preferred bonding techniques are eutectic soldering and In-In thermocompression. During the initial LHC detector development, a bonding pitch in the scale of 50 um was 15 years ahead of industrial demands. This pitch has become a standard today and connection density down to 25 µm has been achieved.

The main benefit of hybrid pixels is that the sensor and the ROIC part can be separately optimized to excel in their specific task: fast, efficient charge collection even after irradiation and low noise,



(a) (b)

Figure 3.1: Hybrid pixel detector: a) Individual pixel layout composed of a dedicated sensor and the readout electronics cell connected through a bump-bond, b) Illustration of a hybrid pixel detector matrix [20]

fast, high rate capable and radiation tolerant electronic readout respectively. The disadvantages of the hybrid approach become evident when addressing the requirements of future particle detectors demanding high granularity (small pixel size), low material budget and large area coverage. The bump-bonding process adds an extra step in module manufacturing, is labor intensive and increases cost considerably. Additionally, the vendors providing the specialized sensor wafers do not usually operate large volume production lines increasing cost and turnaround times. The technological limitations are mostly related to the bumping tecnhology and the power density associated with the minimum achievable pixel capacitance and the fact that the electronics circuitry is confined to the same area as the detecting electrode.

3.1.1 The sensor: signal generation and formation

3.1.1.1 Energy loss of charged particles

When traversing a medium, charged particles lose energy that can be mainly attributed to a sum of contributions from ionization, atom excitation and bremsstrahlung radiation. The energy being lost by each interaction arises from individual stochastic processes. Therefore, the average energy is used instead which is also called stopping power since it describes how particles are stopped in matter. There are different regions, depending on the passing particle mass and momentum where different energy loss processes dominate and are described by a different theoretical description.

As an example, the dependence of stopping power as a function of $\beta\gamma = p/mc$ for μ^+ in copper is given in Fig. 3.2. The energy loss by bremmstrahlung scales with the inverse squared mass of the radiating paricle ($1/m^2$) [35]. For electrons and positrons, due to their small mass, energy loss by bremmstrahlung becomes relevat at relatively small energies. However, in the case of heavier particles, the radiation contribution to energy loss becomes significant at very high energies.

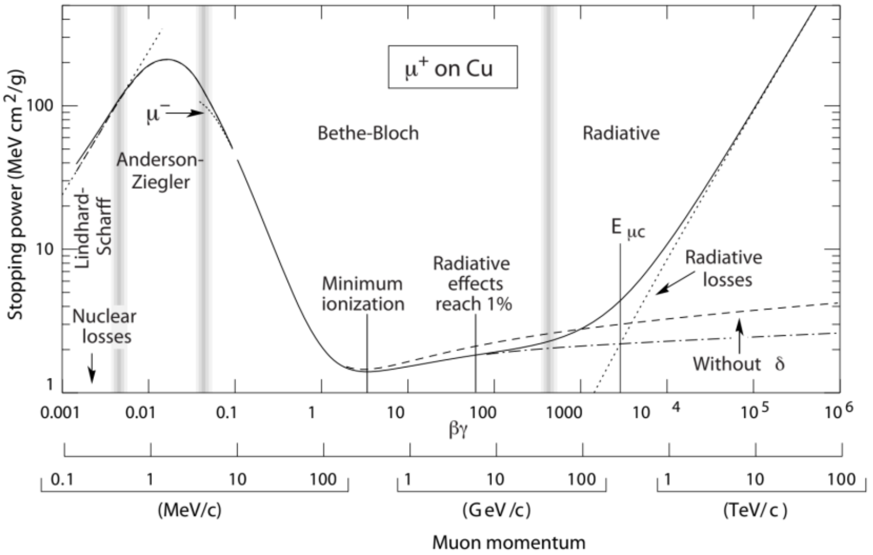


Figure 3.2: Stopping power ($-\,\mathrm{d}E\,/\mathrm{d}x$) for μ^+ in copper as a function of $\beta\gamma$. Each region is described by a different theoretical description. At high energies bremsstahlung (radiative losses) dominate. The center region is described by the Bethe-Bloch formula. For lower energies, the characteristic $1/\beta^2$ dependence of the Bethe-Bloch formula is no longer valid as quantum mechanical interference effects and energy-depended shell corrections become relevant. For very low energies non-ionizing energy loss caused by elastic nuclear recoil dominates [45].

The energy loss by ionization and atom excitation is given by the Bethe-Bloch formula [35] which accurately describes the central region of energies in Fig. 3.2:

$$\left\langle -\frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle = K\frac{Z}{A}\rho\frac{z^2}{\beta^2}\left[\frac{1}{2}\ln\frac{2m_ec^2\beta^2\gamma^2T_{\max}}{I^2} - \beta^2 - \frac{\delta(\beta\gamma)}{2} - \frac{C(\beta\gamma, I)}{Z}\right] \tag{3.1}$$

where:

- $K = 4\pi N_A r_e^2 m_e c^2$. N_A is Avogadro's number, m_e is the electron mass and r_e is the classical electron radius ($\approx 2.8\,\mathrm{fm}$).
- ζ, β are the charge and velocity of the projectile particle.
- Z, A are the atomic mass and number of the medium.
- I is the mean excitation energy.
- $T_{\max}$ is the maximum energy transfer to a shell electron (central collision).
- δ is the so called density correction, important at high energies.
- C/Z is a shell correction factor, relevant for small β values.

At low energies the stopping power is high as the $1/\beta^2$ term becomes dominant. This dependence can be explained by the fact that momentum transfer increases with the effective interaction time, that is longer for slower particles. At high energies the $ln\gamma$ term is dominant. The reason of the stopping power rise at high energies is twofold: The first one is the asymptotic increase of the maximum energy transfer ($T_{\max}$) with γ and the second is the increase of the traverse electric field extension with γ (relativistic effect). In between these regions there is a minimum at about $\beta\gamma \approx 3 - 3,5$ whose exact value depends on Z. Particles with kinetic energies in this range are called Minimum Ionizing Particles (MIPS). Since the energy loss increase for $\beta\gamma > 3 - 3,5$ is moderate, it is common practice to also refer to particles with energies higher that this minimum as MIPS. The concept of MIPS is very important in pixel detector operation and characterization since most relativistic particles relevant to collider experiments have mean energy loss rates close to this minimum that represents the worst case of the generated signal magnitude.

The Bethe-Bloch formula describes the average energy loss per path length. However, since energy loss is statistical in nature, fluctuations occur both in the number of ionization/excitation processes and the emitted energy δE in each process. For relatively thin sensors such as silicon sensors used by pixel detectors, the energy loss probability density function follows a Landau distribution [35] which in practice is often more accurately described by convolution with a Gaussian function. An example of the energy loss distribution in the case of 500 MeV pions in silicon is shown in Fig. 3.3. The energy loss fluctuations lead to a distribution that is asymmetric and is composed of a Gaussian part that corresponds to many ionization processes with small energy loss and a tail that corresponds to high energy loss values. Large energy losses occur due to hard collisions with shell electrons that transfer a high amount of energy to individual electrons, the so called delta electrons. It is worth noting that the Most Probable Value (MPV) Δ_p of the energy loss is smaller than the average value ($\langle \mathrm{d}E / \mathrm{d}x \rangle$) depending on the symmetry of the distribution.

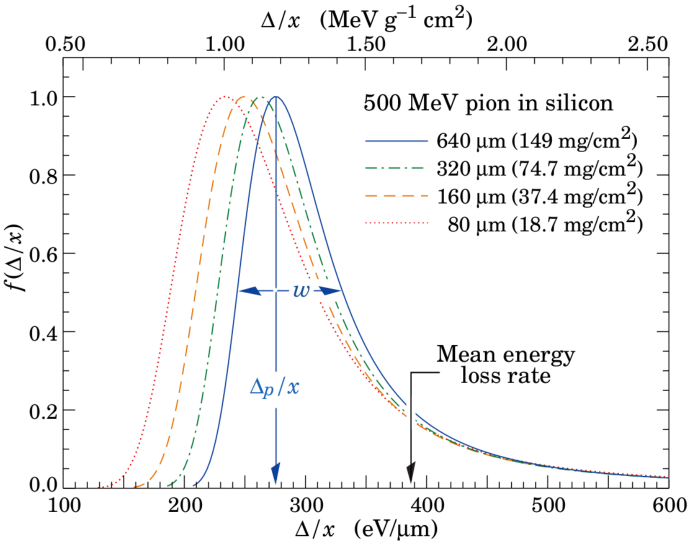


Figure 3.3: Energy loss distribution of 500 MeV pions in silicon, normalized to unity at the MPV. Each curve corresponds to a different sensor thickness. [45]

3.1.1.2 Charge generation

When charged particles lose energy by ionization or a photon is absorbed by a semiconductor detector, part of the released energy is used to generate charge carriers in the form of electron-hole (e/h) pairs. In order to create an e/h pair, at least the bandgap energy (in silicon $\Delta E_{\mathrm{gap}} = 1.1$ eV) is should be provided. However, since silicon is an indirect band-gap semiconductor, due to momentum conservation, part of the energy is absorbed by lattice excitations called phonons. The average energy w_i required to create an e/h pair is temperature dependent. For silicon in 300 K, $w_i = 3.65\,\mathrm{eV}$ which is about three times higher than the band-gap energy. For a given amount of deposited energy, the average number of e/h pairs generated in silicon is $N = E/w_i$. In the case of a minimum ionizing particle traversing a silicon sensor, the MPV of the energy loss is approximately 0.28 keV/um, and in first order does not depend on the sensor thickness. Therefore, the average number of produced e/h pairs is given by:

$$N_{\Delta_p} = \frac{\Delta_p}{w_i} = \frac{0.28\,\mathrm{keV/\mu m}}{3.65\,\mathrm{eV}} \cdot d(\mathrm{\mu m}) \approxeq 80\frac{e/h}{\mathrm{\mu m}} \cdot d(\mathrm{\mu m}) \tag{3.2}$$

where d is the sensor thickness.

For each event, the deposited energy is randomly split between an amount used to generate e/h pairs and an amount that is absorbed by the lattice. Therefore, the number of generated e/h pairs is subject to fluctuations that usually follow a Poisson distribution, which for a large number of processes (>10) resembles a Gaussian distribution. Thus, the standard deviation of the carrier number ($N_{\mathrm{e/h}}$) fluctuation is equal to $\sigma_{\mathrm{e/h}} = \sqrt{N_{\mathrm{e/h}}}$. A special case is the complete absorption of a particle, for example an X-ray photon. Under this constraint, the standard deviation is given by:

$$\sigma_{\mathrm{e/h}} = \sqrt{N_{\mathrm{e/h}} \cdot F} \tag{3.3}$$

where F is called the Fano factor [35] and determines the ultimate limit of energy resolution for semiconductors. F depends on the material and temperature and for silicon is equal to $F_{\mathrm{Si}} \approx 0.115$. Therefore, the energy resolution is better than expected from the sheer number of fluctuations.

3.1.1.3 Charge transport

The generated e/h pairs are subsequently transported and collected, inducing a signal at the collection node. There are two mechanisms of charge transport in semiconductors, diffusion and drift [35]. Owing to thermal motion, a gradient in the concentration of charge carries leads to a diffusion current which for the density n and p of the electrons and holes, respectively, takes the form:

$$\vec{j}_{n,\mathrm{diff}} = -eD_n\vec{\nabla}n, \qquad \vec{j}_{p,\mathrm{diff}} = -eD_p\vec{\nabla}p \tag{3.4}$$

where $D_{\mathrm{n,p}}$ is the diffusion coefficient for electrons and holes, respectively, which depends on the semiconductor material. Charge transport by diffusion is slow and does not depend on the electric field. Additionally, due to the thermal random walk, the charge carrier path is usually long. As mentioned in section 2.2, diffusion leads to the spread of the charge cloud and limits the minimum useful pixel size.

In the presence of an electric field, electrons and holes (free charge carriers) are accelerated in the direction of the field while scattering off lattice phonons and crystal defects. In this case the equation of motion for the average electron and hole movement is described by the drift velocity which is equal to:

$$\vec{v}_{\mathrm{Dn}} = -\mu_n\vec{E}, \qquad \vec{v}_{\mathrm{Dp}} = -\mu_p\vec{E} \tag{3.5}$$

where μ_n and μ_p is the mobility of electrons and holes respectively. For low fields, the mobility is relatively constant. However, for high fields (approximately higher than $10^5\,\mathrm{V/cm}$) it gradually degrades leading to drift velocity saturation (v_{sat}). In the case of silicon, typical values for the electron and hole mobility are $\mu_n(Si) = 1\,450\,\mathrm{cm^2/Vs}$ and $\mu_p(Si) = 500\,\mathrm{cm^2/Vs}$ respectively while the saturation velocity is equal to $v_{\mathrm{sat}} \approxeq 10^7\,\mathrm{cm/s}$. The drift motion of charge carriers leads to a drift current $\vec{j}_{\mathrm{drift}} = \frac{1}{\rho}\vec{E}$, where ρ is the electric resistivity given by:

$$\rho = \frac{1}{e(n\mu_n + p\mu_p)} \tag{3.6}$$

For a doped semiconductor, depending on the doping type, the charge carrier density (n,p) is usually defined by the doping concentration (N_A,N_D) that is usually significantly higher than the intrinsic carrier density. As will be discussed in the following, the resistivity is an important parameter of the sensor that affects the efficiency of charge collection.

Fast charge collection is crucial for high rate and high radiation applications. In the case of the LHC (p-p), all charge carriers have to be collected in a time slot of 25 ns, that is only possible by drift transport. Furthermore, the charge carriers should be collected as fast as possible to avoid signal loss due to trapping or recombination, especially after irriadiation because the induced bulk damage increases the trapping probability.

3.1.1.4 The p-n junction and sensor depletion

The number of intrinsic or extrinsic (by doping) charge carriers in a silicon crystal is orders of magnitude higher that the amount of charge generated by traversing particles. Therefore, in order to construct a practical sensor, an area depleted of free carriers combined with an electric field to accelerate the generated charges is required. This can accomplished by a reverse biased p-n junction which is the basic building block of silicon particle sensors. The structure of an (abrupt) p-n junction is illustrated in Fig. 3.4 (a). It is formed by bringing in contact an n-doped and a p-doped silicon crystal. At the boundary between the two types, the majority carriers diffuse to the opposite part where they recombine. Therefore, a space charge region depleted of free carriers is formed. The positively charged donor ions and negatively charged acceptor ions that remain in the n-type and p-type regions respectively cause an electric field to build up across the junction. The majority charge carrier diffusion is opposed by the generated electric field and the junction reaches an equilibrium state.

The electric field is maximum at $x = 0$ and is given by [35]:

$$E(x) = \begin{cases} \frac{-eN_A}{\epsilon_{\mathrm{Si}}\epsilon_0}(x + x_p)\,, & -x_p < x < 0\,, \\ \frac{+eN_D}{\epsilon_{\mathrm{Si}}\epsilon_0}(x + x_n)\,, & 0 < x < x_n \end{cases} \tag{3.7}$$

where x is the distance from the junction boundary, x_n and x_p are space charge region widths and N_D and N_A are the donor and acceptor doping concentrations. The potential difference across the junction, called the built-in potential V_{bi}, is expressed by [35]:

$$V_{\mathrm{bi}} = \frac{k_B T}{e} ln\left(\frac{N_D N_A}{n_i^2}\right) \overset{Si}{\approx} 0.4 - 0.8\,\mathrm{V} \tag{3.8}$$

where k_B is the Boltzmann constant, T is the sensor temperature and n_i is the intrinsic carrier concentration.

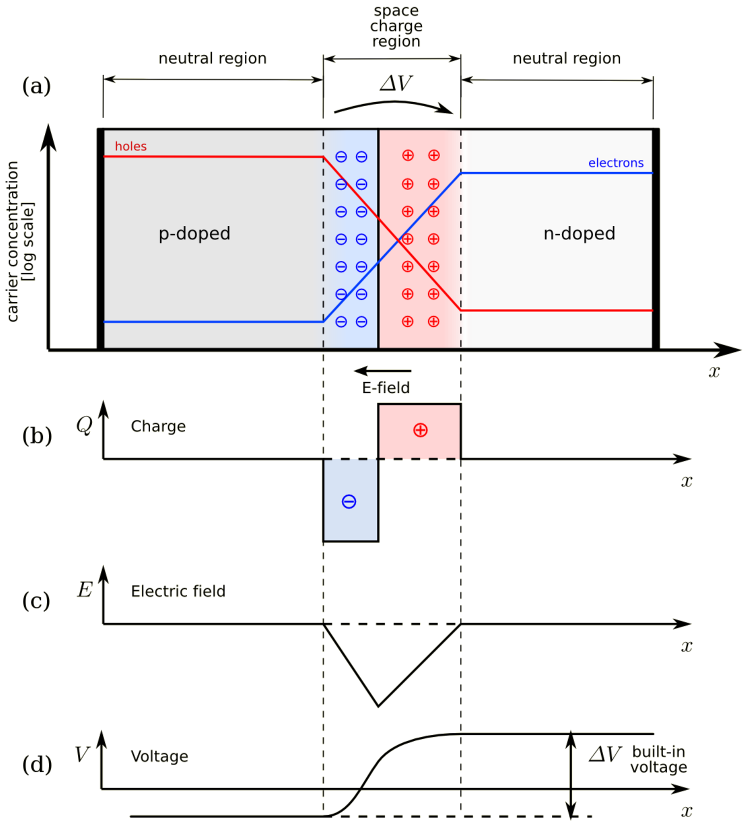


Figure 3.4: The p-n juction: a) structure, b) space charge density, c) electric field distribution and d) potential distribution. [46]

In the case of a particle detector, the doping of one side of the junction is very high while the other side is lightly doped. Thus, the depletion region grows from the junction into the more lightly doped bulk. Since the electron mobility is about 3 times higher than hole mobility, highly doped n-type (n^+) electrodes are usually preferred for fast charge collection. Electrons generated in the depleted region of the bulk are accelerated by the drift field towards the n^+ collection electrode while holes flow in the opposite direction. It is desirable to extend the depletion region to the full sensor volume (if possible) in order to collect a large number of charges by drift and increase signal to noise ratio (SNR). The depleted volume can be extended with the help of an externally applied reverse bias voltage $V_{\text{ext}} = V_{\text{dep}} + V$, where V_{dep} is the voltage that needs to be applied for full depletion. The depletion depth d is in this case equal to [12]:

$$d = x_n + x_p \cong x_p = \sqrt{\frac{\epsilon_{\text{Si}}\epsilon_0}{e}\frac{1}{N_A}(V_{\text{bi}} + V_{\text{ext}})} \cong \sqrt{\frac{\epsilon_{\text{Si}}\epsilon_0}{e}\frac{1}{N_A}V_{\text{ext}}} \tag{3.9}$$

Usually sensors are operated in over-depletion that results in a non-zero field at the back side which is beneficial for fast charge collection (Fig. 3.5). The doping concentration of the bulk (N_A) is inversely proportional to the resistivity ρ (eq. 3.6) of the bulk. Therefore, large depletion can be achieved by: 1) high resistivity (lightly doped) bulk and 2) high reverse bias voltage:

$$d \propto \sqrt{\rho V_{\text{ext}}} \tag{3.10}$$

Even though a high reverse bias voltage is desirable, if it becomes sufficiently large, breakdown will occur resulting in a large current flow though the p-n junction. The breakdown voltage depends on the doping concentration and the layout of the sensor. In order to increase it, layout techniques such as multiple guard rings and smooth edges are often employed.

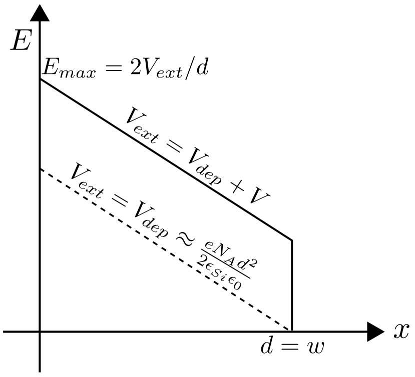

E
$E_{max} = 2V_{ext}/d$
$V_{ext} = V_{dep} + V$
$V_{ext} = V_{dep} \approx \frac{eN_A d^2}{2\epsilon_{Si}\epsilon_0}$
x
$d = w$

Figure 3.5: Electric field shape of a silicon sensor where the depletion depth d is equal to the sensor thickness w. The dashed line corresponds to the case of full depletion while the solid line corresponds to the case of over-depletion.

One of the most critical properties of the sensor is its capacitance as it enters the SNR and rise time equations and has a significant impact of the performance and power consumption of the detector as will be discussed later on. The reverse biased p-n junction can be considered as a parallel plate capacitor with the depletion (space charge) region determining the distance between the plates. The sensor capacitance C normalized to the junction area A can be expressed by:

$$\frac{C}{A} = \frac{\epsilon_{\text{Si}}\epsilon_0}{d} \propto \frac{1}{\sqrt{\rho V_{\text{ext}}}} \tag{3.11}$$

Even in the absence of carrier generation due to energy loss from charged particles, there are thermally generated e/h pairs in the depleted region that are strongly influenced by the presence of impurities which act as generation/recombination centers. These carriers generate a constant leakage current that is proportional to the depleted volume of the sensor $A \cdot d$ and is equal to [35]:

$$I_L^{vol} = eAd\frac{n_i}{\tau_g} \tag{3.12}$$

where τ_g is the carrier generation lifetime. The intrinsic carrier concentration strongly depends on temperature ($n_i \propto T^{3/2}$), hence I_L is also temperature dependent. The leakage current increases noise (shot noise) and impacts the operating point of the electronics, which often have to include leakage compensation circuitry. Although the volume leakage current I_L is the most significant, there are smaller contributions from other sources such as the reverse saturation current I_S generated by minority carrier diffusion.

3.1.1.5 Signal formation

As mentioned, a pixel detector consists of multiple segmented electrodes in a 2D plane. The movement of the separated charges (e,h) in the electric field of the sensor induces a signal on one or more electrodes that is described by the Shockley-Ramo theorem [35]. A signal is already detectable the moment the charge carriers start moving. The instantaneous current induced by a single charge carrier ($q = e$) with drift velocity v is expressed by:

$$i_{\mathrm{e/h}} = -\frac{\mathrm{d}Q}{\mathrm{d}t} = e\vec{E}_w \cdot \vec{v} \tag{3.13}$$

The weighting field $\vec{E}_w$ is a property used to determine how the charge movement couples to a specific electrode depending on the electrode configuration and has nothing to do with the electric field inside the sensor. The signal that id detected by the electronics (integrating amplifier) is proportional to the total charge Q induced on an electrode:

$$Q = \int_{t_1}^{t_2} i(t)\,\mathrm{d}t \tag{3.14}$$

where $t_2 - t_1$ is the integration time. When the movement stops, a net signal (Q) will appear only at the electrodes that collected the charge.

Due to the spread of the charge cloud by diffusion (or in the case of tilted tracks) the total charge will be shared by more than one electrodes. If charge sharing is significant, it can impact the performance of the detector since the SNR of the individual pixels for the same total generated charge will be decreased. In the worst case (charge generated at the intersection of 4 neighboring pixels) the signal is divided by four. Therefore, the electronic circuit (pre-amplifier) noise should be low enough to allow for the detection of a hit at least for the pixel with the highest signal. Since charge sharing depends on the sensor thickness, a thin sensor will be less affected from it, but will generate a lower signal compared to a thicker sensor (assuming full depletion). A carefully designed system (sensor & electronics) can even benefit from charge sharing to improve position resolution (refer to section 2.1.1).

3.1.1.6 Radiation effects - bulk damage

A particle detector is exposed during its operation to an intense flux of charged and neutral particles as well as $\gamma$and X-rays. At the HL-LHC, about 50 particles will traverse every silicon lattice cell during its operation lifetime close to the interaction point. Particles that collide directly with the atomic nuclei of the crystal lattice can dislocate atoms from their position, creating vacancies and interstitials in between the lattice atoms, damaging the silicon crystal. This is a Non-Ionizing Energy Loss (NIEL) process and the damage caused to the lattice is usually referred to as substrate (bulk) NIEL damage. Some of these primary defects are illustrated in Fig. 3.6. The primary released atoms can have enough energy to kick-off other atoms before they come to rest, which can also dislocate additional atoms creating regions with high defect density called defect clusters.

NIEL damage depends on the particle type and energy. Charged particles create more single point defects (and smaller clusters) than neutrons due to Coulomb scattering off lattice nuclei. To be able to easily quantify and compare the damage caused by particle of different types and energies, the displacement damage is normalized to the damage effect of 1 MeV neutrons and described in terms

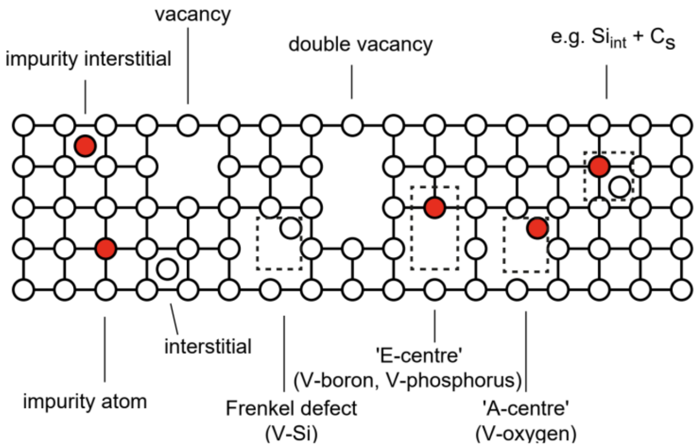


Figure 3.6: Defect types in a silicon lattice caused by traversing particles. Additional types of defects can form when vacancies and interstitials are close to impurity atoms (red circles). [35]

of neutron equivalent fluence (n_{eq}/cm^2). The scaling parameter κ defined as the ratio of the NIEL damage of a particle type with specific energy to the NIEL damage of 1 MeV neutrons is called the hardness factor.

The defects caused by NIEL damage to the substrate, result in new energy levels within the silicon band-gap that may become electrically active and cause (depending on their type): a) generation of donor and acceptor centers, b) leakage current through generation-recombination centers and c) creation of trapping centers. Therefore we identify three main implications of bulk damage to the operation of an Si detector [35]:

(1) *Change of the substrate doping:* The effective doping concentration N_{eff} of the detector bulk is influenced by two mechanisms: the generation of donor or acceptor like states and the deactivation of donor or acceptor atoms (removal). The doping concentration change is a function of the radiation fluence and can even lead to type inversion in which initially n-type silicon becomes effectively p-type after irradiation as shown in Fig. 3.7. The bias voltage required for full depletion of the detector volume (V_{dep}) follows the change of N_{eff} and becomes minimum at the inversion point. After type inversion (for a p$^+$-in-n sensor), the junction boundary moves to the opposite side and depletion grows from the non-pixellated bottom. Sensors fabricated from p-type substrate material generally do not type-invert.

(2) *Leakage current:* Deep level defects created near the middle of the band-gap act as generation-recombination centers that increase the probability of e/h pair thermal generation. Thus, the leakage current increases proportionally to the radiation fluence. Apart for increasing electronic noise, leakage current also heats up the detector. Proper cooling of the sensor is important to avoid thermal runaway which is a positive feedback loop between the leakage current and the generated heat and can ultimately even destroy the detector.

(3) *Charge trapping and collection efficiency:* Electrons and holes can be trapped by defect levels and be released again after some time or even be lost by recombination at a deep level. This results in decreased carrier lifetime and mean free path length. If the detraping time constant is longer than the time of signal formation, charge is being effectively lost resulting is a reduced

signal amplitude. Charge loss due to trapping is described by the help of the effective time $\tau_{\mathrm{eff,e/h}}$:

$$Q(t) = Q_0 \exp\left(-\frac{t}{\tau_{\mathrm{eff,e/h}}}\right) \tag{3.15}$$

The Charge Collection Efficiency (CCE) is an important metric of the sensor performance and is defined as the ratio of produced ionization charge to the amount of collected charge. In silicon, CCE is close to 100% for charge generated in the depleted area of the sensor even for NIEL fluences of $\phi_{\mathrm{eq}} = 10^{15}\, n_{eq}/\mathrm{cm}^2$. In order to achieve high CCE (>97% to comply with the LHC requirements), the sensor should be fully depleted to avoid collection by diffusion that would result in a significant portion of the charge being trapped. For higher fluences, high reverse bias voltage is usually required to maintain full efficiency by creating a strong drift field (over-depletion).

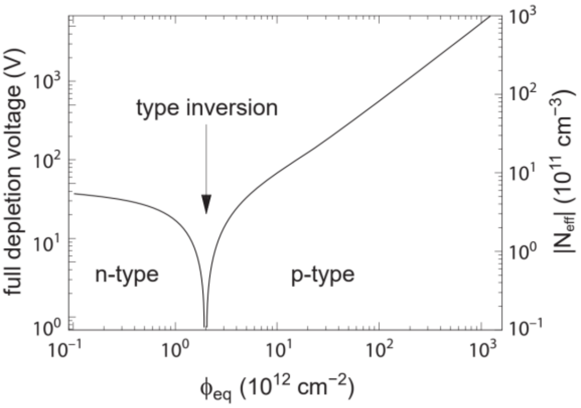


Figure 3.7: Effective charge carrier concentration and full depletion voltage of an initially n-type silicon sensor as a function of normalized fluence ϕ_{eq}. [35]

It has been found that sensors using p-substrate material with electron collection on n^+ electrodes (n^+-in-p) are more radiation tolerant than p^+-in-n configurations. They can also be produced more easily and with lower cost compared to n^+-in-n[1] sensors for which double-side processing is necessary. In order to further increase radiation tolerance, silicon with high oxygen content supplied in the growth process is used that has been shown to reduce the negative effects of bulk damage. The current trend for radiation fluences up to and above $10^{16}\, n_{eq}/\mathrm{cm}^2$ in the context of the HL-LHC upgrade goes to thin (100 – 150 um), high resistivity ($\gtrsim 2\,\mathrm{k\Omega\,cm}$) n^+-in-p sensors operated at high bias voltages (500 – 700 V) and low temperatures [20].

3.1.2 The readout ASIC: signal detection and processing

The signal induced in the collection electrodes by the charge carriers generated in the sensor (typically $< 3\,\mathrm{fC}$) is quite small and therefore sensitive to noise and cross-coupling and is not suitable to be directly transmitted to the Data Acquisition (DAQ) system. A Readout Application Specific Integrated Circuit (ASIC) also refered to as ROIC provides the functionality needed to readout, process, and

[1] Diodes of the same type such an n^+-n configuration also create a potential gradient similar to a p-n structure.

transmit the hit information. It contains dedicated amplification and discrimination channels per pixel (front-end) that operate in parallel followed by parallel processing of the output signals which are most commonly digitized in the pixel. The ROIC technology has been enabled by Moore's Law towards deep sub-micron Complementary Metal Oxide Semiconductor (CMOS) technologies and has evolved to meet the rate and radiation tolerance demands. CMOS ROICs can be classified in generations, which have a rough correlation with the feature size of the process node they are built upon [20].

The first generation of pixel ROIC chips include those used in the original ATLAS and CMS detectors which were typically designed in 0.25 µm CMOS process nodes. They were mainly composed of an analog pixel matrix with full-custom digital circuits to manage hit buffering and readout (analog-on-top design flow). Most of the processing and data buffering was done in the periphery which therefore occupied a relatively large area. The second generation of pixel chips are those currently in operation and are designed in 130 nm CMOS technology. They are characterized by the use of synthesized logic side by side with analog front-ends, enabled by higher logic density, but with an organization in columns and small stepped blocks. The third generation, developed in 65 nm will proceed further in order to meet the challenges of the HL-LHC being essentially complex, high logic density digital chips with embedded analog front-ends in a sea of digital gates (digital-on-top design flow).

A typical structure of a ROIC is illustrated in Fig. 3.8. The pixel chip is usually divided into an active area which is composed of a repetitive matrix of identical pixels, each containing a dedicated front-end, and the chip periphery which controls and supports the active matrix and processes the generated hit data. The active area is organized in columns (first and second ROIC generation) in order to achieve a modular design and optimize signal routing.

A particle impinging at the sensor diode, induces a short current signal at the preampifier input. The preamplifier should be as close as possible to the sensor contact to avoid noise pickup and minimize the influence of stray capacitances. It is usually realized by an active integrator (amplifier with capacitive feedback loop) and generates a voltage proportional to the signal charge. The reset to baseline is accomplished by an element that discharges the feedback capacitor, often realized by a constant current reset circuit. A shaper, which is essentially a band-pass filter with properly tuned time constant, is usually employed to optimize the signal shape and bandwidth in order to reduce noise and pile-up. The discriminator compares the pre-amplifier output signal to a reference voltage V_{ref}. Whenever the signal amplitude exceeds V_{ref}, a digital pulse with width proportional to the amount of collected charge is produced. Typical signal waveforms at the input, the pre-amplifier output and the discriminator output are illustrated in Fig. 3.9. The in-pixel digital readout circuitry which is part of the chip readout architecture processes, stores and transfers data from the pixels to the ROIC periphery. It is important to sufficiency isolate the digital readout circuitry from the sensitive analog part to suppress crosstalk interference due to digital switching. Therefore, the analog and digital domains must have separate power domains and multiple nested well processes must be used in order to reduce substrate noise coupling.

The sensor leakage current, especially after irradiation, is often compensated to avoid pushing the pre-amplifier out of its operating region by shifting the input baseline (if the sensor is DC-coupled to the ROIC). Leakage compensation circuits are realized based on the concept of adjusting a controlled current source (MOSFET) in order to provide a current equal to the sensor leakage. The bandwidth of the control circuit should be very low in order to react only to slow changes, such as leakage, and not affect fast transient currents induced by particles. The front-end readout chain also contains additional features for testability and ease of operation. To test the ROIC functionality, even before the sensor is bump-bonded, and characterize its performance without the need of radioactive sources or particle beams, a small capacitance driven by a pulsing circuit is used to inject a configurable amount of charge to the input. Additionally, the possibility of masking the discriminator output is included to avoid readout saturation by excessive noise hit rate from certain pixels due to manufacturing defects.

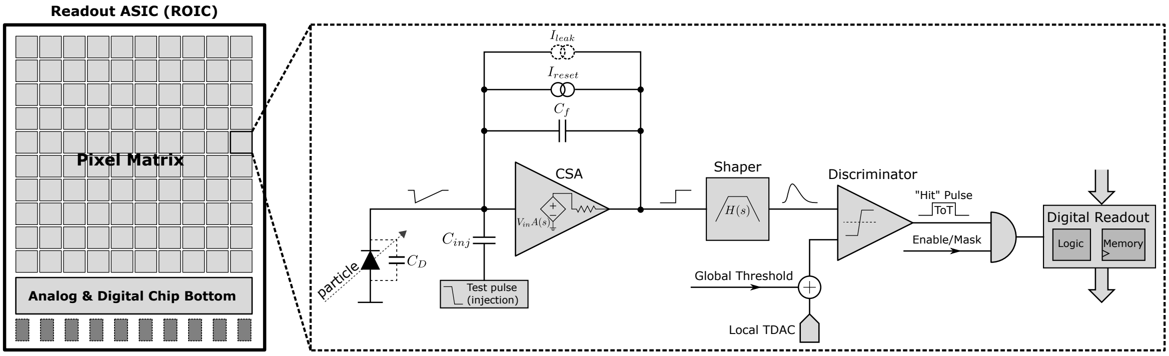


Figure 3.8: A typical front-end readout scheme often used in a detector readout ASIC. The signal processing chain includes amplification, shaping, discrimination and finally digitization.

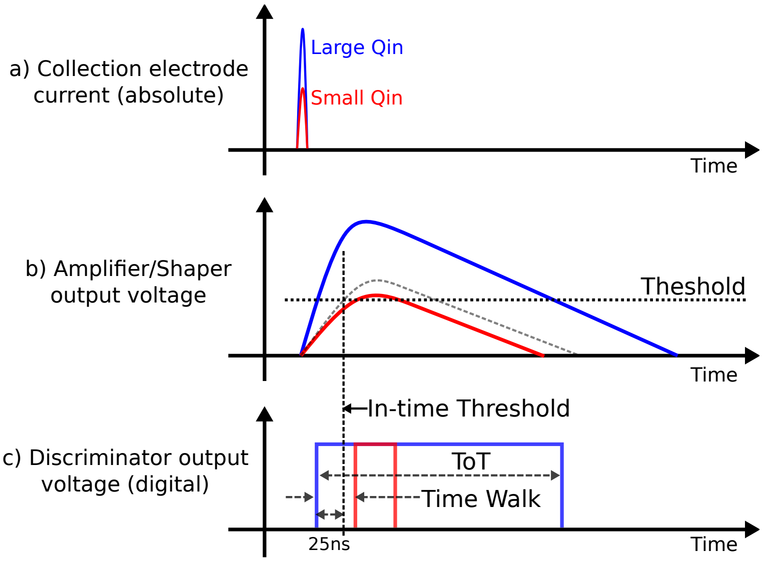


Figure 3.9: Typical signal waveforms of a) the current induced at the collection electrode, b) the preamplifier output voltage and c) the discriminator output voltage for a small (red) and a large (blue) deposited energy.

3.1.2.1 Performance criteria and design aspects

The ROIC must provide low noise amplification, fast hit discrimination and an efficient, high-speed readout architecture, while consuming as low power as possible. Furthermore, robust operation with a well-defined threshold and high radiation tolerance should be guaranteed. The following performance criteria and design aspects are mainly considered in the development and testing of a readout ASIC.

- *Noise - equivalent noise charge*: Electronic noise from the sensor diode and the pre-amplifier transistor (or other) devices result in time-varying voltage fluctuations at the pre-amplifier output. The noise performance of the pre-amplifier (first) stage is the most crucial since after amplification, the signal amplitude is high compared to additional noise introduced by the following stages. Voltage noise can be described by the Root Mean Square (RMS) of the voltage fluctuations at a specific node. A more intuitive quantity is often used to quantify the noise of a readout channel, called equivalent noise charge (ENC), which is defined as:

$$\mathrm{ENC} = \frac{\text{noise output } v\text{oltage (V)}}{\text{output } v\text{oltage of a signal of } 1\, e^{-}\ (\mathrm{V}/e^{-})} = \frac{\sqrt{\langle v_{\mathrm{out}}^{2}\rangle}}{\text{gain } (\mathrm{V}/e^{-})} \tag{3.16}$$

Noise is mainly influenced by the total input capacitance and the system bandwidth (speed). If the bandwidth is constant, noise can be reduced by increasing the input transistor transconductance (and consequently power consumption). The ENC can be directly calculated by the Cumulative Distribution Function (CDF) (s-curve) obtained from the discriminator "hit" pulse response to multiple charge injections as shown in Fig. 3.10. The threshold is defined as the charge where 50% of the injections fire the discriminator and the ENC is determined from the slope s at this point [5]:

$$\mathrm{ENC} = \frac{1}{\sqrt{2\pi}}\frac{1}{s} \tag{3.17}$$

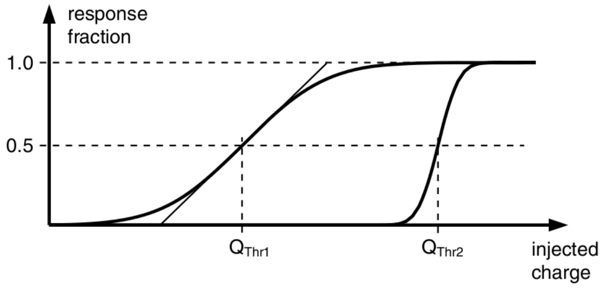


Figure 3.10: Cumulative distribution function (s-curve) of the discriminator response for a channel with high noise at a low threshold and for a channel with low noise at a higher threshold. The y-axis represents the fraction of hits for a given amount of injections at a specific input charge value. [5].

- *Operating hreshold and threshold variations*: The discriminator threshold (Fig. 3.9) is set to a value as low as possible in order to maximize the detection efficiency but not too low in order to keep the noise hits at an acceptable level and adhere to the noise hit rate specification. Therefore the operating threshold value (in e^{-} units) is ideally located between the noise peak around the baseline and the signal distribution. Another important consideration is that possible crosstalk from digital switching can activate a positive feedback mechanism that sets off a number of chain reactions causing all pixels to fire. Therefore, the minimum threshold should be high enough to guarantee stability for a given hit occupancy. The threshold of individual pixels is subject to random variations due to component mismatch and possibly systematic variations due to voltage drops along the power distribution network and layout patterns. Therefore, apart from the global threshold, the possibility to fine tune the threshold locally per pixel is usually included in order to compensate these variations and is often realized by simple digital to analog converters (DACs). The dispersion of threshold after tuning is inversely proportional to the tuning DAC (TDAC) number of bits:

$$\sigma_{\mathrm{THR}_{\mathrm{tuned}}} \approx \frac{\sigma_{\mathrm{THR}}}{2^{n_{\mathrm{TDAC}}}} \tag{3.18}$$

where σ_{thr} is the RMS of the threshold spread (threshold dispersion) before tuning. Typically, a TDAC resolution of 3-7 bits is used depending on the application [5].

The front-end noise sets a lower bound to the threshold, but does not determine the stable operating threshold value which depends additionally on how the threshold varies with time and from pixel to pixel. The operating threshold lower limit should be at least approximately 5-6 times higher than the combined standard deviation of noise and threshold dispersion:

$$Q_{\mathrm{THR}} = 5 - 6 \cdot \sqrt{\sigma_{\mathrm{THR}}^2 + \mathrm{ENC}^2} \tag{3.19}$$

- *Timing response and in-time threshold*: The timing precision with which the arrival of a hit can be determined is crucial in experiments such as the LHC where hits must be associated with one particular bunch crossing with a precision ≤ 25 ns. Apart from the charge collection time, the timing response depends on the pre-amplifier rise time, the shaper bandwidth and the discriminator speed. Furthermore, it is a function of the input charge and follows the so called "time-walk" curve shown in Fig. 3.11. Hits with high amplitude lead to a fast response. In contrast, hits with amplitudes close to the threshold require a longer time to be detected (Fig. 3.9). The response time for a sufficiently large charge is aligned to the collision event time (beginning of the BCID 25 ns interval) by adjusting the global BCID delay. A time window $\varDelta T < 25$ ns (to allow for jitter and systematic delays across the column) is therefore selected on the time-walk curve that fixes an interval $\varDelta Q$ of charges which are detected in-time. The lowest amplitude that can still be associated with the correct BCID is defined as the in-time threshold and is higher than the operating threshold.

If the charge signal amplitude information is available (non-binary readout), hit timing can in principle be corrected on the basis of the known time-walk characteristic. Time-walk correction requires additional data processing and its accuracy is limited by the precision of charge measurement and time-walk variation from pixel to pixel. If the shape of the pulses does not vary with time, time-walk can be avoided by employing techniques such as zero-crossing triggering or constant fraction discrimination [35]. However these methods require complex circuitry that occupies large area in the pixel and are used for applications demanding high timing precision.

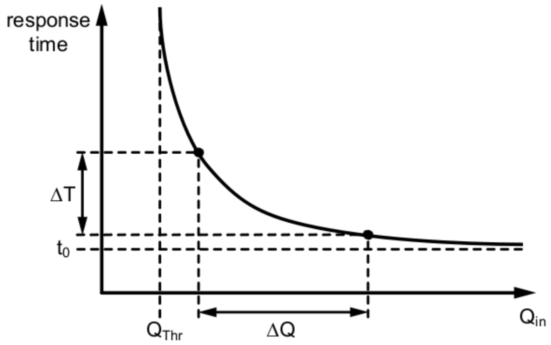


Figure 3.11: Timing response of a detector signal processing chain with discrimination, also refered to as the time-walk curve. [5].

- *Analog charge measurement*: Analog charge measurement is desirable in order to improve the spatial resolution (see chapter 2), for time-walk correction and for monitoring purposes. A straight-forward approach is to sample, store and readout the peak signal amplitude or convert it to a digital word by an Analog to Digital Converter (ADC). However, since a resolution of a few bits is adequate in most cases, charge information is usually obtained by measuring the width of the discriminator output signal, also called Time-Over-Threshold (ToT), as illustrated in Fig. 3.9. The ToT method is essentially a Time to Digital Conversion (TDC) using the BCID time stamp to measure the number of clock cycles during which the signal is higher than the discriminator threshold. The BCID time stamp is latched into local registers when the leading edge and trailing edge transitions are detected, and the ToT is obtained by their difference. Since the leading edge time stamp is already used to measure the Time of Arrival (ToA) of an event producing a hit, the added complexity is relatively small. The ToT should ideally be an almost linear function of the input charge, that is usually achieved by a constant current reset mechanism. The discharge current that determines the ToT slope is adjusted according to the desirable resolution, the BCID time stamp frequency and number of bits and the readout architecture capabilities.

- *Dead time and readout efficiency*: For every readout channel there is a minimum amount of time after registering a signal for which the detector can no longer process new hits. The so-called "dead-time" (τ) is caused by both the analog processing chain, e.g. the recovery time required by the pre-amplifier to discharge the feedback capacitor and the digital readout architecture that requires a certain amount of time to register and process the hit information. Dead time is important since it limits the maximum possible event rate the ROIC can process without significant hit losses. In order to theoretically estimate the dead time influence on the readout efficiency for a given average rate of events uniformly distributed in time, two special cases can be distinguished: a) a system with updating dead-time (paralyzable) and b) a system with a fixed dead-time interval (non-paralyzable). An example of a paralyzable system is a pre-amplifier with recovery time while digital hit processing is an example of a non-paralyzable system.

  For low event rates compared to $1/\tau$, which is usually the case, the hit loss fraction is approximated for both cases by $n\tau$ where n is the event rate and τ is the system dead time. Therefore, the number of measured events is equal to $m = n(1 - n\tau)$. The dead-time τ is generally not constant since it varies with the signal amplitude and depends on the waiting time distribution of the readout architecture. However, the average dead-time can be used instead to estimate the readout efficiency. As an example, the influence of the ToT measurement time on the readout efficiency is explored, assuming all other contributions negligible, since for a well designed system it can be dominant. For an average ToT of 8 in bunch crossing units (25 ns), and 75 kHz event rate per pixel, the hit loss due to ToT would be: $n\tau = 75\,\mathrm{kHz} \cdot 200\,\mathrm{ns} = 1.5\%$.

- *Power consumption*: The maximum allowed power consumption depends on the cooling system capabilities and power delivery scheme. Although a change in power consumption does not linearly translate to higher material due to recent advancements in CO_2 cooling, low power is important especially for high granularity detectors. The analog power consumption is determined by the signal to noise ratio and timing requirements and a for a given CMOS technology it scales with the detector capacitance (refer to section 3.1.2.4). The digital power consumption depends on the logic complexity, the switching frequency the signal rise and fall times. One of the most important contributors to digital power consumption (for synchronous architectures) is the distribution of the BCID clock across the active pixel matrix, due to the large associated capacitance of the distribution lines.

- *Radiation tolerance*: The readout electronic circuitry is affected by ionizing radiation in two different ways. Surface damage (more details in section 3.1.2.6) alters the transistor characteristics (threshold, transconductance) and can lead to high leakage currents or even activate parasitic transistor structures. In addition to long term degradation due to the accumulated dose, charge deposition by strongly ionizing particles can lead to single event effects (SEE). A single event upset (SEU) occurs if the charge happens to be deposited close to a storage node (RAM cell, register) and causes a bit value to flip leading to corrupt information. Single event upsets are usually mitigated by using techniques such as redundant memory (e.g. a DICE cell [5]), spatially separated triplicated circuits and error correcting codes. Furthermore, if the deposited charge is high enough it can trigger parasitic thyristor structures, intrinsic in CMOS technologies, and lead to a single event latchup (SEL) that causes short-circuit currents. SEL can be prevented by careful placement of frequent well taps (contacts) in order to achieve low resistance and a stable well potential.

3.1.2.2 Noise Sources

Electronic noise[2] is a result of stochastic fluctuations in the number N and velocity v of charge carriers and is usually quantified by the variance (σ^2) or RMS value (σ) of voltage and current signals. Often, it is useful to describe noise by means of a frequency spectrum using its spectral power density. Depending of the physical process, the following noise sources can be distinguished [35]:

- *Thermal noise*: Thermal noise has its origin in thermal velocity fluctuations of charge carriers due to their thermal kinetic energy. The noise spectral power density of the current flowing through a conductor with resistance R and temperature T is equal to:

$$\mathrm{d}\langle i^2\rangle_{\mathrm{therm}} = \frac{1}{R} 4kT\,\mathrm{d}f \tag{3.20}$$

[2] In this section only inherent noise sources are considered and external noise due to crosstalk and electromagnetic interference is neglected.

where k is the Boltzmann constant. Thermal noise is independent of the frequency (white noise) and the current flowing through R.

- *Shot noise*: Shot noise is a statistical fluctuation occurring when charge carriers are emitted independently of each other over a potential barrier, as for example in the case of e/h pair generation and recombination in semiconductors. The noise spectral power density in this case is given by:

$$\mathrm{d}\langle i^2\rangle_{\mathrm{shot}} = 2eI_0\,\mathrm{d}f \tag{3.21}$$

  where I_0 is the average current in the noisy system. Shot noise is directly proportional to I_0 and has a white noise spectrum. In the case of semiconductor detectors, the sensor diode (pn junction) can be modeled as an ideal diode with a parallel shot noise current source whose power depends on the leakage current I_0. Therefore its contribution to the detector system noise can be significant after irradiation.

- *Flicker (1/f) noise*: Flicker noise can, in a fairly general sense, describe all noise contributions featuring a non-white frequency spectrum according to $1/f^a$ with $a = 0.5...2-3$. In electronic systems, flicker noise is caused by capture/release processes with different time constants such as charge carrier trapping near the gate-silicon interface of MOS transistors. The spectral density for a single trapping process approximates a $1/f^2$ behavior and superposition of only a few $1/f^2$ spectra with different time constants yields an $1/f$ dependence. The noise power density is generally written as:

$$\mathrm{d}\langle i^2\rangle_{\mathrm{flicker}} = K_\alpha \frac{1}{f^\alpha}\,\mathrm{d}f \tag{3.22}$$

  where K_α is a constant that depends on the specific process.

It is important to determine the noise in the case of MOSFET (Metal Oxide Semiconductor Field Effect Transistor) transistors since they are the building blocks of modern CMOS circuits and their performance is crucial in the case of the sensitive pre-amplifier input stage. The output noise of a MOSFET can be described by two parallel noise current sources, due to thermal and flicker noise. In order to simplify the noise analysis of a detector/pre-amplifier system, it is useful to express these noise contributions as an effective voltage noise source in series with the input (gate). The conversion can be performed using the transistor transconductance g_m as follows:

$$\langle i_D^2\rangle = \langle (g_m v_{\mathrm{in}})^2\rangle \tag{3.23}$$

where I_D is the drain current. For a MOSFET in saturation, the equivalent input series noise is given by [35]:

$$\frac{\mathrm{d}\langle v_{\mathrm{in}}^2\rangle}{\mathrm{d}f} = 4kT\frac{2}{3}\frac{1}{g_m} + K_f\frac{1}{C_{\mathrm{ox}}^2 WL}\frac{1}{f} \tag{3.24}$$

The first term describes the equivalent thermal noise voltage that results by replacing the channel conductance g_0 by $\frac{2}{3}g_m$. The second term describes the flicker noise where C_{ox} is the gate oxide capacitance per unit area and K_f is the 1/f noise constant that depends on the technology. K_f is 10 to 25 times lower for PMOS compared to NMOS transistors due to the lower trapping probability of holes. Apart from the transistor technology and type, flicker noise is inversely proportional to the gate area (W/L). When selecting the input transistor type for optimum noise performance the noise comparison ratio $\langle v_{\mathrm{th}}^2 \rangle / \langle v_{1/f}^2 \rangle$ should be considered taking into account the system bandwidth. If flicker noise is dominant, a PMOS input transistor is preferred while if the thermal noise is dominant an NMOS input transistor is prefered due to its higher g_m.

While flicker noise of a MOSFET is well characterized by the appropriate SPICE (Simulation Program with Integrated Circuit Emphasis) models, for small devices the impact of trapping/detrapping events caused by individual defects at the Si/SiO$_2$ interface shows discrete current fluctuations called burst or Random Telegraph Signal (RTS) noise. The spectrum of RTS noise has a Lorentzian shape characterized by a plateau at low frequencies and $1/f^2$ roll-off at higher frequencies as shown in Fig. 3.12. Depending on the nature of the trap, its proximity to the Si/SiO$_2$ interface and the carrier emission and capture time constant, high amplitude drain current steps (conductivity change) due to RTS can occur [47]. Due to CMOS process downscaling and inaccuracies or even absence of RTS noise in vendor SPICE models, it can become an issue in CMOS readout circuits. To reduce the effects of RTS noise, the critical transistors (usually the input MOSFET) have to be identified and their gate area has to be increased until the noise is sufficiently suppressed.

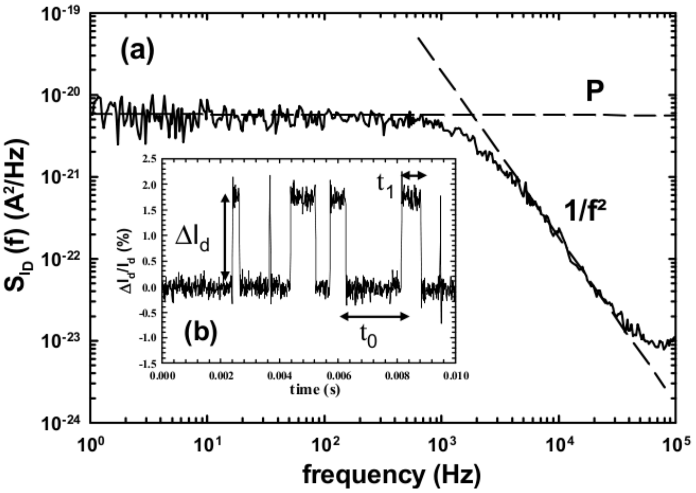


Figure 3.12: a) RTS noise power spectral density measurement example, b) corresponding drain current discrete fluctuations [48].

3.1.2.3 The charge sensitive pre-amplifier

The pre-amplifier implementation depends on the application, the detector capacitance and the optimization of the specified performance metrics. Voltage-sensitive amplifiers are often used when the voltage input signal $V_{\mathrm{in}} = Q_s/C_D$, resulting from the integration of the signal current on the sensor capacitance C_D, is sufficiently large and fast rise times are pursued. An example of a voltage amplifier optimized for small sensor capacitance is the TJ-Monopix front-end pre-amplifier presented on chapter 4. Transimpedance amplifiers are often found in applications requiring high bandwidth such as the acquisition of the signal waveform in transient current technique (TCT) measurements [35]. In the majority of particle detector applications, where the voltage input signal is relatively small, a charge sensitive amplifier (CSA) is preferred.

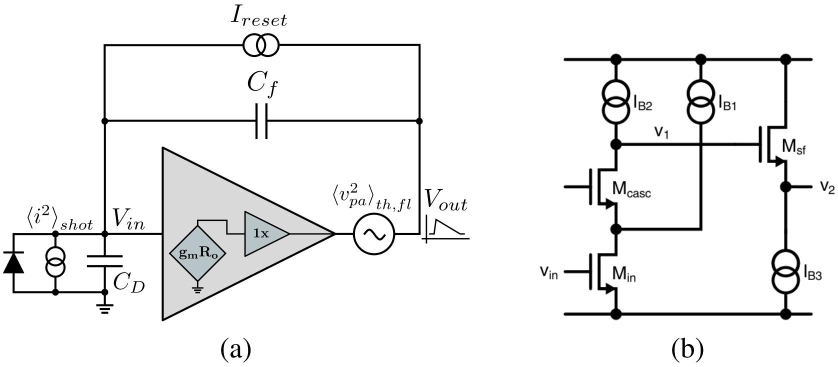


Figure 3.13: a) Charge sensitive pre-amplifer with capacitive feedback, b) Typical straight cascode gain stage implementation followed by a source-follower buffer.

A Charge Sensitive Amplifier (CSA) is realized by a capacitive feedback loop as shown in Fig. 3.13 (a). Its operation principle is the integration of the generated charge (Q_f) on the feedback capacitor C_f, which results in an output voltage approximately equal to:

$$v_f = v_{\text{out}} - v_{\text{in}} = \frac{Q_s}{C_f} \tag{3.25}$$

where Q_s is the signal charge. In order to "force" the input current through the feedback capacitor, a low input impedance is created by the negative feedback loop consisting of the capacitor C_f and an amplifier with open loop gain A_0. According to the miller effect [49], the equivalent capacitance seen at the amplifier input node is equal to $C_f(1 + A_0)$. Therefore, quasi-statically, the signal charge is shared between the sensor capacitance C_D and the miller capacitance $C_M = C_f(1 + A_0)$ yielding a residual charge at the input equal to:

$$Q_{\text{res}} = v_{\text{in}} C_D = Q_s \frac{C_D}{C_D + C_M} \tag{3.26}$$

The feedback capacitor should be small enough to achieve a high charge to voltage conversion gain and the amplifier gain should be sufficiently high to minimize the residual charge. Taking into account the finite gain of the amplifier, the charge to voltage conversion gain is given by [5]:

$$G_Q = -\frac{Q_s}{C_f} \cdot \frac{1}{1 + \frac{1}{A_0} + \frac{C_{\text{in}}}{A_0 C_f}} \tag{3.27}$$

where C_{in} is the sum of the detector capacitance C_D, the amplifier input capacitance C_{amp} and parasitic contributions C_{par}.

A typical CMOS implementation of the amplifier used in CSA circuits is shown in Fig. 3.13 (b). The amplification stage consists of the input transistor M_{in} and the cascode transistor M_{casc} that increases the output resistance. The voltage at the high impedance node (V_1) is equal to:

$$v_1 = v_{\text{in}} g_{m_{\text{in}}} R_{o_1} \tag{3.28}$$

where $g_{m_{\mathrm{in}}}$ is the transconductance of the input transistor and R_{o_1} is the output resistance at the node V_1. M_{sf} acts as a source follower with gain ≅ 1 and low output resistance in order to drive the output V_2 of the amplifier. A real amplifier has a limited bandwidth resulting in finite rise times. The amplifier gain $A(\omega)$ is frequency dependent and the dominant pole that arises from the high impedance node V_1 has a frequency equal to $\omega_c = 1/R_{o_1}C_{o_1}$ where C_{o_1} is the total capacitance at the node V_1. At low frequencies the input impedance Z_{in} of the CSA is purely capacitive (miller capacitance C_M) while at high frequencies the amplifier behaves like a resistor. The CSA time constant is defined by the product of the input capacitance $C_{\mathrm{in}} \cong C_D$ and the input impedance at high frequencies and is given by [35]:

$$\tau_{\mathrm{CSA}} \cong \frac{C_D}{C_f}\frac{C_{o_1}}{g_{m_{\mathrm{in}}}} \tag{3.29}$$

In order to calculate the ENC we consider the noise sources $\langle v_{\mathrm{pa}}^2\rangle_{\mathrm{th,fl}}$ and $\langle i^2\rangle_{\mathrm{shot}}$ that correspond to the output voltage noise of the amplifier (thermal and flicker) and the shot noise of the sensor diode respectively. Thorough noise analysis can be found in the literature (e.g. [5, 35]). However, here an alternative, simplified derivation based on feedback theory will be given to gain insight on the parameters that impact noise performance. To simplify the analysis, we only consider the noise contribution of the input transistor M_{in} to $\langle v_{\mathrm{pa}}^2\rangle_{\mathrm{th,fl}}$ since it is the most dominant. Therefore, $\langle v_{\mathrm{pa}}^2\rangle_{\mathrm{th,fl}}$ arises from the product of input transistor M_{in} total noise current $\langle i_{M_{\mathrm{in}}}^2\rangle$ and the output resistance R_{o_1} of the high impedance node V_1. Furthermore, we make the assumption that the amplifier is not bandwidth limited as the effect of frequency filtering can be later included in the shaper stage.

The CSA output noise voltage due to the sensor shot noise current $\langle i^2\rangle_{\mathrm{shot}}$ can be easily calculated considering that almost all current will flow through C_f due to the feedback action. Hence, taking into account (3.21) we obtain:

$$\frac{\mathrm{d}\langle v_{\mathrm{CSA}}^2\rangle_{\mathrm{shot}}}{\mathrm{d}\omega} = \frac{eI_0}{\pi\omega^2 C_f^2} \tag{3.30}$$

To determine the contribution due to the input (M_{in}) transistor noise at the output of the amplifier $\langle v_{\mathrm{pa}}^2\rangle_{\mathrm{th,fl}}$, which is essentially a disturbance in terms of feedback theory, we must consider the transfer function of $\langle v_{\mathrm{pa}}^2\rangle_{\mathrm{th,fl}}$ to the CSA output noise $\langle v_{\mathrm{CSA}}^2\rangle_{\mathrm{th,fl}}$. We notice that a portion of the output voltage is fed back to the input with a feedback ratio β which is equal to $C_f/(C_f + C_D)$ and is determined by the capacitive divider consisting of the sensor capacitance C_D and the feedback capacitance C_f. Since usually $C_f << C_D$, the feedback ratio is approximately equal to $\beta = C_f/C_D$. Therefore, the noise transfer function is given by:

$$\begin{gathered} -\beta^2 g_{m_{\mathrm{in}}}^2 R_{o_1}^2 \langle v_{\mathrm{CSA}}^2\rangle_{\mathrm{th,fl}} = \langle v_{\mathrm{CSA}}^2\rangle_{\mathrm{th,fl}} - \langle v_{\mathrm{pa}}^2\rangle_{\mathrm{th,fl}} \Rightarrow \\ \frac{\langle v_{\mathrm{CSA}}^2\rangle_{\mathrm{th,fl}}}{\langle v_{\mathrm{pa}}^2\rangle_{\mathrm{th,fl}}} = \frac{1}{\beta^2}\frac{1}{g_{m_{\mathrm{in}}}^2 R_{o_1}^2} = \frac{C_D^2}{C_f^2}\frac{1}{g_{m_{\mathrm{in}}}^2 R_{o_1}^2} \end{gathered} \tag{3.31}$$

Substituting $\langle v_{\mathrm{pa}}^2\rangle_{\mathrm{th,fl}} = \langle v_{M_{\mathrm{in}}}^2\rangle = \langle i_{M_{\mathrm{in}}}^2\rangle R_{o_1}^2$, the CSA output voltage due to the input transistor noise is given by:

$$\langle v_{\mathrm{CSA}}^2\rangle_{\mathrm{th,fl}} = \frac{C_D^2}{C_f^2}\frac{1}{g_{m_{\mathrm{in}}}^2 R_{o_1}^2}\langle i_{M_{\mathrm{in}}}^2\rangle R_{o_1}^2 = \frac{C_D^2}{C_f^2}\frac{\langle i_{M_{\mathrm{in}}}^2\rangle}{g_{m_{\mathrm{in}}}^2} \tag{3.32}$$

To reduce $\langle v_{\mathrm{CSA}}^2\rangle_{\mathrm{th,fl}}$, a high loop gain ($A_0\beta$) is required to suppress the disturbance through the feedback action, which translates to a small detector capacitance C_D (high feedback ratio) and high transconductance (high power consumption). Taking into account (3.23),(3.24),(3.30) and (3.32), the total CSA output noise is equal to:

$$\frac{\mathrm{d}\langle v_{\mathrm{CSA}}^2\rangle}{\mathrm{d}\omega} = \frac{eI_0}{\pi\omega^2 C_f^2} + K_f \frac{1}{C_{\mathrm{ox}}^2 WL} \frac{C_D^2}{C_f^2}\frac{1}{\omega} + \frac{4}{3\pi}\frac{kT}{g_{m_{\mathrm{in}}}}\frac{C_D^2}{C_f^2} \tag{3.33}$$

The first term represents the sensor shot noise contribution, the second term the input transistor flicker noise contribution and the third term represents the input transistor thermal noise contribution. We observe that all terms decrease with the feedback capacitance C_f. However, since charge amplification is proportional to $1/C_f$, C_f disappears during the conversion to the input equivalent noise charge (ENC). For an exact derivation of the ENC, the noise power density spectrum must be integrated taking into account the shaper filtering effect (transfer function). Such a calculation can be found in [35]:

$$\mathrm{ENC}^2 \approx eI_0\tau + \frac{K_f}{C_{\mathrm{ox}}^2 WL} C_D^2 + \frac{4}{3}\frac{kT}{g_m}\frac{C_D^2}{\tau} \tag{3.34}$$

where τ is the shaper time constant (assuming a CR^1-RC^1 filter). Both the ENC and timing performance (rise time) depend on the detector capacitance C_D and the input transistor transconductance g_m that scales with the analog power consumption. It is therefore important to optimize the sensor capacitance in order to achieve better performance and lower power consumption as will be further explained in a more general context in the following section.

3.1.2.4 The importance of the Q/C ratio for high analog performance

The analog power consumption is often determined by the signal to noise ratio (SNR) for a given bandwidth. The signal in this context can be expressed as the voltage amplitude at the sensor defined by the Q_s/C_{in} ratio of the collected signal charge Q_s to the input capacitance C_{in} which is dominated by the sensor capacitance C_D. To obtain the SNR, the signal amplitude (Q_s/C_D) has to be compared to noise expressed as an equivalent series voltage source at the input, or in other words the input charge (Q_s) has to be compared to the ENC. As mentioned, the noise of the input transistor is usually dominant and especially its thermal component which is inversely proportional to the transconductance g_m. Under this assumption, the noise voltage at the input is equal to (see eq. 3.24):

$$N = \frac{\mathrm{ENC}}{C_D} = \sqrt{n\frac{4KT}{g_m}} \tag{3.35}$$

where n depends on the transistor operating region (inversion coefficient) ranges from 2/3 in saturation to approximately 1/2 in weak inversion. Therefore, the SNR is proportional to Q_s/C_D times the square root of g_m [16]:

$$\frac{S}{N} = \frac{Q_s/C_D}{\mathrm{ENC}/C_D} = \frac{Q_s}{\mathrm{ENC}} \approx \frac{Q_s}{\mathrm{ENC_{th}}} = \frac{Q_s}{\sqrt{n\frac{4KT}{g_m}}\cdot C_D} \propto \frac{Q_s}{C_D}\sqrt{g_m} \tag{3.36}$$

The transconductance g_m is proportional to the transistor bias current in weak inversion and propotional to the square root of the bias current in saturation. Assuming that this current dominates the power consumption of the front-end pre-amplifier, the SNR for a certain bandwidth is a function of the bias current I_b and as a consequence the power consumption P:

$$\frac{S}{N} \propto \frac{Q_s}{C_D}\sqrt{g_m} \propto \frac{Q_s}{C_D}\sqrt[m]{I} \propto \frac{Q_s}{C_D}\sqrt[m]{P} \tag{3.37}$$

with $2 \leq m \leq 4$ depending on the transistor operating region. For a given SNR at a specific bandwidth, defined by the specifications, the power consumption strongly scales with Q_s/C_D [16]:

$$P \propto \left(\frac{Q_s}{C_D}\right)^{-m} \tag{3.38}$$

It is evident that Q_s/C_D is the key parameter to reduce the power consumption for a given analog performance. As an example, if the Q_s/C_D is increased by two times, for the same SNR, the analog power consumption can be reduced by more than four times.

3.1.2.5 Readout architecture

The readout architecture refers to the transfer of data from the pixels to the ROIC output and depends on the target application. Data in HEP usually includes the position, time and the corresponding pulse amplitude (charge). For time uncritical and low rate applications the simplest readout scheme can be realized using only a few transistors. The simple 3T (three transistor) readout has a row select, a source-follower buffer and an input baseline reset. The analog voltage of the selected pixel is sampled in the periphery and the readout occurs pixel by pixel following a rolling shutter sequence. For more demanding applications a frame-based readout such as the 3T scheme cannot cope with the high hit rate and a dedicated in-pixel front-end readout chain with discrimination, as shown in Fig. 3.8 is implemented. The digital readout architecture only processes pixels with amplitudes above the threshold set by the discriminator. This readout approach where only a reduced number of pixels that have been hit are processed is referred to as zero-suppression.

One of the most important specifications of the readout architecture is whether the system follows a full (direct) or a triggered readout. A full readout continuously processes and transfers data from every pixel that is hit. Full readout systems do not require high logic density since there is no need to store data on-chip. However, it may not be possible to achieve a sufficiently high output bandwidth in order to handle high hit rates, which would limit the readout efficiency. For demanding applications such as the LHC, data reduction on the detector ROIC is necessary in order to decrease the amount of data the Data Acquisition System (DAQ) has to process. This is achieved by selecting only a subset of interesting events by means of a trigger mechanism. The trigger signal is generated by other detector sub-systems within a fixed latency. All pixel hits must be stored on the ROIC until a trigger decision determines which ones to read out. In this case, the output data bandwidth has to only match the trigger rate. However, the main limitation is the available memory that is determined by the process logic density.

Several readout concepts have been developed to meet the requirements of hit buffering during the trigger latency [5]. A simple timer in the pixel has been succeeded by a "conveyor belt" architecture, used in the initial development of the ATLAS Front End-A (FEA) [50] chip, that uniformly transports hits at each clock cycle to the End of Column (EoC) where a timestamp is assigned and is used for trigger matching. An improved "column-drain" architecture[3] is implemented in the first generation LHC$_{\text{p-p}}$ ROIC chips, the CMS pixel chip [51] and the ATLAS FE-I3 [52]. Hit data buffering is performed in the periphery while only one hit at a time is stored in the pixel. Hit timing association (in the case of the FE-I3 chip) is performed via a Bunch Crossing ID (BCID) timestamp that is distributed across the whole matrix. All pixels share a common data bus that only one pixel at a time is allowed to use. As soon as a pixel is hit, it declares its state and the bus is arbitrated so that one pixel at a time transmits data via the column-bus with a pre-defined priority. Alternatively, the column drain logic can be also implemented off-pixel with the discriminator outputs connected to the periphery with one-to-one correspondence [53]. However this approach is limited by routing density and is not suitable for large matrices. To reduce the number of required lines, pixel grouping has been utilized in the development of the so called "parallel pixel to buffer (PPTB)" architecture [54].

Taking advantage of the higher logic density, local hit storage within the pixel matrix has been implemented starting with the second generation triggered ROICs such as the ATLAS FE-I4 [28]. As a consequence, the required column-bus bandwidth is relaxed and hit rate capability is limited by the amount of memory instead of the transfer rate. In addition to higher hit rates, this approach has enabled the development of larger ROICs. In order to reduce the required memory for a specific buffer depth, pixels have been grouped into regions with shared storage while additionally taking advantage of hit correlation between them. The third ROIC generation continues to evolve and expand region architectures by utilizing a collection of digital cores that are stepped and repeated. Both large size and high rate can be accomplished by replicating a column-drain architecture many times in a matrix of pixel regions, each acting like a "small chip".

The majority of readout architectures are synchronous which means that hit timing information is assigned based on a global timestamp or synchronization clock distributed across the pixel matrix. The drawback of clock distribution across a large area is the increased digital power consumption due to the large total distribution line capacitance. An alternative asynchronous concept that aims to improve power efficiency has been recently pursued [55]. The hit pixel address bits are immediately transferred asynchronously to the EoC via a common bus. To minimize potential conflicts, pixels are arranged in alternating groups across the column and the generated pulse width is short. Due to the lack of ToT, charge information is indirectly obtained by the time-walk between pixels of a specific group. The challenge of this approach is the correct synchronization of the pulses arriving at the EoC. The varying propagation delay across the column has to be corrected while additionally accounting for mismatch and process variations.

3.1.2.6 Radiation effects on the ROIC - surface damage

In contrast to bulk damage effects caused by NIEL, ionizing radiation damages the surface of silicon sensors and CMOS electronic circuitry, especially the silicon dioxide layer and Si-SiO$_2$ interface. Since the damage is proportional to the received dose, it is described by a quantity called Total Ionizing Dose (TID) and is measured in gray or rad units. TID effects in CMOS circuits are induced by charge carriers trapped in oxide structures: the gate oxide, the shallow trench isolation (STI) and gate spacers.[4] Ionizing radiation generates charge carriers within these oxides, which leads to the

[3] The column-drain architecture will be covered in detail in Chapter 4.

[4] Gate spacers are used in modern CMOS technologies to reduce hot carrier effects by creating a lightly doped drain (LDD) region.

accumulation of positive static charge because hole mobility in SiO_2 is 10^6 times lower than electron mobility [20, 35]. Radiation induced positive charge in the gate oxide decreases the NMOS transistor threshold, leading to the formation of a leakage current, and increases the PMOS transistor threshold, hindering its action. However, modern CMOS technologies featuring thin gate oxides (few nm thick) prevent charge build-up because the quantum tunneling effect that allows the escape of the positive trapped charges becomes significant.

STI and spacers are, in contrast, thick oxide structures that accumulate positive charge and affect the radiation tolerance of CMOS circuits. Charge accumulation is additionally influenced by a second ionization induced effect: the formation of dangling bonds at the oxide interface that act as interface traps. Negative charge trapped by these interface states compensates some of the oxide positive static charge. However, the time scales of the two charge accumulation mechanisms is different. At low doses the net result is complex and depends on the technology parameters. At high doses a more stable state of an approximate but not perfect cancellation is reached and a positive net charge always remains. The net positive charge results in lateral gating in the case of the STI and in the modification of source and drain regions in the case of spacers. These effects are more pronounced for narrow and short channel MOS transistors and are called Radiation Induced Narrow Channel Effect (RINCE) and Radiation Induced Short Channel Effect (RISCE) respectively. RINCE affects the sides of the channel causing them to be conductive in the case of an NMOS transistor and non-conductive in the case of a PMOS transistor. Therefore, NMOS transistors develop a parasitic leakage current while PMOS transistors are turned off near the sides reducing the current flow. This effect can be visualized as a radiation dependent width change and has significant impact on narrow transistors, since a significant fraction of the total current flows near the sides. RISCE affects PMOS and NMOS devices in a similar way, impeding the charge flow between the source and drain through the channel. It can be roughly modeled as a radiation induced increase of the channel length and particularly affects short channel devices since the relative change in the effective length is significant. Additionally, apart from the influence on intentional transistor structures, positive charge accumulation on the STI interface can lead to the activation of parasitic NMOS transistors under the thick oxide that affects the circuit functionality. TID effects are influenced by transistor bias and temperature. In general, RINCE and RISCE occur when the transistors are powered and electric fields are generated in the STI and spacer oxides. A high temperature results in the increase of TID damage during irradiation, hence low operating temperatures are desired. After irradiation, high temperatures can be used to anneal part of the damage.

In order to enhance radiation tolerance to TID effects, wide and long transistors as well as specialized radiation hard geometries such as the Enclosed Layout Transistor (ELT) must be used. In the case of analog circuits, large transistors (compared to the minimum width and length) not only can be used but are usually required for high analog performance especially in modern sub-µm technologies due to short channel effects. On the other hand, in the case of high density synthesized logic, designers have less freedom in selecting more tolerant device geometries and increasingly have to rely on accurate modeling of radiation damaged transistors. At critical transistor circuit nodes, the usual linear transistor can be replaced by an enclosed layout geometry, shown in Fig. 3.14, so that the gate completely surrounds the central diffusion and in turn the perimetric diffusion surrounds the gate. Hence, edge effects are avoided and unwanted current paths are strongly suppressed. In order to interrupt parasitic current paths under the STI, p^+ guard rings are additionally used. The disadvantage of ELT transistors is that they occupy a lot of real estate (thus they are not suitalbe for high density logic), and they are not accurately described by the device models provided by the vendor (which are extracted for linear transistors). Furthermore, the W/L (width/length) ratio of ELT transistors cannot be lower than ≊ 2 and due to their large size, the capacitance at the source and drain is relatively high. The central diffusion exhibits less capacitance and is usually used as the drain (most sensitive node).

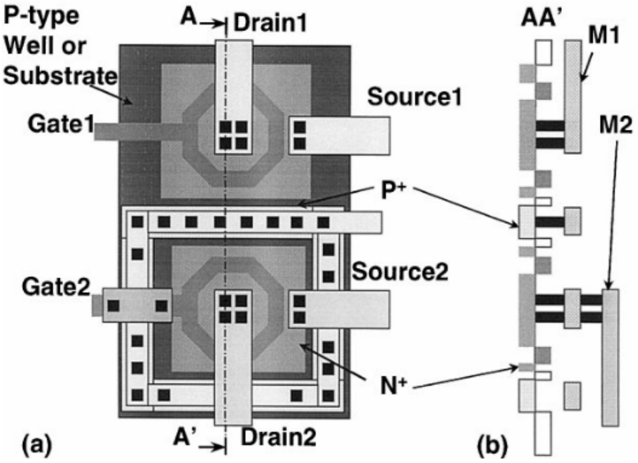


Figure 3.14: Radiation hard transistors laid out in an enclosed geometry. The implementation of a p^+ guard ring prevents leakage between the two NMOS devices [56].

3.2 The monolithic approach: DMAPS

An alternative approach to hybrid pixel technology is to combine the sensor and readout electronics in the same silicon crystal and form a monolithic pixel that can be produced using commercial CMOS technologies. Monolithic pixel detectors do not require complex and laborious bump-bonding and flip-chipping and can be produced in large volumes by commercial semiconductor foundries. Therefore, they are cost-effective and ideal for large area detectors. Futhermore, since they constitute a single entity, their thickness can be small (100 – 150 μm), offering low material budget in the order of 0.5% X/X_0 compared to 1.5% X/X_0 in the case of hybrid pixels. Another advantage is the possibility to realize small pixels (< 30 μm) the size of which is not limited by the capabilities of bump-bonding technology. Low power consumption, that is critical is some applications, can also be achieved due to the very small sensor capacitance and subsequently the high Q_s/C_D ratio of certain monolithic implementations (low fill-factor variants, see below).

Monolithic active pixel sensors (MAPS) have been so far only suitable for experiments with relaxed requirements in terms of rate and radiation tolerance due to their slow and inefficient charge collection, especially after NIEL damage due to irradiation. The next generation of monolithic pixels, called depleted monolithic active pixel sensors (DMAPS) is currently in development. DMAPS exploit high resistivity substrates, high voltage biasing or a combination of both to achieve full depletion and fast charge collection by drift, rendering them suitable for high-rate high-radiation environments such as the $\mathrm{LHC}_{\text{p-p}}$. Multiple nested well technologies employed by DMAPS allow complex CMOS circuitry and high functionality to be integrated in the pixel.

3.2.1 Monolithic active pixel sensors (MAPS)

While monolithic detectors are standard for the detection of visible light, employing CMOS technologies to produce monolithic pixels for particle detection has been first proposed and realized in the early 1990s [13, 57]. However, they often required exotic fabrication steps, such as double sided processing which is incompatible with volume manufacturing in standard semiconductor foundries. As a followup, Monolithic Active Pixel Sensors (MAPS) based on CMOS imaging processes that make use of an epitaxial silicon layer grown on low-cost, low-ohmic substrate wafers have been introduced [14]. The epitaxial layer, that also hosts the CMOS circuitry, is exploited by MAPS as the sensing volume. The epi-layer thickness is typically in the range of $1 - 20\,\mu\text{m}$.

Requirements for particle detectors are somewhat different from those of traditional CMOS imagers, where the signal is generated within a depth of a few microns. Traversing particles generate charge across the full thickness of the epi-layer which has to be collected as efficiently as possible. However, since the allowed biasing voltages in standard CMOS technologies as well as the epi-material resistivity in original MAPS designs are low, the depleted region only extends locally around the collection electrode. Therefore, except for the small depleted region, charge collection mainly occurs by diffusion and hence is incomplete and slow (in the order of 100 ns). Because the deposited charge in the epitaxial material for a typical thickness of $15\,\mu\text{m}$ is small $< 1\,500\,e^-$, the noise of the readout electronics (ENC) must be correspondingly low in order to achieve a reasonalbe SNR. The principle of a MAPS detector is illustrated in Fig. 3.15. Electrons are collected by an n^+ collection electrode (well). In order to prohibit other n-wells from forming competing charge collection nodes and in order to allow full CMOS functionality (NMOS and PMOS) in the pixel, a deep p-well layer offered by multiple well processes is required. MAPS pixel detectors have been successfully used in lower that $\text{LHC}_{\text{p-p}}$ rate and radiation enviroments such as the STAR detector at Relativistic Heavy Ion Collider (RHIC) [58]. The ALICE experiment has also chosen MAPS pixels based on the 180 nm TowerJazz imaging process for its Inner Tracking System (ITS) upgrade that currently in construction [59].

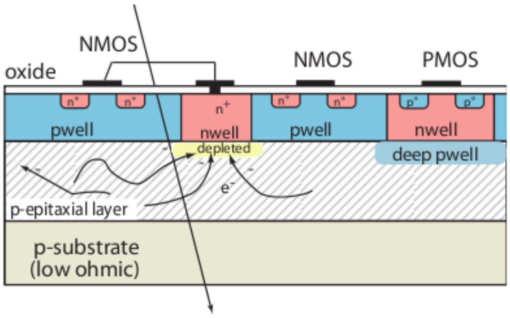


Figure 3.15: Conventional monolithic active pixel sensor (MAPS). Collection of the charge generated in the epitaxial layer is mainly governed by diffusion [35].

3.2.2 Depleted monolithic active pixel sensors (DMAPS)

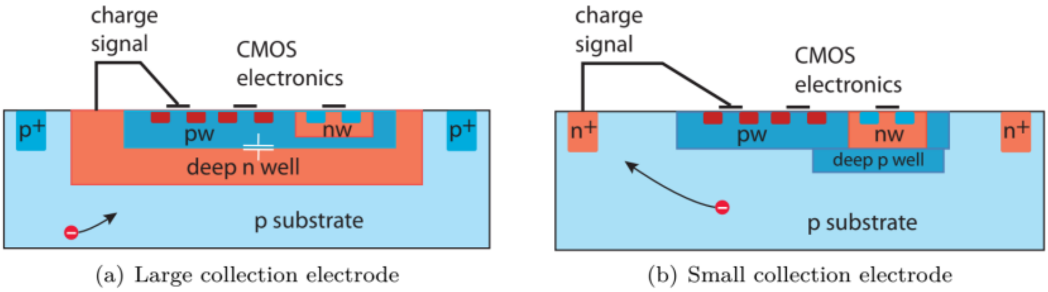
For LHC type high rate and high radiation environments, fast (≈ 10 ns) and complete collection of the deposited charge by drift is mandatory. Improved MAPS development lines that make use of commercial CMOS processes with some add-ons and modifications have been pursued to this end leading to the development of depleted (D)MAPS. R&D of DMAPS has gained significant momentum in the framework of the HL-LHC upgrade due to the potential for low cost and large area monolithic devices for outer tracker layer (ATLAS ITk L4). The technology and sensing properties of these new developments must therefore survive the radiation environment of the HL-LHC, at least in the outer pixel detector layer where the requirements are similar to those previously encountered in the inner layers of the LHC (refer to chapter 2).

The key to radiation tolerance in terms of charge collection and fast timing response is to achieve a depletion region underneath the collection electrode (typically 25 – 15 µm) that provides a sufficiently large and fast signal. In order to extend the depletion depth, high substrate resistivity and/or high reverse biasing voltages are required as described by eq. 3.10. Recently CMOS imaging processes have evolved independently of high energy physics, driven mainly by the smartphone market demands, utilizing high resistivity depleted silicon layers to achieve high collection efficiency for low light conditions. Furthermore, particular vendors interested in market corners away from mass IC production offer more freedom and allow process add-ons and modifications to be implemented. DMAPS detector development relies on these advancements by exploiting the following CMOS technology features:

- High voltage technology add-ons, usually inherited from power management and automotive applications. Reverse bias voltages higher than 200 V have been achieved by designs employing multiple guard ring structures [53].

- Medium to high resistivity (100 Ω cm) silicon substrate wafers, qualified by the foundry. High resistivity wafers can be produced using the Czochralski (Cz) method or by the float-zone (Fz) crystal growth process [35]. Although float-zone silicon has higher purity, Cz wafers have the advantage of lower cost and higher oxygen concentration which can be beneficial for radiation tolerance. Another option is to use wafers that feature a high resistivity epitaxial silicon layer grown on low-ohmic substrates (similar to the original MAPS designs). Due to bulk damage effects (e.g. induced acceptor removal), it has been observed [60] that independent of the starting substrate resistivity, within a large range (10 – 2 000 Ω cm), after a fluence of $10^{15}\ n_{eq}/\mathrm{cm}^2$ NIEL the effective space charge concentration becomes similar and equal to approximately $N_{\mathrm{eff}} \approx 10^{14}\ \mathrm{cm}^{-3}$ that corresponds to a resistivity of 100 Ω cm.

- Multiple nested wells (typically four including deep n,p wells) that are nessesary for transistor isolation and full CMOS functionality by shielding competing charge collection nodes.



(a) Large collection electrode (b) Small collection electrode

Figure 3.16: Two principal DMAPS impementation concepts: a) Large collection electrode: a deep n-well collection electrode encloses the CMOS electronics, b) Small collection electode: a small n-well placed outside the CMOS electronics area is employed as the charge collection node [35].

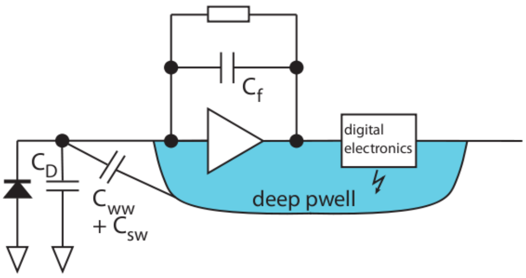


Figure 3.17: Coupling of digital switching noise via the inter-well and side-well capacitances in the case of a large collection electrode arrangement [20].

3.2.2.1 DMAPS implementation concepts

Two principal variants of a DMAPS pixel cell arrangement, shown in Fig. 3.16, can be implemented with different radiation tolerance and power consumption characteristics depending on the configuration and size of the collection electrode[5].

DMAPS with large collection electrode The large collection electrode implementation concept is illustrated in Fig. 3.16 (a). In this case, the collection electrode is realized by a large deep n-well that also hosts the in-pixel readout electronics. Due to the large fill-factor, the field configuration is intrinsically similar to standard planar n$^+$-in-p hybrid pixel sensors. Therefore, the large collection electrode concept provides good charge collection properties over the entire pixel area. Radiation tolerance benefits from the short average drift path to the collection electrode that reduces trapping probability.

The disadvantage of this approach is the high total capacitance at the preampifier input due to the large geometrical size of the collection electrode. In addition to the pixel-to-pixel and pixel-to-backside capacitances also present in any other sensor design, the inter-well capacitance (C_{ww}) arising from the close distance between the deep n-well and the deep p-well needed to shield the embedded electronics can be a significant contributor. The total capacitance amounts to several hundred fF and is comparable to typical hybrid pixel capacitances. As mentioned, a large input capacitance (C_D) results in a low Q_s/C_D ratio and therefore increased noise (ENC) and decreased timing performance (τ_r). To compensate for the high input capacitance, a large transconductance (g_m) value is required that translates to high power consumption. Even if a higher power consumption can be tolerated, the inter-well C_{ww} and side-well capacitances have another undesirable effect. As illustrated in Fig. 3.17, fast transient signals due to digital switching in the embedded electronics can capacitively couple into the sensitive input node and generate additional noise. Careful circuit design and special techniques are required in order to limit transient currents and prevent cross-coupling.

The high-energy physics community has targeted different prototyping designs with various foundries that provide the CMOS technology add-ons that are required to cope with the given demands. DMAPS in high voltage (HV), high resistivity (HR) technologies with large fill-factor are mainly represented by two development lines in the AMS 180 nm (later TSI 180 nm) HV CMOS process and the LFoundry 150 nm HV CMOS process. Fully monolithic demonstrator chips called ATLASpix

[5] The collection electrode size is sometimes refered to as fill-factor inheriting the terminology from CMOS imagers.

[61] and LF-Monopix [53] have been developed in the AMS and LFoundry technologies respectively and have been successfully characterized. The favorable charge collection properties of the large collection electrode implementation is evident from the high detection efficiency measured after irradiation. In the case of LF-Monopix1, a hit detection efficiency of 98.9% has been achieved after irradiation to $10^{15}\,n_{eq}/\mathrm{cm}^2$ NIEL.

DMAPS with small collection electrode An alternative approach is shown in Fig.3.16 (b). A small n-well is used as the charge collection node and is placed outside the CMOS electronics area. The prominent advantage of this arrangement is the very small detector capacitance ($C_D < 5\,\mathrm{fF}$) which results in a very high Q_s/C_D ratio and therefore high analog performance. For a given SNR, very low power consumption can be achieved, as explained in section 3.1.2.4. Furthermore, since the collection electrode is set apart from the in-pixel readout electronics, crosstalk due to digital transients is drastically reduced.

However, it is more difficult to achieve high radiation tolerance compared to the large electrode design due to the on-average longer drift paths (for the same pixel size) and the non-uniform electric field that is stronger near the collection electrode and weaker at the pixel corners. Therefore, a small pixel size is essential in order to achieve high charge collection efficiency in the case of small fill-factor designs. While a small pixel size is advantageous for physics performance, it comes with the expense of increased power density. However, due to the very small sensor capacitance and the exponential power consumption dependence on the Q_s/C_D ratio (eq. 3.38), the power efficiency is still higher compared to hybrid pixels or large fill-factor DMAPS while having the additional benefit of high granularity. A simple comparison can be made assuming a realistic set of design values based on the latest small collection electrode DMAPS (refer to chapter 4): high resistivity 25 µm thick epitaxial layer, 4 fF total input capacitance, 1 mW analog power consumption per pixel and $33 \times 33\,\mathrm{\mu m}^2$ pixel size. These values yield a Q_s/C_D ratio and analog power consumption density approximately equal to 64 mV and $80\,\mathrm{mW/cm}^2$ respectively. In the case of the LF-Monopix1, a large depletion depth can be achieved before irradiation due to the high biasing voltage (> 250 V) which results in the generation of a large amount of charge over $10\,\mathrm{k}e^-$. However, due to the high input capacitance ($C_D \approxeq 400\,\mathrm{fF}$ [53]), the Q_s/C_D ratio is more than a factor of 10 lower ($Q_s/C_D = 4\,\mathrm{mV}$). The analog power density is equal to $P = 290\,\mathrm{mW/cm}^2$ for the pixel size of $50 \times 250\,\mathrm{\mu m}^2$. The corresponding values for the ATLAS pixel detector are $Q_s/C_D \approxeq 10\,\mathrm{mV}$ and $P = 120\,\mathrm{mW/cm}^2$ due to the higher sensor thickness and larger pixel size ($50 \times 400\,\mathrm{\mu m}^2$) [16].

Achieving sufficient radiation tolerance with small fill-factor designs has so far been challenging. Recent developments, based on a novel modification of the TowerJazz 180 nm process, that aim to combine low noise and fast timing due to the small C_D with radiation hardness suitable for LHC type environments are in the scope of this thesis and are presented in detail in the following chapters.

CHAPTER 4

Development of the TJ-Monopix small collection electrode DMAPS in 180 nm CMOS

The TJ-Monopix chip series is a large scale, small fill-factor DMAPS with integrated standalone fast readout architecture and ToT capability fabricated in a modified ToweJazz (TJ) 180 nm CMOS imaging process. It is part of a DMAPS development line in TJ 180 nm, shown in Fig. 4.1, pursued by a collaboration between the University of Bonn and the CERN detector group with the goal to combine high analog performance (low noise, low power consumption) due to the very small sensor capacitance with fast charge collection by drift, hence high radiation tolerance.

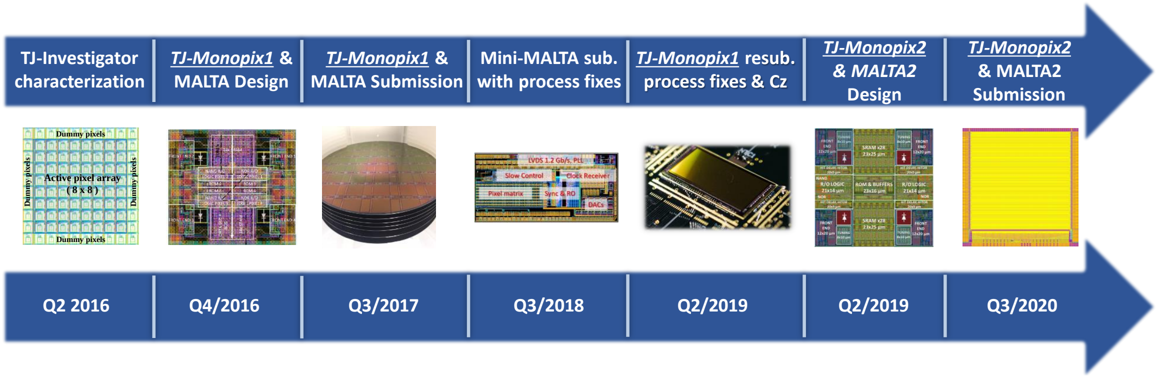

TJ-Investigator characterization
TJ-Monopix1 & MALTA Design
TJ-Monopix1 & MALTA Submission
Mini-MALTA sub. with process fixes
TJ-Monopix1 resub. process fixes & Cz
TJ-Monopix2 & MALTA2 Design
TJ-Monopix2 & MALTA2 Submission
Dummy pixels
Active pixel array (8 x 8)
Dummy pixels
Dummy pixels
Dummy pixels
LVDS 1.2 Gb/s PLL
Slow Control
Clock Receiver
Pixel matrix
Sync & RO
DACs
Q2 2016
Q4/2016
Q3/2017
Q3/2018
Q2/2019
Q2/2019
Q3/2020

Figure 4.1: Timeline of small fill-factor DMAPS development in the TowerJazz 180 nm process

The standard TJ 180 nm CMOS process has been used for the development of the ALPIDE monolithic active pixel sensor [59, 62], selected for the ALICE Inner Tracking System (ITS) upgrade. The new tracker will be fully constructed with MAPS, reducing the material budget to only 0.3% X/X_0. The ALPIDE is the first MAPS with pixel front end (amplifier and discriminator) and sparsified zero-suppression readout within the pixel matrix, similar to hybrid pixels. Although sufficient for the modest ALICE requirements, the standard process does not allow full depletion of the active volume resulting in signal degradation after fluences in excess of 10^{12} – 10^{13} n_{eq}/cm^2. Furthermore, it is not capable of fast (25 ns) timing required by the LHC$_{\text{p-p}}$ due to slow signal collection and long shaping time (several µs).

To improve NIEL tolerance to 10^{15} n_{eq}/cm^2 or beyond, a process modification has been developed by CERN in collaboration with the foundry [29] that allows full depletion of the sensitive layer. A dedicated chip called TJ-Investigator [63, 64] has been designed as a test vehicle to characterize the sensor performance. First results that demonstrate the effectiveness of the modified process, encouraged the development of large scale DMAPS with fast integrated readout for high rate and high radiation applications such as the outer layers of the HL-LHC. Two large scale demonstrator chips have been developed called TJ-Monopix1 ($1 \times 2\,\mathrm{cm}^2$) [30–32] and TJ-Malta1 ($2 \times 2\,\mathrm{cm}^2$) [55, 65]. While the two chips share a common sensor implementation and front-end amplifier, with some variations, their main difference is in the readout architecture. TJ-Monopix incorporates a more conventional "column-drain" readout architecture that has been proven by the ATLAS FE-I3 front end chip [52], while an asynchronous readout aiming to reduce digital power consumption by eliminating the Bunch Crossing ID (BCID) time stamp distribution across the full pixel matrix is used by TJ-Malta1.

TJ-Monopix1 has been extensively measured and characterized. Full functionality and high analog performance has been demonstrated by laboratory tests, radioactive source measurements and test-beam campaigns. Several TJ-Monopix1 and TJ-Malta1 chips have been irradiated up to $10^{15}\ n_{eq}/\mathrm{cm}^2$ NIEL in order to access the radiation tolerance of the modified process in fully monolithic large matrices with integrated readout. Although full functionality was retained, the hit detection efficiency after irradiation decreased from $\approxeq 97\%$ to $\approxeq 70\%$. The main reason for the efficiency drop has been discovered to be related to the weak lateral electric field at the pixel edges. To solve this issue, the process has been further optimized for faster and more efficient charge collection [66]. These modifications have been successfully evaluated with the help of a small scale chip, called mini-Malta [67], showing significant improvements in charge collection and were afterwards implemented in the original TJ-Monopix1 and TJ-Malta1 designs.

The pixel size is critical in order to take full advantage of field shaping through process modifications, and should be minimized for fast and complete charge collection. Furthermore, to achieve full efficiency and high performance the pixel design and front-end electronics have to be further optimized in order to obtain a lower operating threshold. TJ-Monopix2 is a full scale ($2 \times 2\,\mathrm{cm}^2$) next generation successor chip that has been completely re-designed and features:

- Enhanced charge collection in both epitaxial and Czochralski substrate materials.
- Smaller pixel size ($33.04 \times 33.04\,\mu\mathrm{m}^2$).
- Improved analog front-end that results in an expected reduction of the operating threshold by a factor of $\gtrsim 3$.
- More sophisticated digital periphery and fast data transmission.

TJ-Monopix2 chips have been recently fabricated (Q1 2021) and are currently in the initial testing phase.

4.1 Small collection electrode DMAPS pixel concept in the TowerJazz 180 nm CMOS process

4.1.1 Monolithic pixels in the standard TowerJazz 180 nm CMOS process

The TowerJazz 180 nm CMOS imaging process is a quadruple well technology originally oriented to CMOS camera applications. The foundry offers the possibility to use different starting materials, including high resistivity options, which makes it particularly interesting for high energy physics applications. The standard process employed by the ALPIDE MAPS makes use of a 18 – 30 µm thick, high resistivity (> 1 kΩ cm), p-type epitaxial layer grown on a low-ohmic p^+ substrate wafer. Futhermore, following the trend observed in many deep submicron CMOS technologies, tolerance to TID radiation effects in increased due to the thin gate oxide (3 nm).

The cross-section of a monolithic pixel in the TJ 180 nm standard process is illustrated in Fig. 4.2. Charge generated in the sensitive p-epitaxial layer is collected by a small n-well electrode (which yields a small input capacitance), separated from the readout electronics following the small fill-factor design concept. Any other n-well inside the pixel, such as the wells containing PMOS transistors, is shielded from the epitaxial layer by an deep p-well preventing competing collection nodes. Therefore, full CMOS functionality that allows for more complex readout circuitry, is possible. Furthermore, the deep p-well helps to reduce interference from electronic switching. Although the process allows the implementation of both deep p-well and deep n-well, they cannot be nested since they are formed at the same depth. Outside the pixel matrix, it is possible to use a deep n-well to obtain a standard triple well structure that shields the periphery circuitry from the reverse bias voltage applied to the epitaxial layer and also provides increased immunity to substrate noise coupling.

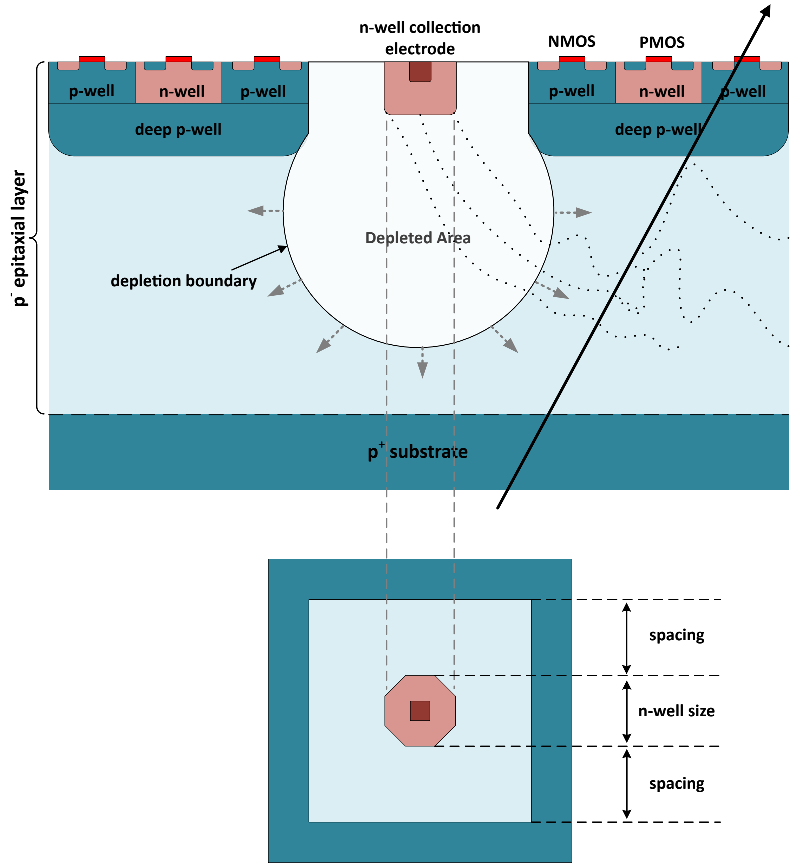


Figure 4.2: Pixel cross section of the MAPS design concept in the TowerJazz 180 nm CMOS process. A deep p-well shields the circuitry from the sensor and allows full CMOS in the pixel. In the standard process the epitaxial layer is not depleted over its full width.

Depletion and influence of reverse bias Due to the high resistivity of the p-epitaxial layer, a depletion layer can be formed around the collection electrode by applying relatively small reverse bias voltages. In order to bias the sensor three potentials can be adjusted: the collection electrode n-well, the in-pixel p-well (and deep-pwell) and the p-substrate which is accessed by the periphery surrounding the pixel matrix or through a backside contact. However, in the case of the standard process (ALPIDE) sensor the epitaxial layer is not fully depleted. Therefore, the p-well and p-substrate potentials should be equal to avoid excessive current draw (and consequent ohmic voltage drop) through the epitaxial layer.

The maximum reverse bias voltage that can be applied to the p-well is $\cong 6V$ since for higher (absolute) values the source and drain junctions of the NMOS transistors begin to breakdown. By applying reverse bias, the depletion region grows as indicated by the arrows in Fig. 4.2 but it is limited to the region around the collection electrode. Hence, the generated electrons are transported by diffusion before being collected by the strong drift field in the depleted zone around the small collection electrode. Reverse bias is also necessary to reduce the sensor capacitance. With the application of only a few volts, the area around the collection electrode is depleted leading to a very low sensor capacitance $C_D < 3\,\mathrm{fF}$ and a corresponding high Q_s/C_D ratio and low power consumption.

Influence of pixel geometries The sensor performance depends on the spatial extension of the depleted region (because it influences both charge collection and the sensor capacitance) and is a characterized by different parameters such as the charge conversion gain (capacitance), charge collection time, charge spread and total collection efficiency. Apart from the reverse bias voltage, the following key geometric parameters influence charge collection and signal size:

- *Pixel pitch*: A small pixel pitch results in on-average shorter path lengths traveled by generated charges, especially near the pixel corners where the epitaxial layer is not fully depleted, and is crucial to achieve fast and efficient charge collection in the case of a small fill-factor design. Furthermore, it increases the ratio of the depleted region volume to the total sensitive volume of the pixel. The minimum pixel size is determined by the sensor geometry (collection n-well size and spacing) and the area required by the in-pixel electronics.

- *Epitaxial layer thickness*: A thicker epitaxial layer has the advantage of a larger induced signal since for energetic charged particles it scales with the particle path length inside the sensitive layer. It also leads to increased charge sharing, hence for the standard process where charge sharing can be significant, a careful tuning of the sensor geometry is required.

- *Collection n-well size*: A small collection electrode size helps to minimize the sensor capacitance and consequently the signal to noise ratio. On the other hand, if the collection electrode is very small, the depleted region and the electric field strength are reduced leading to lower charge collection efficiency and higher charge sharing.

- *Spacing between the collection n-well and the surrounding p-well*: A large spacing helps to extend the depletion laterally, hence results in a more uniform electric field and reduces charge sharing between neighboring pixels. However, it reduces the available area for the readout circuitry. For small spacing values, the volume around the collection electrode is fully depleted and the side-well capacitance is in first order approximation inversely proportional to the distance between the collection electrode n-well and the surrounding p-well "parallel plates". However, as the spacing increases, the collection electrode depletion boundary begins to expand increasing the capacitance. Therefore, there exists an optimum spacing value for a given geometry and reverse bias voltage that minimizes the sensor capacitance.

The influence of pixel geometries and reverse bias voltages on charge collection properties have been studied with the help of the TJ-Investigator test chip. The TJ-Investigator chip contains a large number of mini pixel matrices with varying pixel pitch, collection electrode size and spacing geometry and has been produced using three epitaxial layer thicknesses: 18, 25 and 30 µm. Each mini-matrix, containing an 8 × 8 pixel array, can be selected by a chain of switches within the signal path. All 64 pixel outputs of a mini-matrix are individually connected to a source-follower buffer circuit driving the analog signal observed at the chip output, allowing direct real-time observation of all 64 pixel signals in parallel.

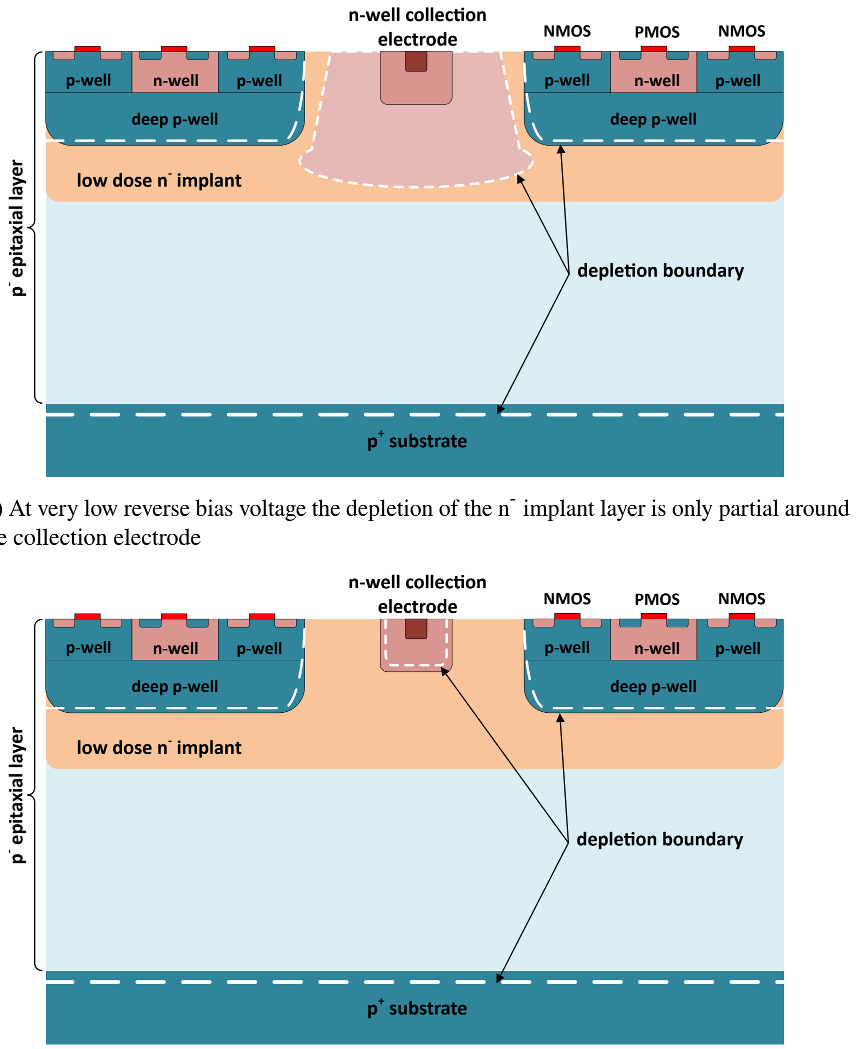
To identify the optimum sensor geometry, the signal output of $28 \times 28\,\mu m^2$ pixels with varying collection n-well size from 1.2 µm (minimum) to 5 µm and spacing from 1 – 5 µm were measured and analyzed. The signal is produced by the absorption of 5.9 keV (K_α) and 6.5 keV (K_β) photons emitted during the electron capture decay of a ^{55}Fe source illuminating the test chip. The extreme case of the smallest collection n-well size and spacing has been measured to yield low charge collection efficiency, as expected. Increasing the size and spacing parameters leads to different trade-offs between signal amplitude and charge sharing. According to the measurement results, the optimum combination is a collection electrode size of 2 µm and spacing to the surrounding p-well equal to 3 µm [68].

4.1.2 Process modification towards full depletion of the sensitive layer

Since depletion in the standard process starts at the junction of the small collection electrode at the pixel center, it is difficult to laterally extend the depletion far into the epitaxial layer in between the low resistivity substrate and the deep p-well, even if higher reverse bias voltages are applied, as this requires a potential gradient (or electric field) configuration similar to a planar junction. In order to achieve full depletion without compromising the sensor capacitance, a process modification, depicted in Fig. 4.3, has been developed which implements a large, even, planar-like junction within the epitaxial layer [29].

A low dose, deep, n-type layer is implanted in the epitaxial layer within the pixel matrix below the wells containing circuitry, while the collection n-well and deep n-type implant profiles overlap sufficiently to form a continuous n-type layer. As a result, depletion starts from two p-n junctions (deep p-well/n$^-$ layer, p-epi/n$^-$ layer) and immediately extends over the full pixel area, even at 0 V reverse bias. A potential minimum for electron collection is formed in the sensor volume with a field component towards the n$^+$ collection electrode, strengthening collection of charges laterally [35]. An important advantage of this approach is that the process modification does not require any layout changes in the design of the sensor, therefore the same designs can be fabricated in both the standard and the modified process. At low biasing voltages (Fig. 4.3(a)) the region around the collection electrode is not fully depleted, resulting in a higher sensor capacitance. In order to fully deplete the area around the collection electrode and maintain a low sensor capacitance ($C_D \approxeq 3$ fF), a moderate reverse bias of the collection electrode (up to 6 V) is required (Fig. 4.3(b)).

In contrast to the standard process, since depletion extends over the full pixel, the p-well/deep-pwell and p-substrate are isolated and can be biased independently provided that a sufficiently large potential



(a) At very low reverse bias voltage the depletion of the n$^-$ implant layer is only partial around the collection electrode

(b) For higher reverse bias voltages, the depletion reaches the n-well implant of the collection electrode yielding a low sensor capacitance

Figure 4.3: Cross section of a DMAPS pixel in the modified process. A low-dose n-type implant is used to implement a planar junction and deplete the epitaxial layer over the full pixel area.

barrier prevents the holes in the p-well from entering the epitaxial layer and hence avoid punchthrough[1]. At approximately 20 to 30 V (depending on the epitaxial layer thickess) reverse p-substrate bias V_{PSUB}), punchthrough sets in and the p-well and p-substrate current starts to exponentially increase. Having the ability to separately bias the p-substrate is beneficial to enhance the vertical electric field component and further extend the depletion region if thicker substrate materials are used (see section 4.1.4). However, a high V_{PSUB} can also "flatten" the potential landscape under the deep-pwell and weaken the lateral field component.

A critical parameter is the dose of the n-implant. In order to form the n^{-} layer, the dose has to be high enough to dominate the p-epitaxial layer doping concentration which is in the order of 10^{13} cm^{-3}. The n^{-} layer doping concentration should be sufficiently low to be fully depleted at reasonable voltages, but high enough to prevent punchthrough between the p-well and the substrate. A low dose results in lower capacitance values that increase the signal to noise ratio and reduce the operating threshold. However, a sufficiently high dose is critical in order to prevent type inversion of the n-implant layer after irradiation to high NIEL fluences ($> 10^{15}$ n_{eq}/cm^2).

4.1.2.1 Modified process sensor first tests

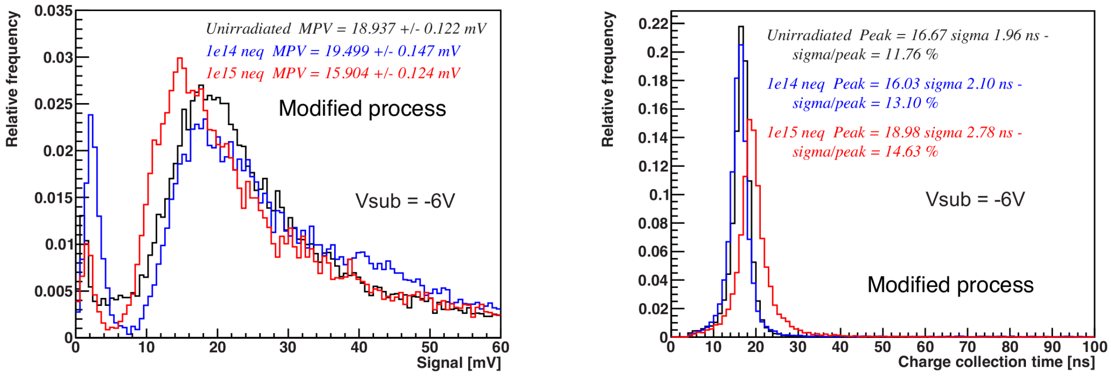
Before the development of the large scale demonstrator chips, the charge collection properties and detection efficiency of the modified process sensor was studied to verify whether it carries the potential to be used in high radiation environments such as the ATLAS ITk outer layers [64]. To this end, the TJ-Investigator test chip originally designed for the ALICE ITS upgrade, has been fabricated in the modified process and several samples have been irradiated up to a fluence of 10^{15} n_{eq}/cm^2 in the Triga reactor facility, Slovenia [69]. The epitaxial layer thickess of the measured samples is equal to 25 µm and reverse bias voltage of −6 V was applied during the measurement.

Signal response characterization using radioactive sources In order to study the signal spectrum and response time, ^{55}Fe and ^{90}Sr radioactive sources were used. The produced photo-electon by 5.9 keV (K$_{\alpha}$) photons emitted by the ^{55}Fe source deposits an ionization charge of approx. 1 616 e^{-} in the sensor and is used to calibrate the signal response (conversion gain) of different pixel designs. The ^{90}Sr source emits electrons that traverse the sensor and generate a signal similar to minimum ionizing particle (MIPS).

The measurement results of ^{90}Sr β-source tests for an unirradiated sample, a sample irradiated to 10^{14} n_{eq}/cm^2 and a sample irradiated to 10^{15} n_{eq}/cm^2 are compared in Fig. 4.4. The pixel size is equal to $50 \times 50\,\mu\text{m}^2$ with 3 um collection electrode size and 18.5 µm spacing. The signal amplitude spectrum is shown in Fig. 4.4 (a) and as expected follows a Langau-Gauss distribution. The red curve shows the excellent signal response that is maintained after a fluence of 10^{15} n_{eq}/cm^2. In comparison, no useful signal could be acquired from a sensor produced in the standard process after the same irradiation fluence. The slight reduction of signal amplitude is attributed to a gain reduction in the signal processing chain used to readout the sensor output. After calibration with an ^{55}Fe source, the charge distribution of the unirradiated and irradiated sensors yield comparable most probable charge values (MPV) of ≊ 1 740 e^{-}, while the charge spectrum is well separated from the noise peak around the baseline. The signal rise time is the convolution of the charge collection time and the amplifier

[1] The junction punchthrough is a phenomenon that occurs when the depletion layer boundary moves from one junction and touches another junction as it expands.

response and its distribution is shown in Fig. 4.4 (a). Despite a slight increase of the charge collection time from 16.7 ns to 19 ns and its spread from 1.96 ns to 2.78 ns, the sensor maintains a fast signal response after irradiation and charge collection is still faster compared to the standard process, even after irradiation.



(a) (b)

Figure 4.4: Signal response to an ^{90}Sr source for a 50 × 50 µm^2 pixel manufactured in the modified process before irradiation (black curves), after 10^{14} n_{eq}/cm^2 NIEL (blue curve) and after 10^{15} n_{eq}/cm^2 NIEL (red curve). The signal amplitude distribution is shown in (a) and signal rise time distribution is shown in (b) [64].

Test beam measurement results In order to measure the sensor efficiency, the modified process CMOS sensors were installed in a test beam setup at the CERN Super Proton Synchnotron (SPS) which delivers 180 GeV/c pions. The tracks have been reconstructed with an FE-I4 based telescope with three reference planes upstream and three downstream with 9 µm position resolution. The hit detection efficiency is defined as the ratio of the number of events with telescope tracks and a corresponding sensor hit to the total number of events with a telescope track and is calculated as a function of extrapolated hit position in X and Y coordinates. For each measured mini-matrix variation, signals from a 2 × 2 pixel subset were recorded and corrections due to acceptance edge effects were applied in the efficiency calculation.

The measured hit detection efficiency for a 30 × 30 µm^2 pixel with 3 um collection electrode size and 3 µm spacing after irradiation to 10^{15} n_{eq}/cm^2 is shown in Fig. 4.5. To minimize errors, the efficiency measurement area has been restricted to the area between the four pixel centers. For this sensor, the measured efficiency is uniform across the pixel cell with an average value of 97.4% ±1.5% (stat) ±0.6% (syst) , while for a 25 × 25 µm^2 pixel with the same collection electrode and spacing geometry it slightly increases (98.5% ±1.5% (stat) ±1.2% (syst)) due to the reduced pixel size. These results demonstrate the substantial improvement of radiation hardness through the process modification.

It is worth noting that because in the case of the modified process charge is collected predominantly by drift, charge sharing is significantly reduced compared to the standard process. The collected charge and cluster size as a function of hit position in the pixel were calculated using the extrapolated hit position and individual pixel waveform data acquired by the test beam setup. While single pixel clusters dominate the 50 × 50 µm^2 pixel sensors, the mean cluster size is increased to ≊ 1.35 for the 25 × 25 µm^2 and 30 × 30 µm^2 pixel sensors due to the smaller collection n-well to p-well spacing (3 µm) of these variants.

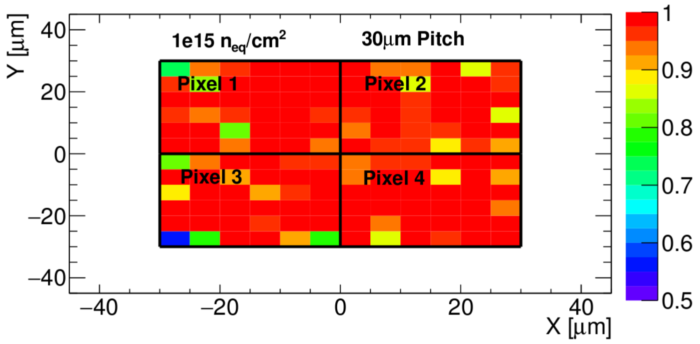


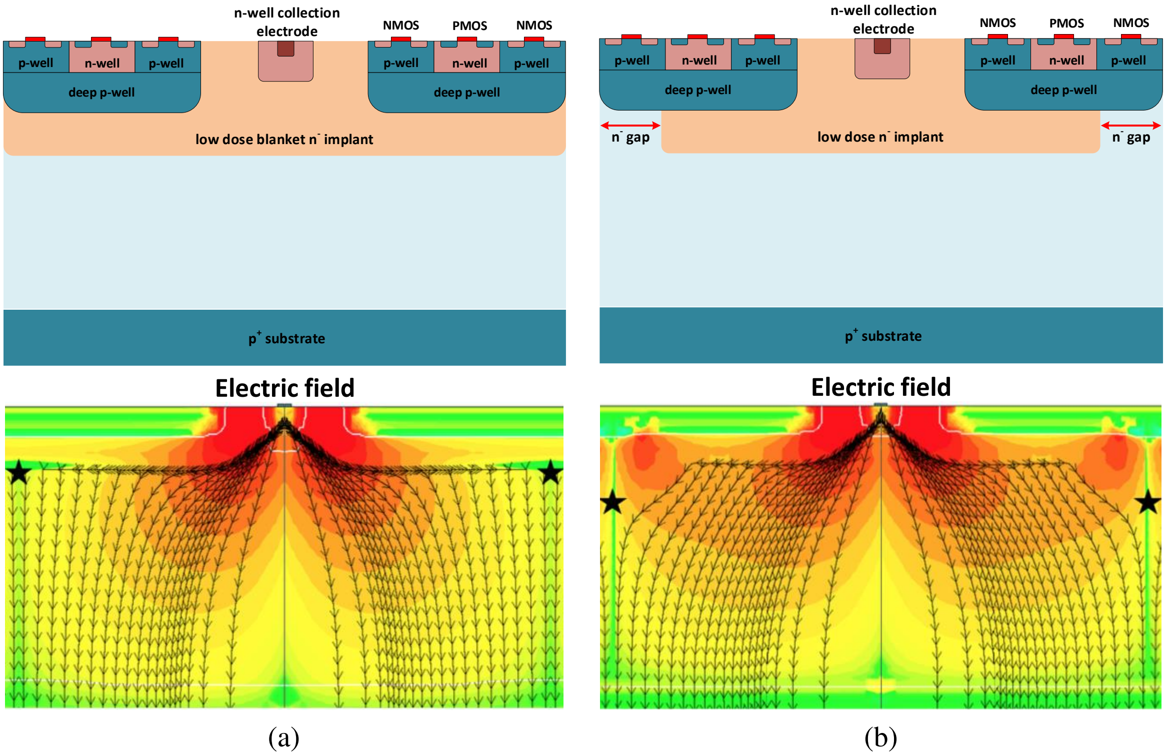
Figure 4.5: Hit detection efficiency across a 4-pixel area as a function of hit position for a $30 \times 30\,\mu\text{m}^2$ pixel manufactured in the modified process after irradiation to $10^{15}\ n_{eq}/\text{cm}^2$ NIEL [64].

4.1.3 Optimization for lateral field enhancement

Although the process modification by implanting a deep planar n⁻ layer allows to achieve full depletion of the sensitive layer and significantly improves radiation tolerance, reduced efficiency measured in the pixel corners of TJ-Monopix1 and TJ-Malta1 [33, 34, 65, 70] revealed the significance of the non-uniform electric field which drops to a minimum at the pixel corners. Since this effect is more pronounced for larger pixels, fast (few ns) and efficient charge collection is challenging for pixel sizes of approximately $40 \times 40\,\mu\text{m}^2$ or higher.

Technology Computer Aided Design (TCAD) simulations have been carried out to study the sensor performance and develop improvements of the process modification [66, 71]. The lateral electric field of a $36.4 \times 36.4\,\mu\text{m}^2$ pixel in the modified process, reverse biased at 6 V is shown in Fig. 4.6 (a) [66]. As observed, the lateral field component becomes minimum (zero) at the pixel corners under the deep p-well. Combined with the vertical component of the electric field along the sensor depth which reaches a zero value close to the top junction boundary they form a "ridge" in the potential landscape under the p-well/deep p-well across the pixel and a saddle point at the pixel corners, indicated by a star symbol. Even though charge collection happens predominantly by drift, part of the generated electrons are pushed by the sensor vertical field into this potential minimum where the drift field is almost zero and hence the trapping probability is high. The lateral field strength around the potential minimum depends strongly on the pixel size and becomes significant for the charge collection properties of pixels larger that $\approx 20 \times 20\,\mu\text{m}^2$. If a high enough reverse bias voltage is applied to the p-substrate the vertical field component is increased, however the lateral field component decreases resulting in a flattening of the potential landscape and longer drift paths.

Increasing the lateral field component near the pixel corners is key to achieve fast charge collection and high radiation tolerance. This can be achieved by introducing junctions along the sensor depth that create a potential gradient and electric field lines towards the collection electrode at the center of the pixel. One way to form vertical juction segments is by altering the deep p-well layout in order to remove a portion close to the collection electrode, thus creating a "step" in the p$^+$-implant depth close to the pixel corners. However this approach is limited by the in-pixel electronics layout since every n-well (apart from the collection electrode) has to be shielded by deep p-well. A similar effect can be achieved more controllably and independent of the electronics layout, by minimal adjustments to the manufacturing process, illustrated in Fig. 4.7. Either a gap is created in the deep n⁻ implant near the



(a)

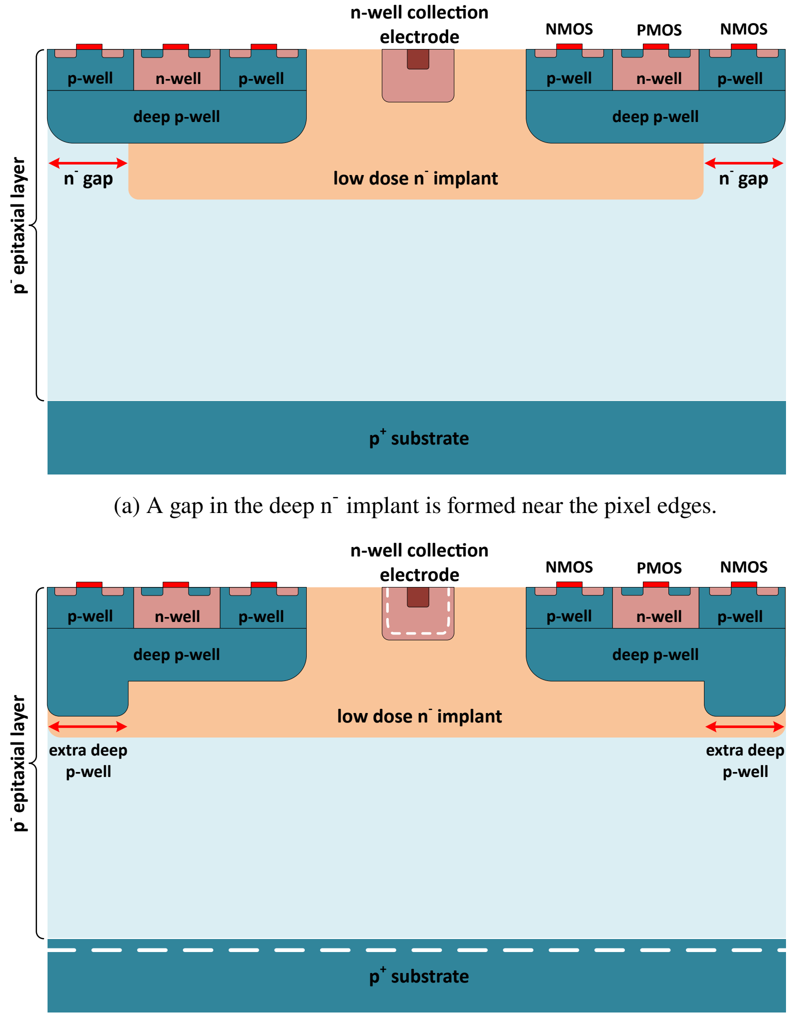
(b)

Figure 4.6: Comparison of the lateral electric field for a 36.4 × 36.4 μm^2 pixel implementation in a) the modified process and b) the modified process with a gap in the deep n$^-$ implant. The collection electrode is biased at 0.8 V and the p-well and p-substrate at −6 V. The electrostatic simulation results are reproduced from [66].

pixel edges (Fig. 4.7(a)) or an extra deep p-well is implanted instead (Fig. 4.7(b)). In the case of the gap in the n$^-$ layer only a mask modification is required, while in the case of the extra deep p-well an additional mask that is already available by the foundry is required. An electrostatic simulation of the lateral electric field in the case of a pixel with a gap in the n$^-$ layer is shown in Fig. 4.6 (b). Apart from a significant increase in the lateral field component at the pixel corners, the potential minimum is shifted deeper in the sensor volume bending the field lines toward the collection electrode and reducing the drift path length.

A side-effect of the process modification enhancements is the weaker separation between the p-well and p-substrate since the potential barrier is reduced. Therefore, punchthrough begins at lower reverse p-substrate bias voltages (V_{SUB}). For a p-well voltage equal to $V_{\mathrm{PW}} = -6$ V, punchthrough occurs approximately at $V_{\mathrm{SUB}} = -10$ V in the case of the extra deep p-well and even lower, at $V_{\mathrm{SUB}} = -8$ V, in the case of the gap in the n$^-$ layer since the small depletion of the epitaxial layer does not provide enough isolation. However, at least for the 25 μm thickness of the epitaxial layer, higher V_{SUB} voltages are not required as they would not provide any further improvement of the charge collection properties for pixels manufactured with the lateral field enhancement modifications.

The effectiveness of the process enhancements in terms of charge collection time and efficiency has been studied by transient 3D TCAD simulations for a minimum ionizing particle traversing the pixel corner for both non-irradiated sensors and sensors with modeled radiation damage equivalent to $10^{15}\ n_{eq}/\mathrm{cm}^2$ [66]. The simulated charge collection time for a pixel size of 36.4 × 36.4 μm^2 is equal to < 5 ns, which is an improvement by at least a factor of two. In comparison, in the case of the original planar n$^-$ layer process modification not all charge is collected within 25 ns, even before irradiation, and a large fraction of the charge is being lost due to trapping after irradiation. After the implementation of the process modification enhancements, the total collected charge after irradiation is increased by approximately a factor of three.



(a) A gap in the deep n$^-$ implant is formed near the pixel edges.

(b) An extra deep p-well is implanted near the pixel edges.

Figure 4.7: Process modification enhancements to increase the lateral electric field at the pixel borders.

4.1.4 High resistivity thick Czochralski substrate alternative

Although processing on a p-epitaxial layer is standard for the TowerJazz 180 nm process, the foundry has accepted thick, high-resistivity ($> 800\,\Omega\,\mathrm{m}$) p-type Czochralski (Cz) wafers that have been used for the fabrication of TJ-Monopix1 and TJ-Malta1 chips [72]. The manufacturing of small fill-factor DMAPS with special implant geometries on a high-resistivity Cz substrate allows to combine the advantages of a small collection electrode with the advantages of thick sensors. While the low sensor capacitance and low noise is maintained, the signal amplitude is significantly increased due to the higher ionization charge, given that a sufficiently large depletion depth is achieved. The higher signal to noise ratio can improve radiation tolerance as well as timing performance. However, charge sharing is also expected to increase due to charge cloud spread by diffusion.

A high substrate bias voltage V_{PSUB} is essential to achieve a depletion depth higher than 25 µm (which is the thickness of the epitaxial layer sensors). In the case of the original planar n$^-$ process modification substrate voltages down to −20 V can be applied. After irradiation to 10^{15} n_{eq}/cm^2, even higher bias voltages, down to −50 V, have been achieved due to the effective doping change by NIEL damage. Similarly, while for the process modification enhancements, only moderate reverse substrate voltages ($\cong$ 8 – 10 V) can by applied before irradiation, V_{PSUB} bias down to −20 V has been achieved after irradiation.

4.2 The column-drain readout architecture

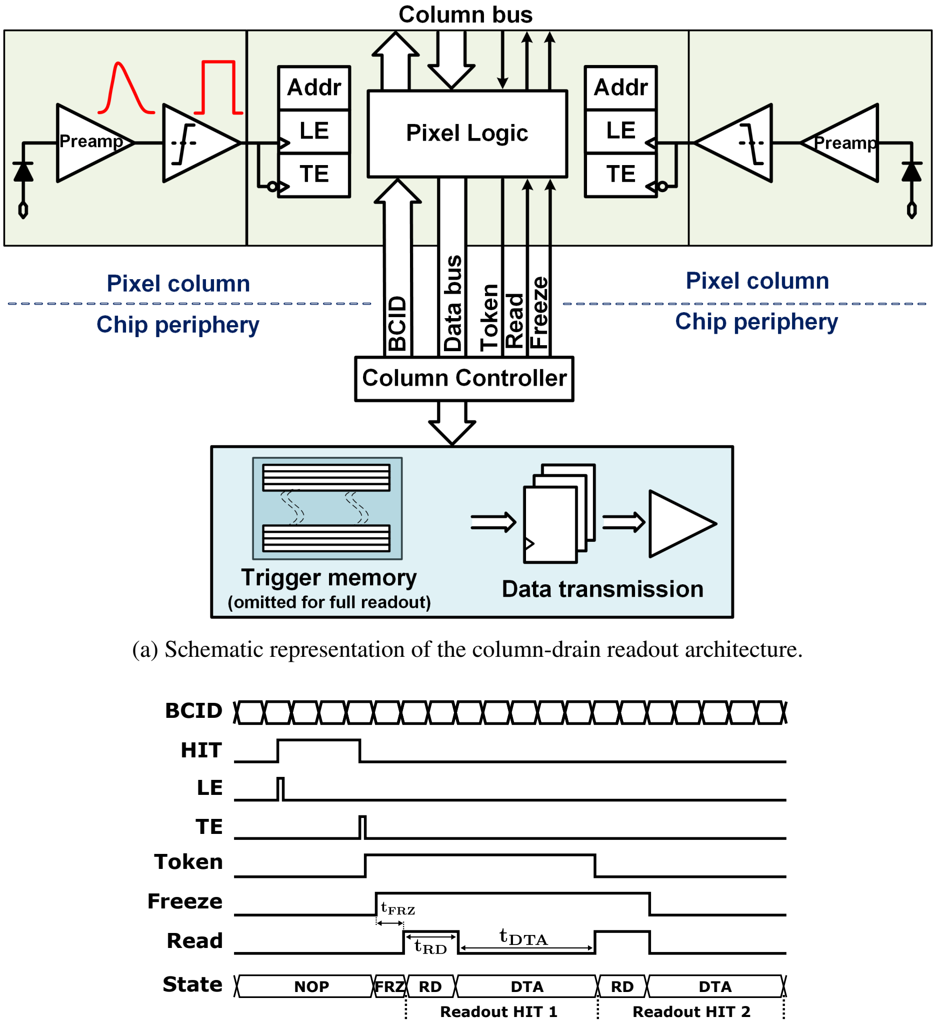
To cope with high hit rates, a capable, zero-suppression readout is required. The column-drain readout architecture, derived from the ATLAS FE-I3 ROIC [52], has been selected to be implemented in TJ-Monopix due to the following advantages:

- Fast readout with ToT capability, proven by the ATLAS FE-I3 ROIC.
- Capability of handling the expected hit rate of ATLAS ITK outer layer (100 MHz/cm^2).
- Simple implementation, therefore smaller pixel size and reduced crosstalk due to less switching activity.

The TJ-Monopix column-drain readout implementation is based on the experience gained from the LF-Monopix1 [53], which is the first fully monolithic DMAPS chip to employ this architecture. The operating principle of the column-drain readout is illustrated in Fig. 4.8. The in-pixel circuitry consists of Random Access Memory (RAM) cells to store the hit pulse Leading Edge (LE) and Trailing Edge (TE) timing information, a Read-Only Memory (ROM) to store the pixel address and the control and arbitration logic. The LE corresponds to the hit ToA (time of arrival) while the ToT can be calculated from the TE and LE difference. The buffer depth is equal to one, hence only one hit can be stored in the pixel until the hit data is transferred to the periphery. The pixel matrix readout is column based. All pixels of the same column (or double column) share a common column-bus that includes the BCID timestamp, the data-bus (LE,TE and pixel address) and the necessary control signals. The column data-bus can be accessed by one pixel at a time with a pre-defined priority. Therefore, each column can be viewed as a queuing system operating at a specific speed (bandwidth) that transfers hit information to the chip periphery.

The periphery includes the End of Column (EoC) block and Digital Chip Bottom (DCB) circuitry. The EoC block supports the transmission and readout of the column-bus signals in the physical (electrical) level. The digital chip bottom process the hit information and its implementation differs by the choice of a triggered of full readout. In the case of a triggered readout, the received hit data from each column readout operation is stored in a trigger memory (usually with intermediate buffering in-between), where it remains until a trigger arrives or the trigger latency is expired. Hit data belonging to the BCID requested by the trigger is serialized and transmitted to the DAQ, and the rest is discarded. In the case of a full readout, hit data is continuously transmitted. An intermediate memory, such as a

First in First out (FIFO) register file can be used to temporarily buffer hits until they are transferred off-chip. If no indeterminate memory is included, a control logic (similar to the one used in the pixel) is necessary to arbitrate access to the chip data output in a column level which leads to a significant reduction in the total chip readout speed.



(a) Schematic representation of the column-drain readout architecture.

(b) Readout sequence timing diagram. In this example two hits are being processed.

Figure 4.8: Column-drain readout architecture. One hit at a time can be stored in the pixel, which immediately attempts to grab the common column-bus. Access to the column-bus is arbitrated with the help of a priority token, while a column controller controls the readout sequence. Hit data in the periphery is directly transmitted (full readout) or stored in the trigger memory (triggered readout).

The 40 MHz BCID timestamp is distributed across the pixel matrix. The propagation delay through the metal wires causes a dispersion of the BCID phase (timing). Since the total timing uncertainty, including the front-end time-walk and systematic effects such as the BCID propagation delay should be smaller than 25 ns, BCID distribution should be carefully designed to minimize the timing dispersion.

A typical requirement in the case of the LHC is that the BCID timing dispersion should be smaller than $\approxeq 4$ ns. When a hit pulse is produced, the timestamp of its LE and TE are stored in the in-pixel RAM. The readout sequence starts with the TE. A hit flag is set and the token signal is asserted to signify that the pixel is available to be read out. The token signal is used to arbitrate the pixel readout and propagates across the column following the pixel priority. The token at the column output is received by a readout controller which uses two signals, called "freeze" and "read", to control the readout sequence. Freeze is activated during the readout phase and prohibits new hits from disrupting the pixel priority logic. The read signal controls pixel access to the data-bus and times the readout of each hit. While the read signal is active, the pixel with an asserted hit flag that has the highest priority (i.e. no higher priority pixel outputs a token signal) accesses the data-bus and transmits the hit information. At the falling edge of read, hit data is latched at the EoC and is ready to be stored in the trigger memory (triggered readout) or to be serialized and transmitted off-chip (full readout).

The column controller is based on a Finite State Machine (FSM), schematically shown in Fig. 4.9, which is composed of four states called no-operation (NOP), freeze (FRZ), read (RD) and data trasfer (DTA). NOP is the default starting state where the controller stays waiting for the token to start the readout sequence and returns to when there are no more hits to be read out. After the token becomes active, the FSM enters the FRZ state (at the next clock positive edge) and the freeze signal is asserted. After a delay time equal to (t_{FRZ}), the FSM enters the RD state and the read signal is produced. It remains in the RD state according to the desired read duration (t_{RD}) and then enters the DTA state. During the DTA state (t_{DTA}), the hit data is stored in the periphery or directly transmitted off-chip. At the end of the DTA phase, the token signal value is examined. If the token is still active (more hits to readout) the controller moves back to the RD state, otherwise it returns to the NOP state and the column freeze is withdrawn. The duration of each state (except NOP) is selected according to the timing constraints derived from the column and periphery specifications and is realized by a delay counter. The following guidelines and constraints apply to the readout sequence timing, shown in Fig. 4.8 (b), and should be taken into account by the controller design:

- The token signal propagates through each pixel with lower priority until it arrives at the EoC accumulating a propagation delay that depends on the pixel position, the priority logic gate delay and the layout parasitics. Until the freeze signal is applied, new hits can still be recorded. After the freeze signal is asserted and before the first read, the token state at each pixel needs to settle in order to avoid issues such as two pixels accessing the data-bus at the same time. Therefore, $t_{\mathrm{FRZ}} \geq t_{\mathrm{del}}^{\mathrm{token}} + t_{\mathrm{del}}^{\mathrm{freeze}}$, where $t_{\mathrm{del}}^{\mathrm{token}}$ is the token delay of the highest priority pixel and $t_{\mathrm{del}}^{\mathrm{freeze}}$ is the freeze signal propagation time through the column.
- The read pulse duration depends on the data-bus speed, i.e. the time that is required until the data-bus signal amplitude at the EoC is high enough to be detected with no error by the sense amplifiers. Therefore, $t_{\mathrm{RD}} \geq 1/BW_{\mathrm{bus}} + t_{\mathrm{del}}^{\mathrm{read}}$, where BW_{bus} is the data-bus speed (bandwidth) and $t_{\mathrm{del}}^{\mathrm{read}}$ is the read signal propagation time through the column.
- The DTA phase duration is depends on the readout type. In the case of a triggered readout, 1 BCID clock cycle (25 ns) is enough to store the hit data in the trigger memory. In the case of a full readout, $t_{\mathrm{DTA}} \geq n_{\mathrm{bit}} \cdot 1/f_{\mathrm{TX}} + t_{\mathrm{del}}^{\mathrm{ser}}$, where n_{bit} is number of data bits to be transmitted per hit, f_{TX} is the output data link frequency and $t_{\mathrm{del}}^{\mathrm{path}}$ is the propagation delay of the data path to the serializer. During the DTA phase, the data-bus baseline is restored, a process that typically requires ≤ 25 ns to complete.

- Since the pixel hit flag is cleared at the start of the read phase, if $t_{\text{del}}^{\text{token}} \geq t_{\text{RD}} + t_{\text{DTA}} - 25\,\text{ns}$ an extra (empty) read cycle will be added to the end of each readout sequence. Therefore, in the case of a triggered or buffered readout, where t_{DTA} is small, the token propagation delay must be optimized.

The maximum column readout speed (or else column bandwidth BW_{col}) is determined by the maximum number of read cycles per unit of time, assuming that data at the end of the read phase can be stored in the periphery as fast as possible (1 clock cycle). Assuming a 40 Mhz base block and that the data-bus baseline is restored within 25 ns after the read phase, the column bandwidth is equal to:

$$BW_{\text{col}} = \frac{1}{t_{\text{RD}} + 25\,\text{ns}} \tag{4.1}$$

The actual readout speed is equal to:

$$f_{\text{readout}} = \frac{1}{t_{\text{RD}} + t_{\text{DTA}} + \frac{t_{\text{OH}}}{\langle N_{\text{hit}}^{\text{RO}} \rangle}} \approxeq \frac{1}{t_{\text{RD}} + t_{\text{DTA}}} \tag{4.2}$$

where t_{OH} is the overhead time of each readout sequence until start of the first read cycle due to the token propagation and FSM synchronization and $\langle N_{\text{hit}}^{\text{RO}} \rangle$ is the average hit number to be read out in each readout sequence.

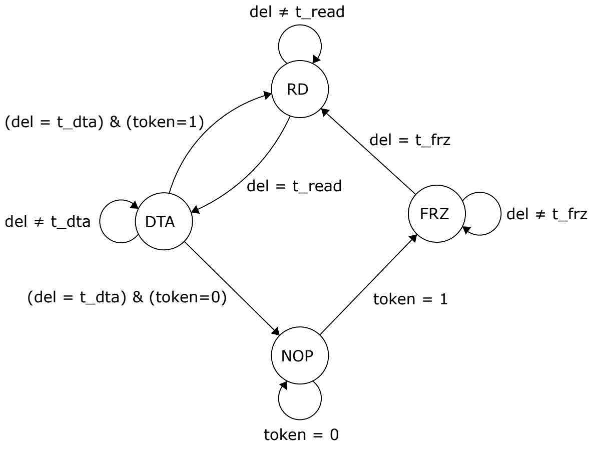


Figure 4.9: Diagram of the column controller finite state machine (FSM).

4.2.1 Hit rate capability simulation

The column-drain readout architecture has been proven by the FE-I3 ROIC to be capable of coping with the hit rate of the ATLAS ID inner layers, which is similar to the expected hit rate of the ATLAS ITk outer layer. However, since the TJ-Monopix and FE-I3 pixel parameters are different, architecture efficiency simulations were carried out [73] in order to validate the hit rate capability of the TJ-Monopix column-drain implementation . The purpose of this study is to determine the raw TJ-Monopix pixel matrix performance without being limited by the hit processing ability at the periphery (e.g. full readout) and determine the feasibility of a complete ATLAS-ready chip based on the same pixel matrix. Therefore, a full-size chip with triggered readout and no trigger memory pile-up is assumed. Each column has a dedicated readout (RO) unit that consists of the EoC circuitry and the column controller. Hit data is stored in the trigger memory within 25 ns (t_{DTA} = 25 ns).

The architecture simulation is based on a simple framework built using the Python programming language. Two different types of simulations were carried out depending on the provided input hit event data. Initially a random hit injection was performed in order to determine the architecture efficiency for different input rates and investigate the limit of the column-drain readout implementation. Subsequently, a hit event dataset generated by a physics monte-carlo simulation of the ATLAS ITk detector was used to verify the TJ-Monopix pixel matrix readout performance in an environment as close as possible to the experiment.

The readout efficiency is measured by comparing the number of injected events to the number of hits having been successfully read out. The simulation framework provides the possibility to decompose the total data loss into the individual contributions due to different hit loss mechanisms in order to provide a better insight of the simulation results. The following hit loss mechanisms can be classified (excluding the trigger memory pileup):

- *Analog pile-up*: Hit loss due to analog pile-up occurs when a new hit arrives during the pre-amplfier response to the previous event, and is therefore not registered separately. In the context of the architecture simulation, to simplify the system and reduce the simulation time, the injected hit data encapsulates charge information translated to a certain hit pulse width (ToT) value. Therefore a dedicated pre-amplifier and discriminator model is not included. Whether analog pile-up occurs is determined based on the ToT value of the previous hit and can be represented by a non-paralyzable model. The analog dead-time model type (paralyzable or non-paralyzable) does not affect the simulation accuracy since the expected hit rate per pixel is low due to the small TJ-Monopix pixel size.
- *Digital pile-up*: Hit loss due to digital pile-up occurs when a new hit arrives while the information of the previous hit has not yet been transferred to the periphery.
- *Data loss due to late copy*: Hit loss due to late copy happens when the hit data stays long enough in the pixel before being read out and the BCID counter rolls over. Therefore, even though the hit is read out, it is assigned a wrong timestamp in the periphery and can be considered as "noise". Apart from the column size and bandwidth, data loss due to late copy depends on the resolution (number of bits) of the BCID timestamp that is distributed to the pixel matrix and the ToT distribution since the hit ToA is recorded at the hit pulse LE while the readout sequence starts at the TE.

A theoretical description of the column-drain readout can be given based on queuing theory. Each (double) column can be described as an M/D/1 queue with a single server and random customer arrival time and to be viable its bandwidth must exceed the incoming hit rate per column. Therefore, for a given hit rate, the distribution of the hit waiting time in the pixel can be analytically calculated. Hits with waiting time $t_{\mathrm{wait}} > 2^{n_{\mathrm{BCID}}} \cdot 25\,\mathrm{ns} - \mathrm{ToT}$, where n_{BCID} is the BCID number of bits will be lost due to late copy. The mean waiting time $\langle t_{\mathrm{wait}} \rangle$ can be used to calculate digital pile-up by following a non-paralyzable model.

A summary of the parameters used in the architecture simulation is given in Table 4.1. Since high ToT values can lead to increased data loss due to late copy, it is beneficial to clip the ToT so that it cannot be higher than a maximum value. Clipping can be achieved in the analog domain by modifying the pre-amplifier circuit (see section 4.5.1.1). In the case of random hit injection all hits induce the maximum ToT (after clipping) equal to 16 bunch crossing (BX) periods (400 ns), which is a worst case senario. In the case of monte-carlo physics based hit injection, all ToT values up to the clipping point of 16 BX are possible.

Table 4.1: Column-drain readout architecture simulation parameters

Parameter	Value
Matrix size	512×512 pix.
Pixel size	$36 \times 40\,\mu\text{m}^2$
BCID resolution	6-bit
Column bandwidth	10 – 20 MHz
Hit ToT	16×25 ns

Simulation with random hit injection In this simulation scenario, input events are produced by a random hit generator with a specified hit rate, hence following a Poisson distribution. Simulation results for $BW_{\text{col}} = 20$ MHz and different input hit rates are depicted in Fig. 4.10. The total data loss remains very low (below 1%) until the hit rate becomes ≊ 600 MHz/cm^2, and steeply rises for higher hit rates. Below 600 MHz/cm^2, data loss is almost equal to analog pile-up contribution, while after 600 MHz/cm^2 it is dominated by late copy and to a lesser extent digital pile-up. If the column bandwidth is reduced to 10 MHz, the matrix can still cope with hit rates up to approximately 250 MHz/cm^2 without any significant hit loss. Therefore, the hit rate capability significantly exceeds the ATLAS ITK L4 expected rate of 100 MHz/cm^2. A similar study of the FE-I3 column-drain implementation can be found in [74]. Due to the larger pixel size, analog pile-up in this case becomes more important.

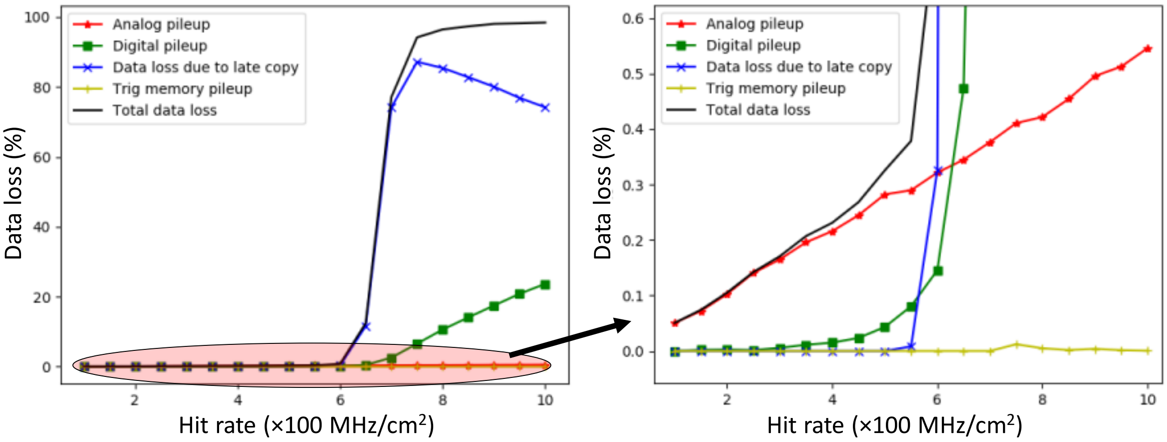
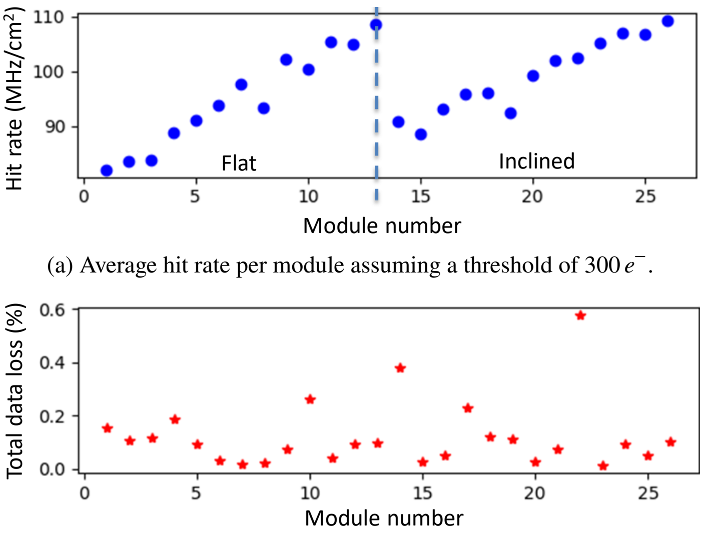


Figure 4.10: TJ-Monopix column-drain architecture efficiency simulation with random hit injection. The column bandwidth is set to 20 MHz [73].

Simulation using physics monte-carlo hit data A more realistic simulation requires an input hit event dataset, representative of the ATLAS experiment, which can be generated by performing a physics based monte-carlo simulation using specialized software such as the ATHENA framework [75]. An ATLAS ITk L4 event dataset using the TJ-Monopix1 pixel parameters ($36 \times 40\,\mu m^2$ and 25 µm depletion) has been provided by E. Zaffaroni (University of Geneva) and includes data from 26 barrel modules (flat and inclined). The average hit rate, after applying a threshold "cut" equal to $300\,e^-$, varies between $\cong 75 - 110\,\mathrm{MHz/cm^2}$ depending on the module type and position along the beam axis as shown in Fig. 4.11(a).

The architecture simulation is performed assuming a column bandwidth of 20 MHz and a maximum (clipped) ToT equal to $16\,BX$. The total data loss per module is shown in Fig. 4.11(b) and remains in all cases well below 1%. The data loss is dominated by hit patterns mainly caused by energetic delta electrons. If a delta electron traverses the column, it can activate a significant number of pixels of the same column causing a temporary data loss. However, these events are usually not interesting in terms of physics and can be discarded. Additional logic to "clear" a column saturated by such events can be implemented, but would lead to an increase in pixel area and complexity.



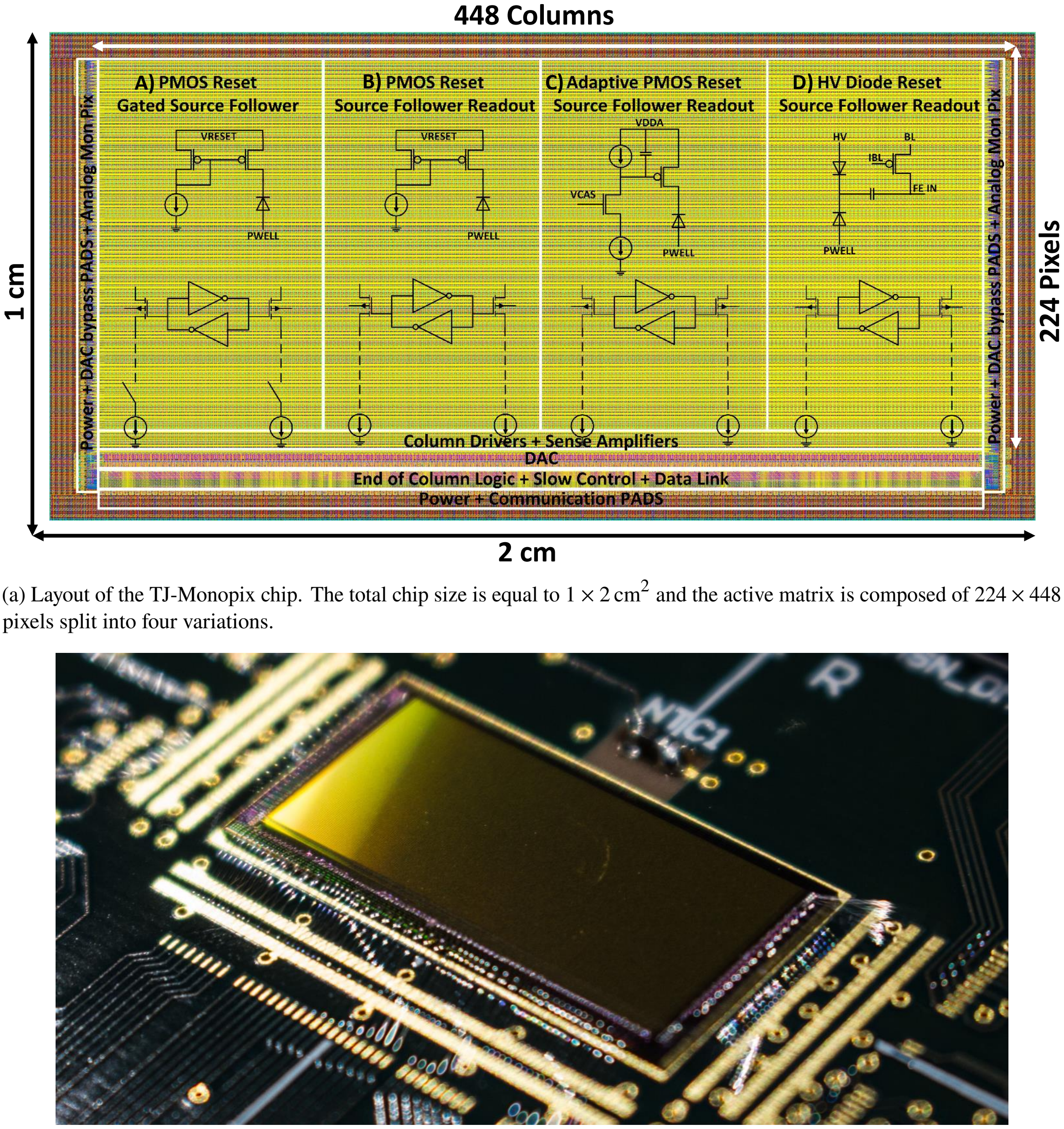
(a) Average hit rate per module assuming a threshold of $300\,e^-$.

(b) Total data loss per module.

Figure 4.11: TJ-Monopix column-drain architecture efficiency simulation using an ATLAS ITk L4 monte-carlo event dataset [73].

4.3 Design of the TJ-Monopix1 demonstrator chip

The TJ-Monopix1 demonstrator chip is the first small collection electrode DMAPS featuring a standalone column-drain readout architecture. It has been designed according to the requirements of the ATLAS ITk outer layer (see section 2.2.2) to explore the feasibility of a large scale pixel matrix based on the modified TowerJazz 180 nm process for high rate and high radiation environments. The chip layout and floorplan are illustrated in Fig. 4.12. The pixel matrix contains 224×448 pixels of $36 \times 40\,\mu m^2$ size yielding a total active area approximately equal to $145\,mm^2$. The small sensor capacitance and hence high SNR is exploited by a compact, low-power analog front-end (FE) which has been designed to operate at a nominal threshold of ≊ 300 e^-. The pixels are arranged in a double column layout consisting of 112 2×2 pixel cores (448 pixels) each, in order to utilize the pixel area more efficiently and reduce crosstalk interference. The column design is capable of achieving 20 MHz bandwidth, but is normally operated at a more conventional 10 MHz speed (50 ns read duration). The 40 MHz BCID timestamp is gray-encoded with 6-bit resolution and analog charge information is measured using the time over threshold technique (6-bit ToT).



(a) Layout of the TJ-Monopix chip. The total chip size is equal to $1 \times 2\,\mathrm{cm}^2$ and the active matrix is composed of 224×448 pixels split into four variations.

(b) Manufactured TJ-Monopix1 chip, photographed on the carrier PCB.

Figure 4.12: The TJ-Monopix1 fully monolithic, small collection electrode DMAPS prototype.

The pixel matrix is split into four sectors (flavors) in order to explore different variations of the data-bus readout circuitry and front-end input reset and coupling. Additionally, the front-end analog output of four special pixels at each side, placed next to the matrix, is buffered and can be monitored for characterization and debugging purposes. Each flavor is independent having a separate readout and data transmission. Flavors B,C and D feature a "source-follower" column-bus readout derived from the LF-Monopix1 chip [53] that aims to reduce crosstalk. However, it leads in a significant amount of static power being drawn at the EoC periphery. To eliminate static power consumption, a modified gated source follower readout has been included in flavor A, which is otherwise identical to flavor B. Flavor B, refereed to as the "PMOS reset" flavor, is the standard (reference) variation and features a DC-coupled PMOS input reset. Flavor C incorporates a novel leakage compensation circuit and flavor D, also called the "HV" flavor, explores the possibility of applying a high front-side bias voltage to the collection electrode which in this case is AC-coupled to the front-end input. Additionally, throughout the whole matrix, pixels are split in two variations across the column according to their deep p-well coverage. The electronics area of pixels belonging to the bottom half of each column (rows 0 to 111) is fully covered by deep p-well (FDPW), while part of the deep p-well is removed (RDPW) for pixels belonging to the top half of the column (rows 112 to 223) in order to enhance the lateral electric field component.

The chip periphery includes all the required support blocks to bias, configure, test and readout the pixel matrix. A 7-bit Digital to Analog Converter (DAC) is used to generate the analog bias voltage and current levels. No trigger memory is implemented in this prototype, hence the readout is full (continuous). Pixel data is transferred to the End of Column (EoC) logic, serialized and then immediately transmitted. Digital Input/Output (I/O) CMOS transceivers, operating at 40 MBps, are used for communication and data transmission to the external DAQ testing system. Chip control and configuration is based on a simple Serial Peripheral Interface (SPI) implementation. Power delivery is split in four domains: 1) matrix analog, 2) matrix digital, 3) biasing DAC, 4) periphery in order to minimize noise coupling through the power grid. Matrix power pads are distributed at the sides, such that the analog static voltage drop gradient is horizontal, which simplifies voltage drop compensation scheme layout implemented by the biasing circuit (see section 4.3.2.3). The total power consumption is approximately equal to 280 mW, excluding the source follower readout static power which can be eliminated using the gated source follower variation. A decomposition of the different power contributions is presented in Table 4.2.

Table 4.2: TJ-Monopix1 power consumption summary

Contributor	Power consumption
Matrix analog	70 mW/cm^2
BCID distribution	60 mW/cm^2
Matrix readout dynamic (ATLAS ITk L4 hit rate)	< 5 mW/cm^2
Column-bus readout static (flavors B,C,D)	190 mW/cm
Periphery	20 mW

4.3.1 Pixel design and layout

The layout of the 2×2 pixel core is shown in Fig. 4.13. It is comprised of the sensor (collection electrode) structure, the analog front-end and the digital readout logic. The collection electrode size is 2 µm in diameter and its spacing from the electronics area is equal to 3 µm, following the guidelines mentioned in section 4.1. In order to reduce possible crosstalk, the digital and analog areas are physically separated and layout techniques have been employed to effectively shield the sensitive nodes. Apart from providing isolation, grouping the analog and digital circuitry area of each four neighboring pixels, simplifies the pixel layout and routing and reduces the pixel size. Although the analog and digital power domains are separate, the p-well substrate housing every NMOS transistor in the pixel is common since deep n-well isolation is not allowed in the pixel matrix area as it would compete charge collection. Therefore, the data-bus readout circuit is designed to limit the digital peak transient currents and suppress noise due to substrate coupling using the source-follower readout scheme. All four pixels belonging to the same pixel core share the same column-bus and the two physical columns (left and right) constitute a double column, which is logically (from a readout perspective) equivalent to a single column with double the amount of pixels. The 2×2 pixel core silicon well structure is illustrated in Fig. 4.14 and varies according to the deep p-well geometry. While the electronics area of FDPW pixels is fully covered by deep p-well, RDPW pixels have a portion of the deep p-well removed, which is constrained by the n-well required for PMOS transistor fabrication. Due to the electronics layout, the deep p-well removal is not symmetric in the analog and digital area around the collection electrode.

The analog front-end (FE) is based on a novel voltage pre-amplifier design and has been optimized for fast timing response (< 25 ns), low noise and low power consumption. Due to the relatively low simulated threshold dispersion ($\sigma_{\mathrm{THR}} \cong 25\, e^{-}$), a threshold trimming DAC has not been included in this prototype to reduce complexity and area. The digital readout logic includes the 6-bit LE and TE memory (12 SRAM cells per pixel), a 9-bit pixel address ROM and the readout (R/O) control logic. To optimize the token propagation speed and reduce the total number of gates, the token pass logic follows a NAND-NOR design. Therefore, "NAND" and "NOR" type pixels alternate across the double column. The BCID timestamp, data lines and read/freeze control signals use differential signaling in order to reduce electromagnetic interference, reject common mode noise and effectively double the signal to noise ratio. As a result, the BCID timing uncertainty (jitter) is reduced and the data-bus speed (column bandwidth) is increased. The drawback of differential signaling is the higher routing complexity that requires double the amount of metal wires and can impact the pixel size.

Additionally, test features required for the correct operation, characterization and debugging are implemented. Noisy pixels due to process variations and defects can be masked using a three vector (horizontal, vertical and diagonal) projection scheme. Hit events due to charged particles can be emulated by using an injection capacitance driven by a pulsing circuit with configurable pulse height. Furthermore, the discriminator output of pixels belonging to the same physical column are connected through an OR logic (HITOR) with row and column enable i order to be directly measured by an oscilloscope or a Field Programmable Gate Array (FPGA) for precise timing measurements. The remaining pixel area is filled with power decoupling capacitors constructed by MOSFET transistor structures using the thin gate oxide of as a dielectric (MOSCAP).

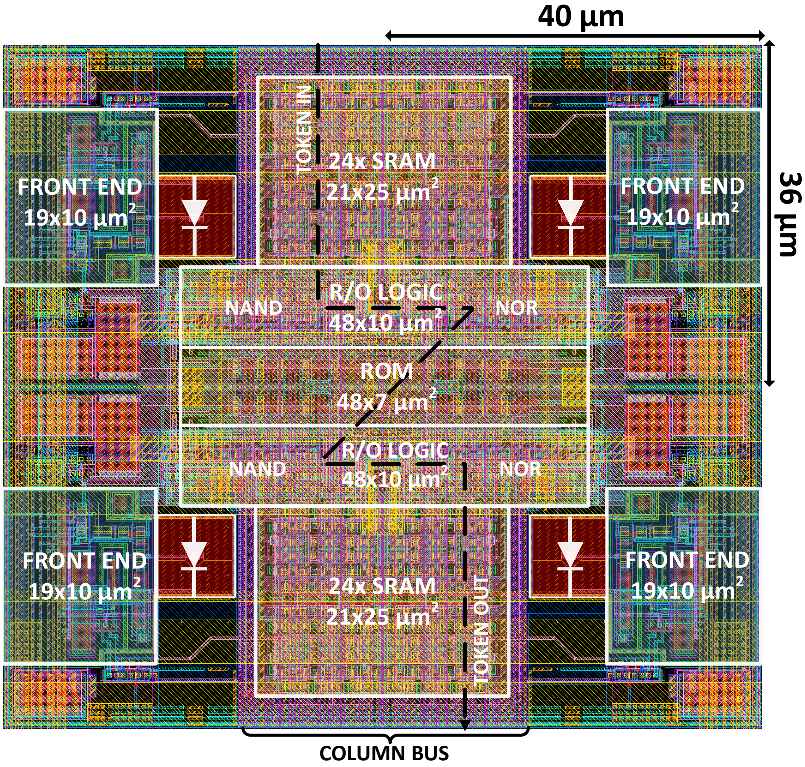


Figure 4.13: Layout of a TJ-Monopix1 2 × 2 pixel group. The sensitive analog front-end region is separated from the common digital area in the pixel group center to suppress crosstalk interference.

4.3.1.1 Front-end design

The prominent advantage of the TJ-Monopix small collection electrode concept is the small detector capacitance (C_D = 3 fF) which results in a high input signal amplitude $V_{\text{in}} = Q_s/C_{\text{in}}$, where $C_{\text{in}} \cong C_D$ represents the sum of all capacitances at the front-end input node. In the case of a fully depleted modified process sensor with 25 µm thick epitaxial layer (approx. 20 µm deep sensing volume), the expected charge MPV is approximately 1 600 e^-. Hence, for $C_{\text{in}} \cong 4$ fF, V_{in} will be equal to:

$$V_{\text{in}} = \frac{1\,600\,e^- \cdot 1.6 \cdot 10^{-19}\,\text{C}}{4\,\text{fF}} = 64\,\text{mV} \tag{4.3}$$

Therefore, the input voltage amplitude, developed on the sensor capacitance, is large enough to be directly amplified by even a simple, one-stage voltage amplifier and achieve a high signal to noise ratio (SNR). Considering that the power supply voltage is 1.8 V, a moderate voltage gain $A \approx 25$ is sufficient as the amplifier maximum voltage swing at the output would be met or exceeded for an input charge equal to the MPV. Thus, due to the low gain and constant gain-bandwidth (GBW) product, the amplifier can be fast while keeping power consumption low. Furthermore, due to the high output voltage amplitude, which is can be significantly higher than the transistor threshold, the discriminator design can be simple and power-efficient while having a fast timing response (< 25 ns).

The TJ-Monopix analog front-end is an evolution of the design used by the ALPIDE chip [76] and

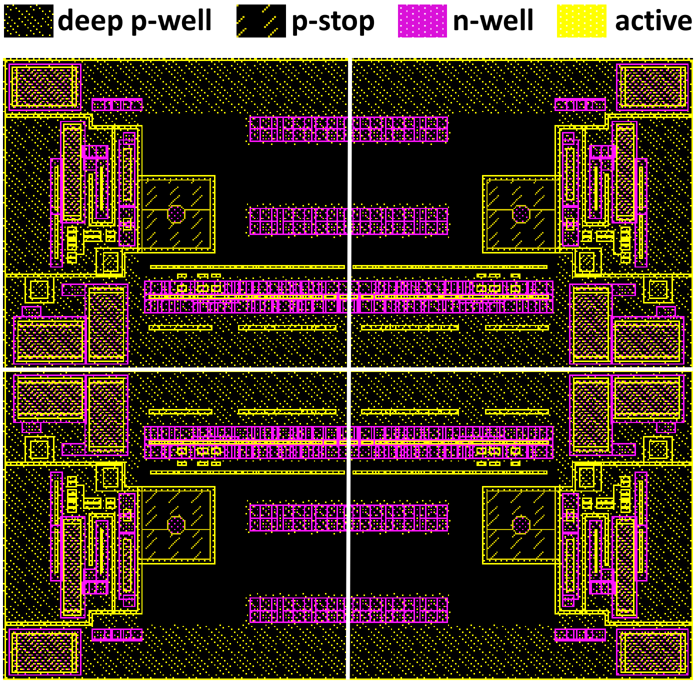


Figure 4.14: Top-view of a TJ-Monopix1 2 × 2 pixel group silicon well layout. A p-stop mask is used to create the spacing around the n-well collection electrode. The pixel electronics area of pixel belonging to the bottom half of each column is fully covered by deep p-well (FDPW), while part of the deep p-well is removed (RDPW) for pixels belonging to the top half of each column, as shown. The "active" mask marks the area where the shallow trench isolation (STI) is removed. The STI local density can affect the implantation depth of the modified process n$^-$ layer.

is based on a voltage amplifier concept shown in Fig. 4.15. Since the pre-amplifier input impedance is very high (MOSFET gate), the collected charge Q_s is essentially integrated on the small input capacitance $C_{\text{in}} \cong C_D$. The output voltage will then be equal to the product of the input voltage and the amplifier gain:

$$V_{\text{out}} = \frac{Q_s}{C_{\text{in}}} A_v(\omega) \tag{4.4}$$

In practice, the pre-amplifier implementation is more involved than a simple voltage gain stage since it has to include an input and output voltage baseline reset mechanism and signal shaping in order to reduce the noise bandwidth. The input baseline is restored by a reset element, which can be a simple diode, a PMOS transistor or a more complex circuit that also compensates for the sensor leakage current. A low frequency (LF) feedback is used to control the output baseline reset by slowly adjusting the gain stage input (V_{GN}) with a configurable time constant that determines the ToT resolution (Q_s-ToT slope). A capacitor C_c provides DC isolation required by the LF feedback in order to independently set the V_{GN} DC potential. Because multiple transistor terminals are connected to the gain stage input due to the LF feedback, its total capacitance is high. Therefore connecting the input directly to the gain stage through C_c would result in a substantial increase of C_{in} that nullifies the advantage of a small detector capacitance. To solve this issue, a source follower is used to buffer the input voltage and provide isolation from the rest of the circuit. Thus, only the small gate-to-drain (C_{gd}) capacitance of the source follower input transistor is added to the input and high signal amplitude is maintained. The C_c capacitor additionally serves as a high-pass filter (along with the G_N node input resistance) that in combination with the gain stage low-pass response behave like a shaper that optimizes the SNR by filtering frequencies outside its bandwidth.

It is useful to compare the TJ-Monopix front-end amplifier concept with an implementation based on a CSA, which is the most widely used amplifier type for particle detection, in order to understand why a voltage amplifier is better suited for this application. As mentioned in section 3.1.2.3, the CSA pre-amplifier operation is based on the integration the generated charge on the feedback capacitance C_f. Therefore, the CSA is very effective in the case of a relatively large detector capacitance because the charge is not integrated on C_D but on the much smaller C_f. Furthermore, the amount of generated charge in these applications is typically significantly higher compared to TJ-Monopix due to the use of a thick, planar sensor. As a result, the output voltage $V_o = Q_s/C_f$ is sufficiently high to be directly used by the discriminator a a second amplification stage. A voltage pre-amplifier, in comparison, is not feasible in this case since the input voltage amplitude would be comparable to noise due to the high C_D.

In contrast, if a CSA pre-amplifier was used by TJ-Monopix, the feedback capacitance would have to be smaller than $C_D \approxeq 3$ fF in order to provide a charge to voltage conversion gain Q_s/C_f higher than the intrinsic gain of the sensor Q_s/C_D. Such a small capacitance value (in the order of a few hundred aF) is not practical and is difficult to be accurately controlled because it is similar, or even smaller than the transistor (gate, junction) and layout parasitic contributions. Furthermore, if C_f is comparable to C_D, a non-negligible residual charge that does not contribute to the output voltage can remain on the detector capacitance. Therefore a CSA approach would have a limited gain, high threshold dispersion and would require a second amplification stage or a more complex discriminator which results in increased power consumption. The TJ-Monopix analog front-end is a superior option because charge integration (Q-V) inherently occurs on the small detector capacitance, yielding a high SNR without the need for a CSA. The voltage amplifier (V-V) acts essentially as a second stage that further amplifies the input signal in order to be used by a simple one-stage discriminator rendering the front-end design compact and power efficient.

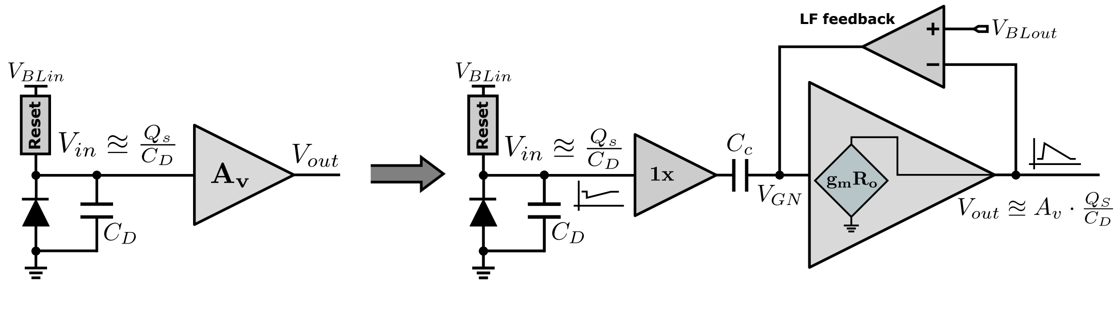

V_{BLin}
Reset
$V_{in} \approxeq \frac{Q_s}{C_D}$
$\mathbf{A_v}$
V_{out}
C_D
LF feedback
+
−
V_{BLout}
V_{BLin}
Reset
$V_{in} \approxeq \frac{Q_s}{C_D}$
1x
C_c
V_{GN}
g_mR_o
$V_{out} \approxeq A_v \cdot \frac{Q_S}{C_D}$
C_D

Figure 4.15: The analog front-end voltage pre-amplifier concept used in TJ-Monopix. The input charge is integrated on the sensor capacitance, and the resulting voltage is amplified while a low frequency feedback is used to restore the output baseline.

Input coupling and baseline reset The collection electrode can be either DC-coupled directly to the readout electronics or AC-coupled through a metal-oxide-metal (MOM) capacitor. AC-coupling allows high positive bias voltage (HV) to be applied to the collection electrode, but suffers from a signal loss penalty mainly due to additional parasitic capacitance introduced at the sensitive input node. Therefore, DC-coupling is the standard approach because it results in a higher input signal amplitude. However, the collection electrode voltage is limited in this case by the maximum gate voltage allowed by the technology ($V_{\mathrm{dd}} = 1.8\,\mathrm{V}$), and is in practice somewhat lower ($\approxeq 1\,\mathrm{V}$) since it sets the front-end input transistor DC operating point. The detector capacitance is slowly discharged through the reset element resistance and the input baseline is restored. Therefore the input voltage is equal to:

$$V_{\mathrm{in}} = \frac{Q_s}{C_{\mathrm{in}}} e^{-t/R_b C_{\mathrm{in}}} \tag{4.5}$$

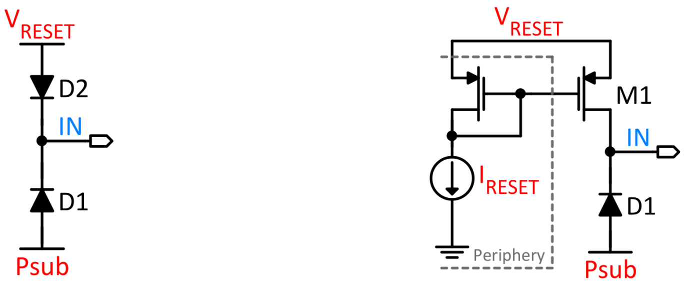
where R_b is the equivalent reset element small signal resistance. The input discharge time constant $\tau_{\mathrm{IN}} = R_b C_{\mathrm{in}}$ must be longer than the pre-amplifier response time in order to avoid signal loss due to the input voltage being reset before the pre-amplifier output reaches its peak (an effect similar to ballistic deficit). Additionally, τ_{IN} should be long enough to avoid limiting the ToT slope by forcing the output to reset faster than the time constant set by the LF-feedback, but sufficiently short in order to keep analog pile-up low. Thus, for $C_{\mathrm{in}} = 4\,\mathrm{fF}$, R_b should be in the order of several GΩ.

Input DC coupling In the DC coupled case, without leakage compensation, continuous baseline reset can be implemented by a forward biased diode or a PMOS transistor as shown in Fig. 4.16. When no charge is collected and the input capacitance is completely discharged, the reset diode (D_2) is biased by the sensor D_1 leakage current I_{leak} which is < 1 pA before irradiation. The input voltage V_{in} will be equal to V_{RESET} (generated by the DAC) minus the forward voltage drop which, following the diode exponential I-V characteristic, is approximately 500 mV for a very low biasing current. The reset diode small signal resistance is equal to:

$$r_D = \frac{nV_T}{I_Q} \tag{4.6}$$

$V_T = kT/q_e$ is the thermal voltage where k is the Boltzmann constant, T is the temperature and q_e the electron charge. n is the diode ideality factor, which for silicon diodes is approximately 1 to 2 and I_Q is the DC bias current which is equal to the sensor leakage. After a particle is detected, the input capacitance is discharged to the baseline voltage through D_2. The simulated transient response of V_{in} for two different leakage current values (1 pA and 50 pA) and two different input charges ($300\,e^-$ and $1\,600\,e^-$) corresponding to the threshold and the most probable value (MPV) is shown in Fig. 4.17. Assuming a linearized diode model, which is a good approximation for small amplitudes, the discharge time constant will be equal to $C_{\mathrm{in}} r_D$, which depends on the sensor leakage current. However, since the diode is a non-linear element, the discharge rate also depends on the collected charge Q_s which leads to a non-linear Q_s − ToT relationship. The advantage of diode reset is its simplicity as it is implemented by a p$^+$ diffusion inside the n-well collection electrode. Thus, it only contributes approximately 0.3 fF due to the p-n junction capacitance to C_{in}.

To avoid the strong dependence on the sensor leakage current and achieve a linear Q_s − ToT relationship, a PMOS transistor can be used instead of a diode. In a first order approximation, the PMOS device acts as a constant current source. Since the discharge current has to be very low (in the



(a) (b)

Figure 4.16: DC-coupled input reset by using a) a forward biased diode and b) a PMOS transistor

order of pA), the transistor operates in weak-inversion and its current is given by [77]:

$$I_D = \frac{W}{L} I_0 e^{\frac{V_{\mathrm{gs}}-V_{\mathrm{TH}}}{mV_T}} \left(1-e^{\frac{V_{\mathrm{ds}}}{mV_T}}\right) \approxeq \frac{W}{L} I_0 e^{\frac{V_{\mathrm{gs}}-V_{\mathrm{TH}}}{mV_T}} \tag{4.7}$$

where W/L is the transistor aspect ratio, V_{gs} and V_{ds} are the gate-to-source and drain-to-source voltages, V_{TH} is the threshold voltage, V_T is the thermal voltage, I_0 is the current for $V_{\mathrm{gs}} = V_{\mathrm{TH}}$ and m is the slope factor ranging from 1.2 to 2 depending on the technology. The PMOS bias current I_{RESET}, which can be configured through the biasing DAC, must be larger than the sensor leakage current, otherwise the input voltage will be pulled down prohibiting the correct operation of the front-end pre-amplifier. If therefore $I_{\mathrm{RESET}} > I_{\mathrm{leak}}$, the DC input voltage will be approximately equal to V_{RESET}. The discharge current is equal to $I_{\mathrm{RESET}} - I_{\mathrm{leak}}$ and does not depend on the collected charge. The disadvantage of the PMOS reset is the slightly higher capacitance contribution to C_{in} (0.4 – 0.5 fF) and its susceptibility to TID radiation effects due to the small drain current. The simulated transient response for $I_{\mathrm{leak}} = 1$ pA, 50 pA and $Q_s = 300\,e^-, 1\,600\,e^-$ is shown in Fig. 4.18. Due to the linear ToT characteristic, the PMOS reset method has been preferred in the design of TJ-Monopix1.

Leakage compensation Although the PMOS reset is capable of providing a constant current I_{RESET} that is equal to the sum of the discharge and leakage current, it has to be manually re-tuned as I_{leak} varies. To achieve a constant input DC baseline voltage and discharge rate independently of the sensor leakage current, a leakage compensation circuit, has been devised. The operation concept is based on a low-frequency feedback (similar to the one used to reset the pre-amplifier output baseline) that slowly adjusts the gate voltage of a PMOS transistor in order to source a current that is equal to I_{leak}. Due to the low bandwidth, the feedback mechanism does not react to fast transients such as the voltage generated by an impinging particle, keeping this current approximately constant.

The circuit design, shown in Fig. 4.19 (a), is based on a flipped voltage follower (FVF) topology [78]. M_2 acts as a current source that provides a current equal to $I_D^{M_2} = I_{\mathrm{leak}} + I_B$. M_1 is used in a common-gate amplifier configuration that controls the high-impedance node (A) at the gate of M_2. Since a very low bandwidth is required, the biasing current I_B must be small (in the order of pA). The feedback action works as follows: Because the current through M_1 is constant and equal to I_B its

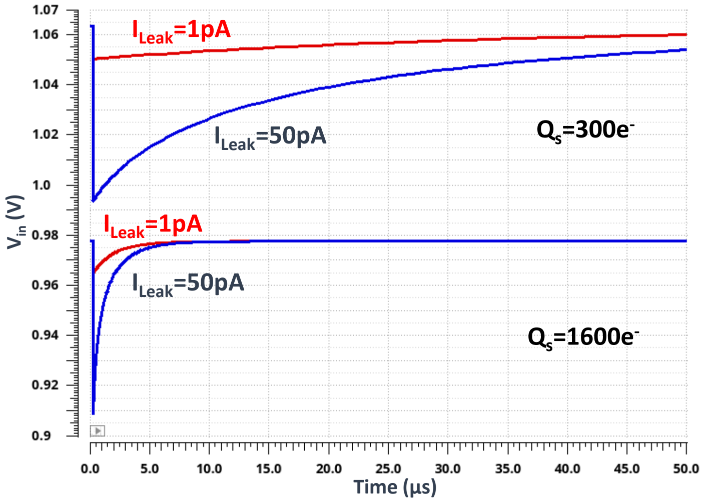


Figure 4.17: Diode reset: Simulated transient response of the input voltage for two different leakage current and collected charge values. V_{RESET} has been set to 1.5 V.

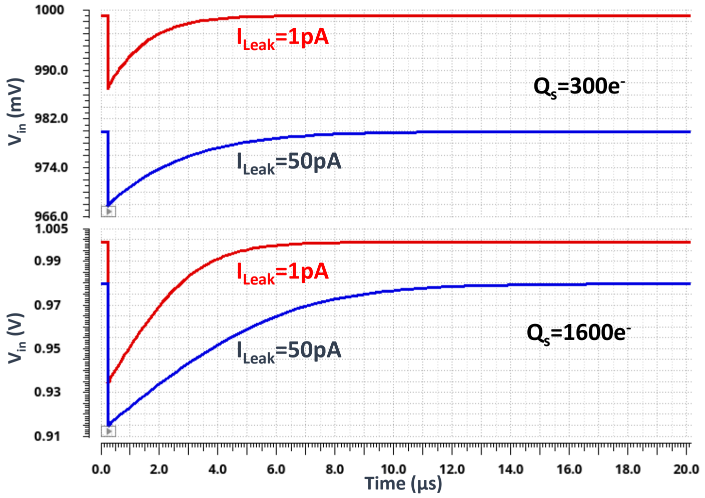


Figure 4.18: PMOS reset: Simulated transient response of the input voltage for two different leakage current and collected charge values. V_{RESET} has been set to 1 V and I_{RESET} is equal to 100 pA.

gate-to-source voltage must also be constant and approximately equal to (weak inversion, eq. 4.7):

$$V_{\mathrm{gs}}^{M_1} \approxeq ln\left(\frac{I_D^{M_1}}{I_0}\left(\frac{L}{W}\right)_{M_1}\right) mV_T + V_{\mathrm{TH}}^{M_1} \tag{4.8}$$

If the leakage current changes, e.g. becomes higher, the input node voltage V_{IN} will start to decrease. Since the gate of M_1 is connected to a constant voltage V_{BL}, $V_{\mathrm{gs}}^{M_1}$ and therefore its drain current will also start to drop. As a result, the gate voltage of M_2 becomes lower and its drain current is increased in order to compensate for I_{leak} causing V_{IN} to rise again. Thus, the input DC voltage is held constant and equal to $V_{\mathrm{BL}} + V_{\mathrm{gs}}^{M_1}$. To improve stability, a compensation capacitor C_c that reduces the bandwidth and slows down the feedback response is added to the high-impedance node A. The loop gain is equal to:

$$G_L = g_{m_1} g_{m_2} Z_{\mathrm{IN}} Z_A \tag{4.9}$$

where Z_{IN} is the impedance at the input node, Z_A is the impedance at node A and g_m is the transistor transconductance. Assuming that the transistor output resistance r_{ds} is high and the compensation capacitor C_c is large enough to dominate Z_A, at high frequencies (compared to the feedback time constant) the closed loop transfer function will be approximately equal to:

$$\frac{i_D^{M_2}}{i_{\mathrm{IN}}} \approxeq \frac{g_{m_1} g_{m_2}}{s^2 C_c C_D + s g_{m_1} C_c + g_{m_1} g_{m_2}} \tag{4.10}$$

For the circuit to be stable, C_c and must be higher than C_D and g_{m_2} must be as low as possible (low W/L ratio). The discharge current, under the same assumptions, flows through C_c and the input resistance at the source of M_1, which is approximately equal to $1/g_{m_1}$. Because the transconductance $g_{m_1} \approxeq I_B/mV_T$ depends on the bias current I_B, it can be used to control the discharge rate.

The input DC voltage must be approximately equal to 1 V, in order to set the operating point of the pre-amplifier input transistor. To keep M_1 in saturation, V_A must be lower than 1 V − V_{DSsat_1}, where V_{DSsat} is the drain to source saturation voltage. However, because M_2 operates in the deep-subthreshold region, its gate voltage (node V_A) will be approximately 300 – 400 mV below the source voltage V_{COMP}. Therefore, the compensation circuit of Fig. 4.19 (a) requires a voltage rail V_{COMP} lower than $V_{\mathrm{dd}} = 1.8$ V. Since all voltage rails of the biasing DAC were used, a folded variation, shown in Fig. 4.19 (b) has been implemented in TJ-Monopix1 to overcome this limitation. By folding, the main branch of the circuit is split into two and the type of M_1 is switched to NMOS. The input DC voltage is equal to $V_{\mathrm{BL}} - V_{\mathrm{gs}}^{M_1}$ and can be set independently from the gate of M_2. M_1 is biased by the current source I_{B1} and M_2 sources a current equal to $I_D^{M_2} = I_{B2} - I_{B1} + I_{\mathrm{leak}}$. Because NMOS transistors can develop parasitic leakage current paths due to TID radiation damage, higher than I_B, M_1 must be an ELT device. As can be observed from the input voltage transient response, shown in Fig. 4.20, the reset rate is independent of the sensor leakage current.

The main disadvantage of the leakage compensation circuit is the significant capacitance increase at the input because it is connected to multiple transistor source or drain junction capacitances. In the case of the folded variation, the signal loss due to the added capacitance at C_{in} is approximately equal to 30%. This high penalty is due to the ELT geometry of transistor M_1 that results in a large source/drain junction area. In contrast, the straight FVF implementation is a better option if an extra voltage rail is available since it contributes a much smaller capacitance at the input.

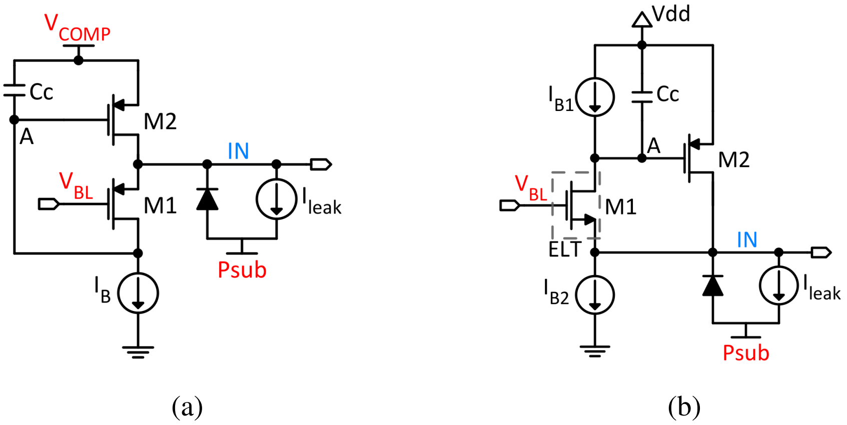


Figure 4.19: Leakage current compensation circuit: a) design based on a flipped voltage follower b) folded variation implemented in TJ-Monopix1.

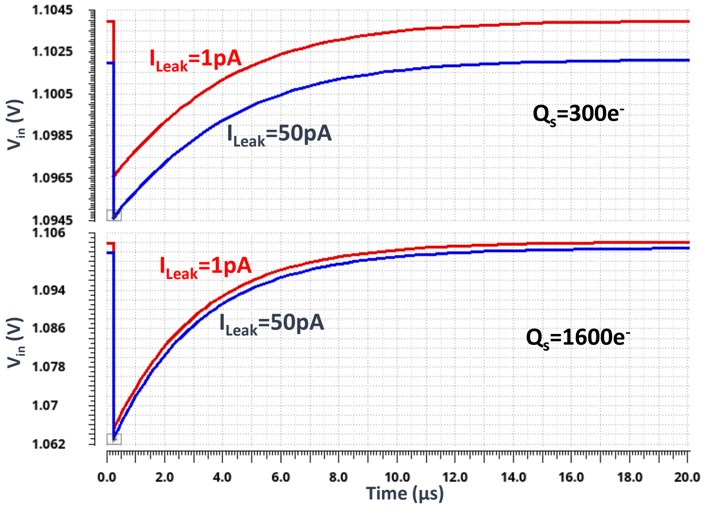


Figure 4.20: Leakage compensation reset: Simulated transient response of the input voltage for two different leakage current and collected charge values. The biasing current is equal to 50 pA.

Input AC coupling In order to apply a positive high voltage (HV) bias at the collection electrode, AC-coupling to the front-end pre-amplifier input is required in order to avoid breakdown of the MOSFET devices. The AC-coupled input circuit implementation is illustrated in Fig. 4.21. The collection electrode biasing voltage is applied by a p$^+$/n-well diode D_2, that also provides the baseline reset current. The AC-coupling capacitor C_{ac} is implemented by a metal-oxide-metal (MOM) structure[2] that makes use of the routing metals and is placed in the sensor area above the collection electrode ($\approxeq 8 \times 8\,\mu\mathrm{m}^2$). The front-end pre-amplifier input DC operating point is set with the help of a the PMOS transistor M_1.

Due to the stray capacitance to the substrate and surrounding metal wires, the MOM structure additionally introduces the parasitic capacitances C_{p1} and C_{p2} between each plate of C_{ac} and the circuit AC ground. Therefore, the total capacitance at the collection electrode node IN and the front-end input node FE_{IN} will be equal to $C_{\mathrm{IN}} = C_D + C_{D_2} + C_{p1}$ and $C'_{\mathrm{FE}} = C_{\mathrm{FE}} + C_{p2}$ respectively, where C_{FE} represents the capacitance at the front-end input including the drain capacitance of M_1. C_{ac} forms a capacitive divider with C'_{FE}, hence the front-end input voltage $v_{\mathrm{FE_{IN}}}$ will be equal to v_{IN} times the coupling ratio r_c:

$$v_{\mathrm{FE_{IN}}} = v_{\mathrm{IN}} r_c = v_{\mathrm{IN}} \frac{C_{\mathrm{ac}}}{C_{\mathrm{ac}} + C'_{\mathrm{FE}}} \tag{4.11}$$

To avoid signal loss due to voltage drop (charge storage) on the coupling capacitor C_{ac}, $C_{\mathrm{ac}} \gg C'_{\mathrm{FE}}$. However a higher C_{ac} value requires a larger MOM structure that will also be closer to the surrounding shielding metals and will lead to higher parasitic capacitance. C_{p1} and C_{p2} are crucial since they can considerably increase the total input capacitance and cause a substantial reduction of the signal amplitude. The equivalent total input capacitance C_{tot} is equal to:

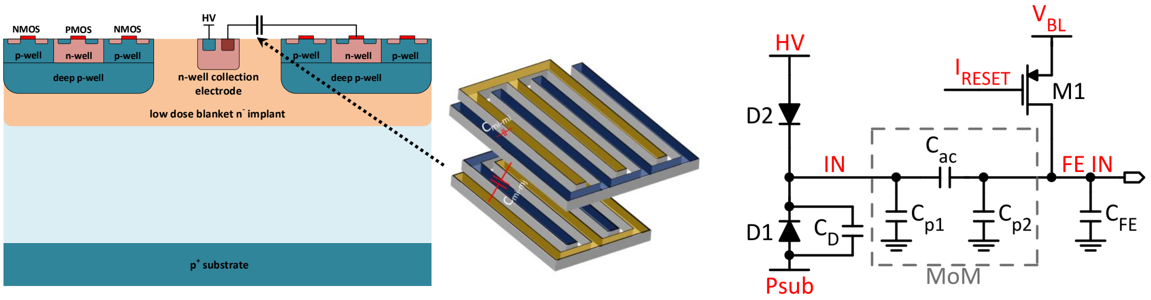
$$C_{\mathrm{tot}} = C_{\mathrm{IN}} + \frac{C_{\mathrm{ac}} C'_{\mathrm{FE}}}{C_{\mathrm{ac}} + C'_{\mathrm{FE}}} \tag{4.12}$$

Therefore, the voltage at the front-end input for a collected amount of charge Q_s is given by:

$$v_{\mathrm{FE_{IN}}} = \frac{Q_s}{C_{\mathrm{tot}}} r_c = \frac{Q_s}{C'_{\mathrm{FE}} + \frac{C_{\mathrm{IN}}}{r_c}} = \frac{Q_s}{C_{\mathrm{FE}} + C_{p2} + \left(C_D + C_{p1}\right) \frac{1}{r_c}} \tag{4.13}$$

For the high voltage biasing scheme to be feasible, the ratio of the coupling to parasitic capacitances $C_{\mathrm{ac}}/(C_{p1} + C_{p2})$ must be high and should be maximized by optimizing the MOM structure layout. To achieve high capacitance density, the MOM capacitor is composed of multiple layers (metal-1 to metal-5) of interleaved metal segments (fingers) as shown in Fig. 4.21 (a). The interleaved finger structure polarity alternates with each metal layer in order to additionally take advantage of the capacitance between the different layers. As a result, C_{ac} is the sum of individual capacitances between fingers of the same layer ($C_{m_i - m_i}$) and between segments of each layer and the layers above and below ($C_{m_i - m_j}$). Futhermore, to reduce the parasitic capacitances to ground, the MOM structure is formed in the shape of an inverse pyramid with increased number of fingers (area) for each metal layer from bottom to top. This layout results, in the case of TJ-Monopix1, to a coupling capacitance $C_{\mathrm{ac}} \approxeq 15\,\mathrm{fF}$ and parasitics $C_{p1} \approxeq C_{p2} \approxeq 1.25\,\mathrm{fF}$. Therefore, for $C_D + C_{D_2} = 3.5\,\mathrm{fF}$ and $C_{\mathrm{FE}} = 1\,\mathrm{fF}$, AC-coupling

[2] A high density metal-insulator-metal (MIM) capacitor using a thin high-k dielectric was not available in the selected technology mask set.



(a) (b)

Figure 4.21: High voltage biasing option through the front-side (collection electrode): a) Layout cross-section and b) Schematic of the input coupling and reset circuit.

results to a signal loss of approximately 50% compared to the DC-coupled case (eq. 4.13). An effort to further optimize the MOM capacitor density for TJ-Monopix2 is analyzed in section 4.5.1.1.

Circuit design and analysis The front-end pre-amplifier schematic is shown in Fig. 4.22. M_1 is the voltage amplifier input transistor that amplifies the voltage at node GN (v_{GN}) with a small signal gain equal to:

$$A_{v_{\text{GN}}} = g_{m_1} Z_{\text{OUT}} \tag{4.14}$$

where Z_{OUT} is the total impedance at the output node. The LF-feedback is implemented by M_2, which acts as a common-gate amplifier that amplifies the "error" between V_{OUT} and V_{CASN} and slowly adjusts V_{GN} in order to restore the output baseline. To achieve a low feedback bandwidth, M_2 operates in deep-subthreshold and is biased by a small current I_{THR} that controls the output reset rate. The output baseline voltage is controlled by V_{CASN} and can be approximated using the sub-threshold current equation (4.7):

$$V_{\text{OUT}_{\text{BL}}} = V_{\text{CASN}} - V_{\text{gs}}^{M_2} \approxeq V_{\text{CASN}} - ln\left(\frac{I_{\text{THR}}}{I_0}\left(\frac{L}{W}\right)_{M_2}\right) mV_T - V_{\text{TH}}^{M_2} \tag{4.15}$$

M_4 is the pre-amplifier input transistor which is used as source-follower, biased by the current I_{BIAS}, that buffers the input signal and isolates the sensitive input node from the gain stage (node GN). The source follower output resistance is typically low and depends on the transistor transconductance, $r_{o_{\text{SF}}} \approxeq 1/g_{m_4}$. Its voltage gain is given by:

$$A_{v_{\text{SF}}} = \frac{g_{m_4}}{g_{m_4} + g_{s_4} + \frac{1}{r_{\text{ds}_4}} + \frac{1}{r_{I_1}}} \approxeq \frac{g_{m_4}}{g_{m_4} + g_{s_4}} \tag{4.16}$$

where g_s is the body-effect parameter and r_{I_1} is the output resistance of the current source I_1. In order to eliminate the body-effect ($g_{s_4} = 0$) and achieve a gain close to unity, M_4 is placed in a separate n-well, connected to its source (node SF). To allow V_{GN} to be controlled at low frequencies by the feedback mechanism, the buffered input signal v_{SF} is AC-coupled to node GN by the capacitor C_c. M_3 is a cascode transistor, which essentially works as an impedance converter and serves two purposes:

First, it is used to increase the output impedance of M_4 by $g_{m_3} r_{\mathrm{ds}_3}$ (where r_{ds} is the transistor output resistance) and therefore increase Z_{OUT}. Second, it isolates the drain of M_4 from V_{OUT} due to the low input impedance at its source, which is approximately equal to:

$$R_{\mathrm{in}_3} = \frac{1}{g_{m_3}} \left(1 + \frac{r_{\mathrm{ds}_1} /\!/ R_{\mathrm{in}_2}}{r_{\mathrm{ds}_3}} \right) \approxeq \frac{1}{g_{m_3}} \tag{4.17}$$

where R_{in_2} is the input impedance at the source of M_2. Therefore, the voltage at the drain of M_4 does not fluctuate significantly despite the high amplitude of v_{OUT}. The introduction of M_3 is very important because otherwise the gate-to-drain capacitance of M_4, C_{gd_4}, would act as a miller capacitor between the input and the output of the pre-amplifier with an effective value of $C_{\mathrm{gd}_4} \left(|A_v| + 1 \right)$, where $A_v = -v_{\mathrm{OUT}}/v_{\mathrm{IN}}$ is the pre-amplifier voltage gain. Apart from increasing the input capacitance, the output voltage would couple to the input and reduce the input signal amplitude or even invert its polarity hampering the circuit functionality. Because the gate-to-source capacitance C_{gs_4} can be neglected due to the source-follower action ($v_{\mathrm{IN}} \approxeq v_{\mathrm{SF}}$), the pre-amplifier contributes to the input capacitance only $C_{\mathrm{gd}_4} \approxeq 0.5\,\mathrm{fF}$.

The gate voltage of M_1 (node GN) is adjusted by the LF-feedback so that it sinks a DC current equal to $I_D^{M_1} = I_{\mathrm{BIAS}} + I_{\mathrm{THR}}$. Because the same main branch current I_{BIAS} is used to bias both the source follower (M_4) and amplifier (M_1) transistors, power is utilized very efficiently. When charge is collected, the input voltage "step" $v_{\mathrm{IN}} = Q_s / C_{\mathrm{in}}$ is buffered by M_4 and is transferred to the gate of M_1 through the high-pass filter formed by C_c and the input resistance at node GN. As a result, the gate-to-source voltage of M_1 ($V_{\mathrm{gs}}^{M_1}$) decreases and the drain current of M_1 becomes less than $I_{\mathrm{BIAS}} + I_{\mathrm{THR}}$. Therefore, the output voltage V_{OUT} increases as the capacitance at the output node C_{OUT} starts to charge up due to the current difference $I_D^{M_1} - \left(I_{\mathrm{BIAS}} + I_{\mathrm{THR}} \right)$. At the same time, as V_{OUT} rises, $V_{\mathrm{gs}}^{M_2}$ is decreased resulting in a net positive current $I_{\mathrm{THR}} - I_D^{M_1}$ that starts to charge the capacitance at node GN (C_{GN}) back up and slowly increase $V_{\mathrm{gs}}^{M_1}$ and therefore $I_D^{M_1}$ until the output is reset and the baseline is restored.

Due to the large input signal amplitude, the pre-amplifier charge to voltage conversion gain $G_e = v_{\mathrm{OUT}}/Q_s \approxeq 0.4 - 0.5\,\mathrm{mV}/e^-$ is high and the output voltage amplitude v_{OUT} quickly becomes very large. Therefore, the large signal behavior of the pre-amplifier circuit is inherently non-linear. Nevertheless, it is useful to perform a linearized small signal analysis that is valid for small signal amplitudes in order to gain more insight into the circuit operation and calculate its transfer function. The pre-amplifier small signal model is illustrated in Fig. 4.23. To simplify the analysis, the source-follower is considered to behave like an almost ideal voltage source with unity gain and very low output resistance. Before proceeding, it is necessary to calculate of the total resistance (R_{GN}, R_{OUT}) and capacitance (C_{GN}, C_{OUT}) at nodes GN and OUT that determine the pre-amplifier frequency behavior. R_{GN} is equal to the parallel combination of the output resistance of the current source I_2 (r_{I_2}) and the total output resistance $r_o^{M_2}$ at the drain of the common-gate transistor M_2:

$$R_{\mathrm{GN}} = r_o^{M_2} /\!/ r_{I_2} = \left(g_{m_2} r_{\mathrm{ds}_2} \left(r_{\mathrm{ds}_1} /\!/ r_o^{M_3} \right) \right) /\!/ r_{I_2} \approxeq \left(g_{m_2} r_{\mathrm{ds}_2} r_{\mathrm{ds}_1} \right) /\!/ r_{I_2} \tag{4.18}$$

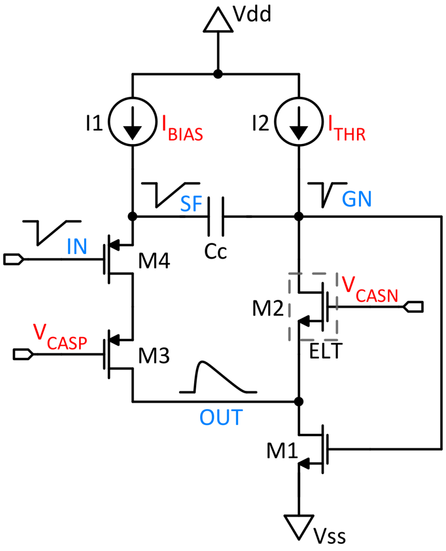


Figure 4.22: Schematic of the TJ-Monopix front-end pre-amplifier circuit.

where $r_o^{M_3}$ is the high total output resistance at the drain of M_3 due to the cascode topology:

$$r_o^{M_3} \approxeq g_{m_3} r_{\mathrm{ds}_3} r_{\mathrm{ds}_4} \tag{4.19}$$

R_{OUT} is equal to the parallel combination of the input resistance at the source of the common-gate M_2 ($r_i^{M_2}$), the output resistance of transistor M_1 (r_{ds_1}) and $r_o^{M_3}$:

$$R_{\mathrm{OUT}} = r_i^{M_2} /\!/ r_o^{M_3} /\!/ r_{\mathrm{ds}_1} = \left(\frac{1}{g_{m_2}}\left(1 + \frac{r_{I_2}}{r_{\mathrm{ds}_2}}\right)\right) /\!/ r_o^{M_3} /\!/ r_{\mathrm{ds}_1} \approxeq \frac{2}{g_{m_2}} /\!/ r_{\mathrm{ds}_1} \tag{4.20}$$

The total capacitance at node GN (expect C_c) is equal to:

$$C_{\mathrm{GN}} \approxeq c_{\mathrm{gd}_{I2}} + \left(1 + A_{v_{\mathrm{GN}}}\right)\left(c_{\mathrm{gd}_1}\right) + c_{\mathrm{gs}_1} + c_{\mathrm{gd}_2} \approxeq A_{v_{\mathrm{GN}}} c_{\mathrm{gd}_1} \tag{4.21}$$

The gate-to-drain capacitance of M_1 is increased by the voltage gain $A_{v_{\mathrm{GN}}} = -v_{\mathrm{OUT}}/v_{\mathrm{GN}}$ due to the miller effect. c_{gs} is the transistor gate-to-source capacitance and $c_{\mathrm{gd}_{I2}}$ is the gate-to-drain capacitance of the current source I_2 PMOS device. Finally, the total capacitance at the output node given by:

$$C_{\mathrm{OUT}} \approxeq c_{\mathrm{gd}_3} + c_{\mathrm{gd}_1} + c_{\mathrm{gg}}^{\mathrm{disc}} \tag{4.22}$$

where $c_{\mathrm{gg}}^{\mathrm{disc}}$ is the gate capacitance of the discriminator input transistor.

The pre-amplifier voltage gain $A_v = v_{\mathrm{OUT}}/v_{\mathrm{IN}}$ can be approximated by the gain at frequencies higher than the LF-feedback, which is equal to the open-loop transfer function $A_{v_{\mathrm{ol}}}(s)$ (the connection between

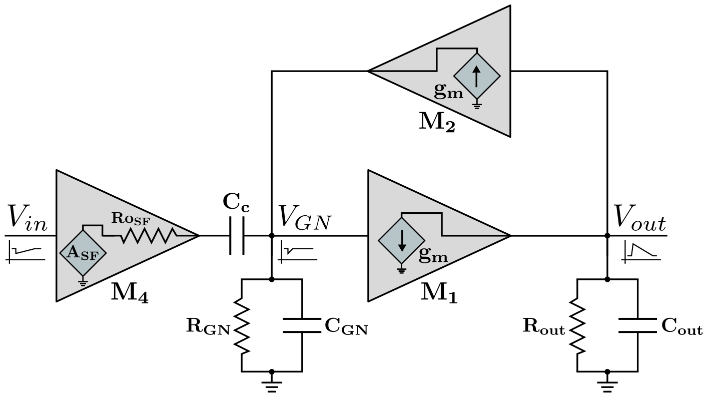


Figure 4.23: Small signal model of the TJ-Monopix front-end pre-amplifier.

M_2 and GN is opened):

$$A_{v_{\text{ol}}}(s) = \left(\frac{v_{\text{GN}}}{v_{\text{IN}}}\right) A_{v_{\text{GN}}} \approxeq \left(\frac{sR_{\text{GN}}C_c}{sR_{\text{GN}}C_c + 1}\right)\left(-\frac{g_{m_1}R_{\text{OUT}}}{sR_{\text{OUT}}C_{\text{OUT}} + 1}\right) \tag{4.23}$$

The first term represents a high pass filter (due to the introduction of C_c) with time constant $\tau_{\text{HP}} = C_c R_{\text{GN}}$. C_{GN} is neglected because it is significantly smaller than C_c. The second term represents a voltage amplifier with gain $g_{m_1} R_{\text{OUT}}$ and gain-bandwidth product $GBW = g_{m_1}/C_{\text{OUT}}$. Therefore, the output voltage is filtered at high frequencies by a low-pass filter with time constant $\tau_{\text{LP}} = C_{\text{OUT}} R_{\text{OUT}}$. Due to the combined high-pass and low-pass terms, the pre-amplifier has built-in pulse shaping (band-pass response) which is essential in order to filter out noise and improve the SNR.

Depending on the timing response requirements, the biasing current I_{BIAS} has to be set accordingly. Increasing I_{BIAS} results in higher g_{m_1} and therefore a higher GBW product. At the same time, since the transistor output resistance is inversely proportional to its drain current ($r_{\text{ds}} \approx 1/LI_D$), R_{OUT} will decrease while the gain will remain approximately constant. Therefore, the output time constant τ_{LP} will be shorter resulting in a higher rise time. Additionally, C_{OUT} has to be kept as low as possible by optimizing the transistor dimensions and minimizing the layout parasitics. The value of I_{BIAS} is also crucial for the performance of the input source follower. The transconductance of the input transistor M_4 (g_{m_4}) has to be high enough to achieve a gain close to unity and low output resistance in order to drive the capacitive load at node GN.

The high-pass cutoff frequency ($f_c \approx 1/\tau_{\text{HP}}$) has to be low to avoid signal loss by excessive filtering within the input signal power spectrum. Therefore, the coupling capacitor C_c has to be large (in the order of 50 fF or higher) and occupies a significant percentage of the total front-end area. Overall, the open-loop system has a low frequency pole at node GN equal to $p_{\text{GN}} = 1/C_c R_{\text{GN}}$ and a high frequency pole at the output equal to $p_{\text{OUT}} = 1/C_{\text{OUT}} R_{\text{OUT}}$.

A low frequency pole at node GN is also necessary to slow down the feedback action. Because the feedback current $i_f = v_{\text{OUT}} g_{m_2}$ charges both C_{GN} and C_c, the feedback time constant is equal to $\tau_f = \left(C_{\text{GN}} + C_c\right) R_{\text{GN}} \approxeq C_c R_{\text{GN}} = 1/p_{\text{GN}}$ and the feedback ratio is given by:

$$\beta = g_{m_2}\left(R_{\text{GN}} /\!/ C_c\right) = \frac{g_{m_2} R_{\text{GN}}}{sR_{\text{GN}}C_c + 1} \tag{4.24}$$

The loop gain is equal to:

$$A_{v_{\mathrm{GN}}}\beta = -\frac{g_{m_1}R_{\mathrm{OUT}}}{sR_{\mathrm{OUT}}C_{\mathrm{OUT}}+1}\frac{g_{m_2}R_{\mathrm{GN}}}{sR_{\mathrm{GN}}C_c+1} \tag{4.25}$$

To calculate the closed-loop transfer function, the superposition principle is used by summing the voltage at node GN for a) $g_{m_2} = 0$ (open loop) and b) $V_{\mathrm{IN}} = 0$ (closed loop with shorted input):

$$\begin{aligned}
&v_{\mathrm{GN}} = v_{\mathrm{IN}}\left(\frac{sR_{\mathrm{GN}}C_c}{sR_{\mathrm{GN}}C_c+1}\right) + v_{\mathrm{OUT}}\beta \implies \\
&A_v(s) = \frac{v_{\mathrm{OUT}}}{v_{\mathrm{IN}}} = -\frac{sg_{m_1}R_{\mathrm{OUT}}R_{\mathrm{GN}}C_C}{(sR_{\mathrm{OUT}}C_{\mathrm{OUT}}+1)(sR_{\mathrm{GN}}C_c+1)+g_{m_1}g_{m_2}R_{\mathrm{OUT}}R_{\mathrm{GN}}} \implies \\
&A_v(s) = -\frac{sg_{m_1}R_{\mathrm{OUT}}R_{\mathrm{GN}}C_C}{s^2R_{\mathrm{OUT}}C_{\mathrm{OUT}}R_{\mathrm{GN}}C_c+s(R_{\mathrm{OUT}}C_{\mathrm{OUT}}+R_{\mathrm{GN}}C_c)+g_{m_1}g_{m_2}R_{\mathrm{OUT}}R_{\mathrm{GN}}+1}
\end{aligned} \tag{4.26}$$

The closed-loop pre-amplifier circuit is also a second-order system. Compared to the open-loop transfer function, the term $g_{m_1}g_{m_2}R_{\mathrm{OUT}}R_{\mathrm{GN}}$ has been added at the denominator which shifts the position of the poles as the feedback strength ($\approx g_{m_2}$) is increased according to the root locus plot shown if Fig. 4.24. The feedback strength and speed can be controlled by the current I_{THR}. A higher I_{THR} value, apart from increasing g_{m_2}, results in a shorter time constant at node GN (τ_{HP}) that shifts the high-pass filter cutoff frequency to higher values. As I_{THR} increases, the feedback action becomes stronger and the high-pass filter bandwidth becomes lower which results to lower gain $|A_v|$ and faster return to baseline. The closed loop poles p_1 and p_2 are initially (for low I_{THR}) close to the open loop poles p_{GN} and p_{OUT} which are real and well separated resulting in an over-damped behavior. As I_{THR} becomes higher, the poles become less separated (critically damped response) and for even higher I_{THR} values they constitute a complex conjugate pair. In this case, the system can be under-damped which results in undershoot of the output voltage as it returns to baseline. To ensure circuit stability and increase the phase margin and damping ratio in order to avoid this effect, the capacitance C_c has to be sufficiently high.

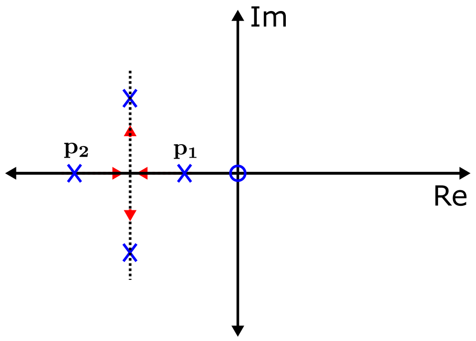


Figure 4.24: Pole shift in the complex s-place as the feedback strength (g_{m_2}) is increased (root locus).

As mentioned, the above analysis is valid for relatively small signal amplitudes. However, if the output voltage amplitude v_{OUT} becomes high, the operating region of transistors M_2, M_3, M_4 and the current source I_1 PMOS device can change due to the reduced voltage headroom, and the circuit behavior becomes significantly non-linear. From a large signal perspective V_{IN}, V_{SF} and V_{GN} remain approximately constant (compared to V_{OUT}). As V_{OUT} approaches V_{CASN}, the drain current of M_2 becomes smaller as it enters the cutoff region and for $V_{\mathrm{OUT}} \gtrsim V_{\mathrm{CASN}}$ it completely stops conducting current. The feedback loop is essentially broken and since $I_D^{M_2} \approxeq 0$ the entirety of I_{THR} charges the total capacitance at node GN (dominated by C_c) with a time constant approximately equal to $\tau \approxeq I_{\mathrm{THR}}/C_c$, similar to a constant current reset scheme. When V_{OUT} drops below V_{CASN}, the feedback loop is closed again and adjusts V_{GN} until the baseline is restored.

As V_{OUT} becomes even larger, transistors M_3, M_4 and the I_1 PMOS are pushed out of saturation and enter the triode region acting as voltage controlled ohmic resistances. Since the total output resistance (R_{OUT}) drops significantly, the output voltage gain is decreased. As a result, V_{OUT} is compressed and is not able to further increase. Therefore, the pre-amplifier voltage gain $|A_v|$ is non linear and quickly drops above a certain input charge. V_{OUT} will remain at the highest possible amplitude and will not start to decrease until C_c is charged by I_{THR}. Hence, although the voltage gain is non-linear, the ToT is not compressed and its linearity is maintained, even for large input charges.

The complete front-end circuit The complete front-end circuit, that includes the pre-amplifier and discriminator is shown in Fig. 4.25. The pre-amplifier biasing current sources I_1 and I_2 are realized by PMOS transistors M_5 and M_7. The capacitance C_c is implemented by a PMOS device connected as a MOSFET capacitor (MOSCAP) that takes advantage of the high capacitance density due to the thin gate oxide ($C_{\mathrm{ox}} = \epsilon_{\mathrm{ox}}/t_{\mathrm{ox}}$). The main branch transistors M_1, M_3 and M_4 operate in moderate inversion ($V_{\mathrm{gs}} \approxeq V_{\mathrm{TH}}$) that offers a good compromise between transconductance efficiency (g_m/I_D) and speed (transit frequency f_T). The input DC potential ($\approxeq$ 1 V) is selected to maximize the output voltage swing by forcing M_5 to operate close to saturation ($V_{\mathrm{ds}} \approxeq V_{\mathrm{dsat}}$). Likewise, the potential at the gate of M_3 ($V_G^{M_3}$) must be set such that M_4 operates close to saturation. However, to optimize routing, $V_G^{M_3}$ is not connected to a separate potential, but is attached to node GN that has a compatible DC level.

The discriminator consists of a simple common-source amplifier stage that operates as a current comparator. The discriminator output (OUTC) state depends on the difference between the drain current of M_{11} and M_{13} ($I_{\mathrm{OUTC}} = I_D^{M_{13}} - I_D^{M_{11}}$). Due to the high impedance at node OUTC, the output voltage V_{OUTC} is driven close to the supply rails (V_{dd} or V_{ss}) for even a small net current I_{OUTC} and hence behaves like a digital signal. The steep slope of the transition between the two states can be observed by the voltage transfer function (large-signal) of the discriminator output (V_{OUTC}) with respect to its input (V_{OUTA}), shown in Fig. 4.26. M_{11} is the discriminator input transistor and its drain current is directly defined by the pre-amplifier output voltage since $V_{\mathrm{gs}}^{M_{11}} \approxeq V_{\mathrm{OUTA}}$. The baseline voltage $V_{\mathrm{OUTA_{BL}}}$ is set below the threshold voltage of M_{11} ($V_{\mathrm{TH}}^{M_{11}}$), but close to the discriminator transition threshold, such that M_{11} only draws a small standby current (typically $<$ 50 nA) that is several times lower than the bias current setting of M_{13}. Therefore, V_{OUTC} is pulled up by M_{13}, which acts as an active load that can be adjusted by the current I_{DISC}. If V_{OUTA} becomes higher than the discriminator transition threshold voltage V_{THR}, M_{11} starts to sink a current higher than I_{DISC} and quickly discharges the total capacitance at node OUTC, pulling down V_{OUTC}. At this state, and while $V_{\mathrm{OUTA}} > V_{\mathrm{THR}}$, M_{11} operates in the triode region with a drain-to-source voltage (V_{ds}) of only a few

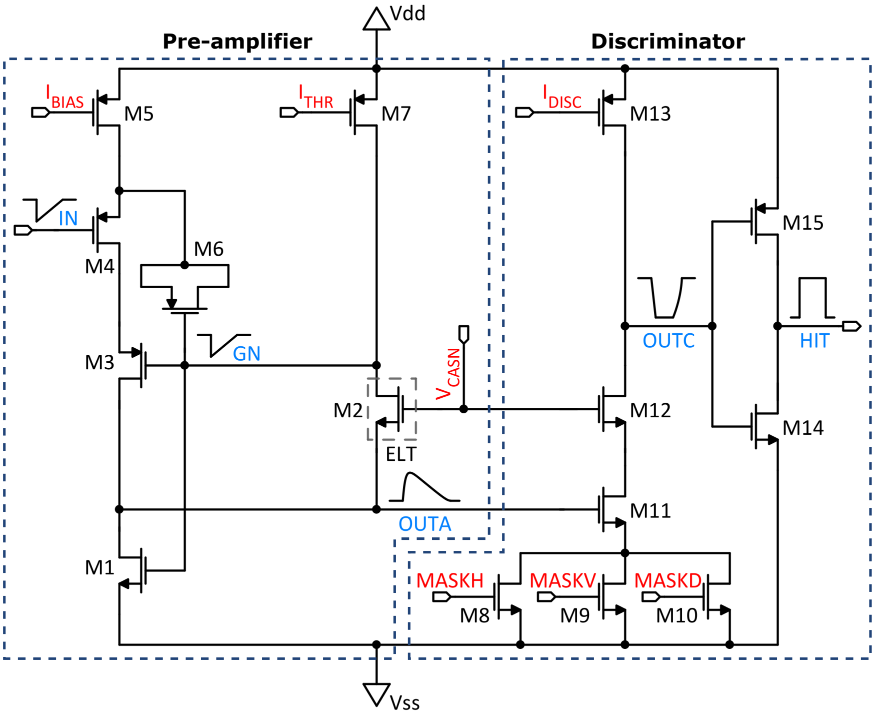


Figure 4.25: Complete schematic of the front-end used by TJ-Monopix1 and includes the pre-amplifier and discriminator stages.

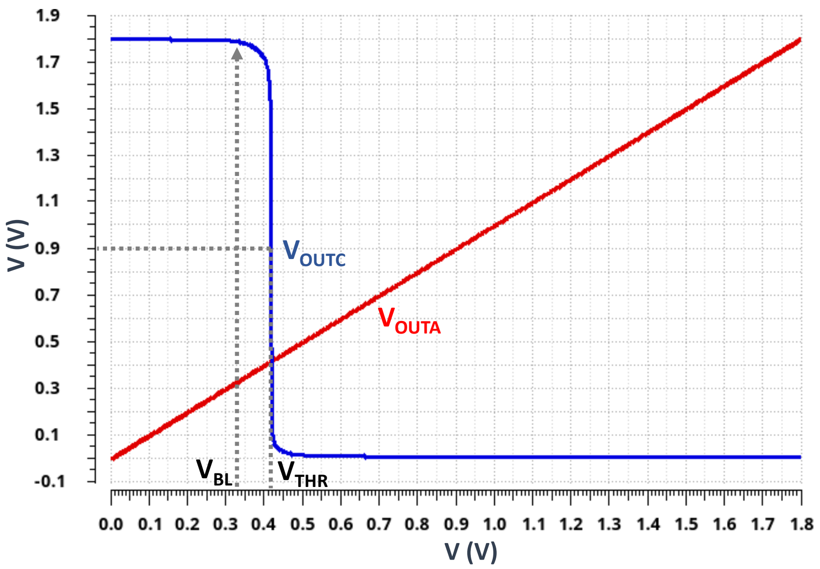


Figure 4.26: Simulated discriminator input-output voltage transfer characteristic.

mV such that $I_D^{M_{11}} = I_D^{M_{13}} \approxeq I_{\text{DISC}}$.

The discriminator transition threshold V_{THR} depends on the geometry (width W and length L) of transistors M_{11} and M_{13} and the discriminator current setting I_{DISC}. In order to flip the discriminator output state and register a hit, a pre-amplifier output voltage (AC) amplitude $v_{\text{OUTA}} > \left(V_{\text{THR}} - V_{\text{OUTA}_{\text{BL}}}\right) = v_{\text{OUTA}_{\text{TH}}}$ is required. Therefore, the threshold amplitude $v_{\text{OUTA}_{\text{TH}}}$ can be controlled by V_{CASN} (that sets the pre-amplifier output baseline) and by I_{DISC}. The threshold input charge Q_{TH} required to induce an amplitude higher than $v_{\text{OUTA}_{\text{TH}}}$ additionally depends on the pre-amplifier gain, and therefore on the feedback current I_{THR} setting.

In order to achieve a faster transition of the discriminator output and improve the timing response, the transconductance of M_{11} ($g_{m_{11}}$) must be high and therefore its width over length ratio (W/L) must be large. Additionally, its gate capacitance must be small in order to increase the pre-amplifier output pole p_{OUTA} frequency and improve the rise time of V_{OUTA}. Therefore, since the gate capacitance scales with the gate area, $C_{\text{gg}} = C_{\text{ox}}WL$, M_{11} is designed with minimum length (0.18 µm). However, because the transistor output resistance is in first order approximation inversely proportional to its length ($r_{\text{ds}} \approx 1/LI_D$), $r_{\text{ds}}^{M_{11}}$ is relatively low. In order to achieve a high discriminator gain and therefore a steep transition slope, transistor M_{12} is introduced to combine the high g_m of M_{11} with a high output impedance. M_{12} is a cascode transistor that works similarly to M_4 in used by the pre-amplifier. The output resistance at the drain of M_{12} is multiplied by $g_{m_{12}}r_{ds_{12}}$ and the discriminator (small-signal) gain is equal to:

$$A_{v_{\text{DISC}}} = g_{m_{11}}r_{\text{OUTC}} = g_{m_{11}}\left(g_{m_{12}}r_{ds_{12}}r_{ds_{11}} /\!/ r_{ds_{13}}\right) \tag{4.27}$$

Apart from increasing the gain, M_{12} is essential to isolate the pre-amplifier output V_{OUTA} from the high amplitude, fast V_{OUTC} signal. Due to the low input impedance at the source of M_{12} the voltage at the drain of M_{11} has a much lower amplitude than V_{OUTC} and coupling to V_{OUTA} is drastically reduced. Furthermore, M_{12} prevents the gate-to-drain capacitance of M_{11} ($C_{\text{gd}}^{M_{11}}$) to be multiplied by the discriminator gain due to the miller effect ($C_{\text{gd}}^{M_{11}}(A_{v_{\text{DISC}}} + 1)$) and significantly increase the total capacitance at the pre-amplifier output (C_{OUT}).

The discriminator output is followed by a digital inverter gate, comprised of transistors M_{14} and M_{15} that buffers the HIT pulse output in order to drive the in-pixel readout logic. Furthermore, the polarity of the HIT pulse is inverted in order to be an active high signal. Transistors M_8, M_9 and M_{10} are used to mask the pixel by disabling the discriminator. They work as switches implementing an "OR" logic i.e. if at least one of the MASKH, MASKV or MASKD signals is high, the discriminator is enabled because there is a current path to ground. The mask configuration signals MASKH, MASKV or MASKD represent the three coordinates (vertical, horizontal, diagonal) of the projection masking scheme that is implemented in TJ-Monopix1 (refer to section 4.3.1.3). The pixel is masked if the masking signals of all coordinates are low.

It should be emphasized that such a simple discriminator circuit is only feasible due to the high charge to voltage conversion gain G_e of the sensor/pre-amplifier and therefore high V_{OUTA} signal amplitude. To calculate the impact of the discriminator noise and transistor parameter mismatch between different pixels, the equivalent RMS noise ($V_{n_{\text{DISC}}}^{\text{rms}}$) and mismatch ($V_{\text{mis}_{\text{DISC}}}^{\text{rms}}$) voltage at the discriminator input (gate of M_{11}) has to be converted to an equivalent input charge. Therefore the

ENC and threshold dispersion ($\sigma_{\mathrm{TH}_{\mathrm{DISC}}}$) due to the discriminator are equal to:

$$\begin{aligned} ENC_{\mathrm{DISC}} &= \frac{V^{\mathrm{rms}}_{n_{\mathrm{DISC}}}}{G_e} \\ \sigma_{\mathrm{TH}_{\mathrm{DISC}}} &= \frac{V^{\mathrm{rms}}_{\mathrm{mis}_{\mathrm{DISC}}}}{G_e} \end{aligned} \tag{4.28}$$

Thus, as the gain is increased (G_e), the discriminator influence in noise and threshold mismatch performance is reduced. Although the discriminator noise is relatively low compared to the pre-amplifier noise and can be neglected, the contribution to threshold dispersion can be significant due to the threshold variation of transistor M_{11}. In contrast to TJ-Monopix1, for applications with a much lower gain more complex discriminator circuits have to be employed, often based in a matched differential pair, to reduce mismatch and noise with the expense of area and power consumption.

The simulated transient response of the complete front-end for two input charges is shown in Fig. 4.27. The charge values of 300 e^- and 1 600 e^- are close to the detection threshold and charge MPV respectively. The voltage Q_s/C_{in} induced at the input by the collected charge is translated to a digital hit pulse with a width (ToT) proportional to Q_s. The sensor leakage current is set to $I_{\mathrm{leak}} = 10\,\mathrm{pA}$, which is an overestimation for an unirradiated sensor. The I_{THR}, V_{CASN} and I_{DISC} parameters have been tuned for a threshold of $\approx 300\,e^-$ and ToT at the MPV approximately equal to 1 µs. Because the phase margin is high ($PM \approxeq 90°$), no undershoot is observed during the return to baseline of the analog output V_{OUTA}.

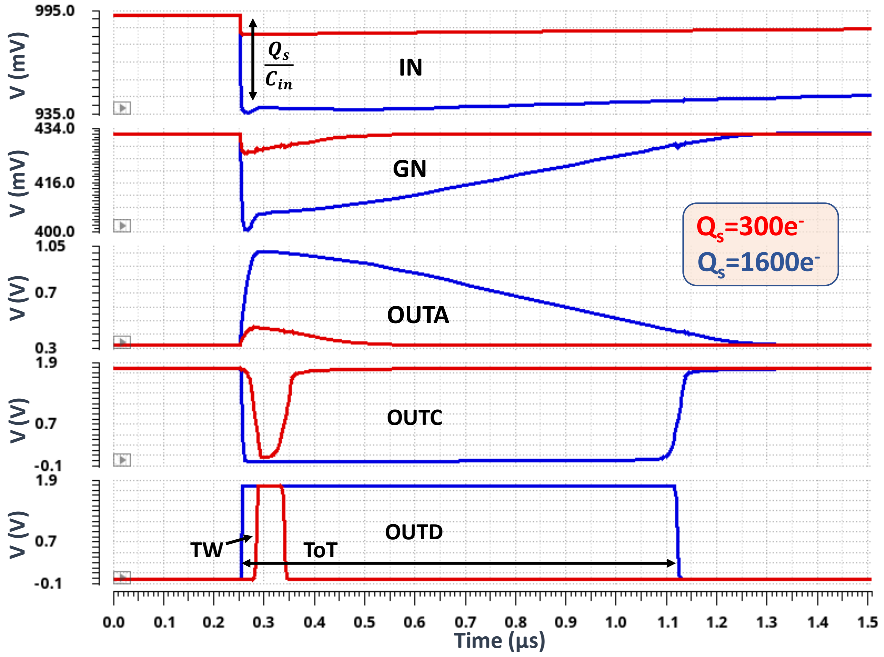


Figure 4.27: Post layout simulation of the TJ-Monopix1 front-end transient response for two input charge values ($Q_s = 300\,e^-, 1\,600\,e^-$).

Pre-amplifier transfer function and gain The simulated pre-amplifier voltage transfer function is shown in Fig. 4.28 for two different feedback bias current (I_{THR}) settings. The zero is not at $f = 0\,\mathrm{Hz}$, as predicted by the transfer function A_v (eq. 4.26), but at a very low frequency $f < 100\,\mathrm{Hz}$ due to the direct path through the high resistance of transistors M_3, and M_4 ($\cong 500\,\mathrm{M\Omega}$) that is not included in the pre-amplifier model. The two poles depend on the value of I_{THR}. For $I_{\mathrm{THR}} = 0.5\,\mathrm{nA}$ they are separated and close to the open loop poles p_{GN} and p_{OUT}. When I_{THR} is increased to 2.5 nA, the poles approach each other and are approximately equal to $p_1 = 6.5\,\mathrm{MHz}$ and $p_2 = 8\,\mathrm{MHz}$. As can be observed, the pre-amplifier bandwidth is decreased, hence the output signal amplitude is reduced. The high frequency pole at $f \cong 200\,\mathrm{MHz}$ appears due to the source-follower (M_4) output resistance. It can be concluded that the theoretical analysis based on the simplified pre-amplifier model accurately describes its behavior.

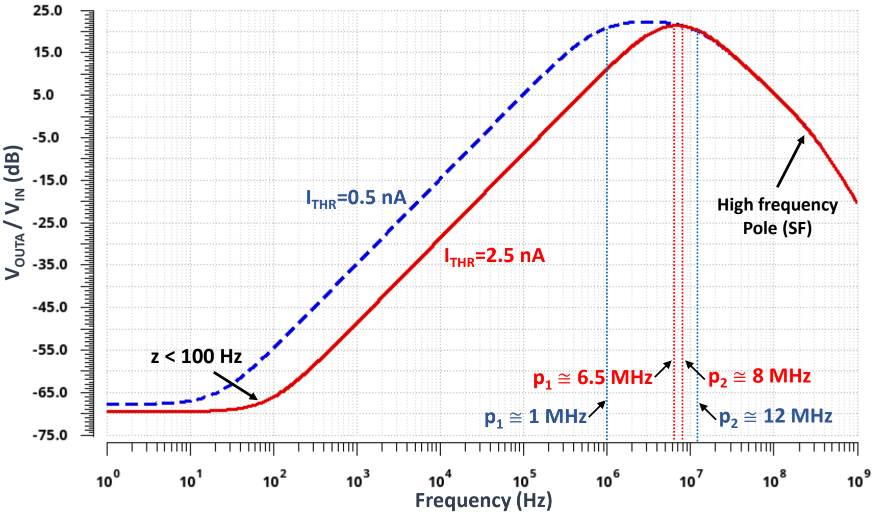


Figure 4.28: Simulated small-signal voltage transfer function $A_v(\omega)$ of the TJ-Monopix1 front-end pre-amplifier for two feedback (I_{THR}) current settings.

The pre-amplifier charge to voltage conversion gain (G_e) has been also simulated and is plotted in Fig. 4.29. Due to the circuit non-linear behavior, as V_{OUTA} increases the transconductance of M_2 decreases until it is completely switched off. Therefore the gain increases until V_{OUTA} is high enough to push transistors M_3 and M_4 out of saturation and then significantly drops since the output swing limits have been reached. The glitch near the threshold ($300\,e^-$) is due to coupling from the discriminator output to V_{OUTA} that is not completely suppressed by M_{12}. The gain value near the threshold is important for the calculation of the equivalent noise charge and threshold dispersion and is approximately equal to $0.4\,\mathrm{mV}/e^-$. It is worth noting that the corresponding voltage gain $G_v = v_{\mathrm{OUT}}^{p-p}/v_{\mathrm{IN}}^{p-p}$ is equal to approximately 12 V/V. Although the pre-amplifier voltage gain is relatively low, G_e is large due to the low sensor capacitance and corresponding high Q_s/C_D ratio at the input.

Timing response and ToT In contrast to a CSA, the timing response of the TJ-Monopix front-end does not depend on the detector capacitance (C_D) because the input is isolated from the pre-amplifier output due to the absence of feedback. Therefore, the timing response depends on the pre-amplifier rise-time and the charge integration time on C_D, which is defined by the charge collection time. It

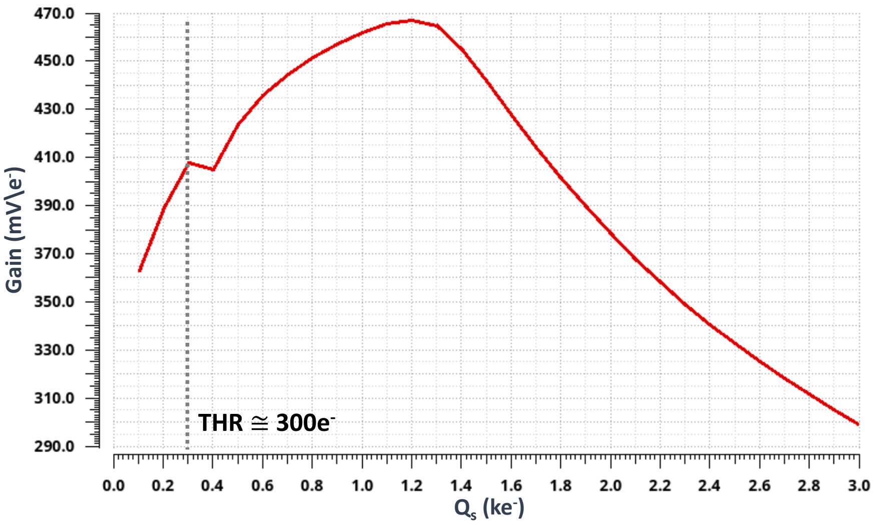


Figure 4.29: Post-layout simulation of the TJ-Monopix1 front-end charge to voltage conversion gain G_e (mV/e^-). The glitch near the threshold is caused by coupling from the discriminator stage. More details are provided in the text.

should be noted that although the detector capacitance does not influence the front-end time-walk, it still affects the detection threshold and as a consequence the in-time threshold. Therefore, apart from a low the pre-amplifier rise-time, the key to high in-time efficiency is to achieve a low operating threshold.

The timing response is one of the most important differences between the ALPIDE and the TJ-Monopix front-end. While the ALPIDE chip has a shaping time of several µs that is adequate for the ALICE experiment, the TJ-Monopix timing response has to be much faster to comply with the ATLAS specifications of 25 ns. As explained, the most important parameter that influences speed i.e. the pre-amplifier output rise time is the main branch bias current I_{BIAS} since it controls the gain-bandwidth product at the output node (g_{m_1}/C_{OUT}). Therefore, the front-end power consumption is constrained by the timing requirements. To achieve a fast timing response, compatible with the ATLAS requirements, I_{BIAS} has to be increased to approximately 500 nA. This modification has significant consequences in several design aspects such as the bandwidth, gain, noise and threshold dispersion resulting in a different front-end behavior. The front-end power consumption is equal to:

$$P_{\mathrm{FE}} \approxeq V_{\mathrm{dd}} \left(I_{\mathrm{BIAS}} + I_{\mathrm{DISC}}\right) = 1.8\,\mathrm{V} \cdot 550\,\mathrm{nA} \approx 1\,\mu\mathrm{W} \tag{4.29}$$

The simulated timing response of the hit pulse leading edge (LE), that describes the front-end time-walk is plotted in Fig. 4.30. The time-walk (TW) can be calculated by subtracting the LE arrival time for high input charge values e.g. $Q_s \geq 3\,000\,e^-$ from the arrival time for $Q_s = Q_{\mathrm{TH}} = 300\,e^-$ and is equal to $TW \approxeq 45$ ns. Due to the steep rise of the LE timing response close to the threshold it is more useful to instead calculate the in-time threshold i.e. the minimum input charge for which the front-end responds within 25 ns. As can be observed, the in-time threshold is only about 340 e^-, which is 40 e^- higher than the detection threshold. This difference is also called the "overdrive" charge and is a good indicator of the front-end timing performance with respect to its threshold. Therefore, even for the worst case of charge being shared between four neighboring pixels, for a deposited charge close to the MPV, all hits will be recorded in-time ($\approx 400\,e^-$ per pixel). In fact, if pixel clusterization is performed during processing, only the cluster's seed pixel (pixel that collects the highest amount of charge) has to be in-time.

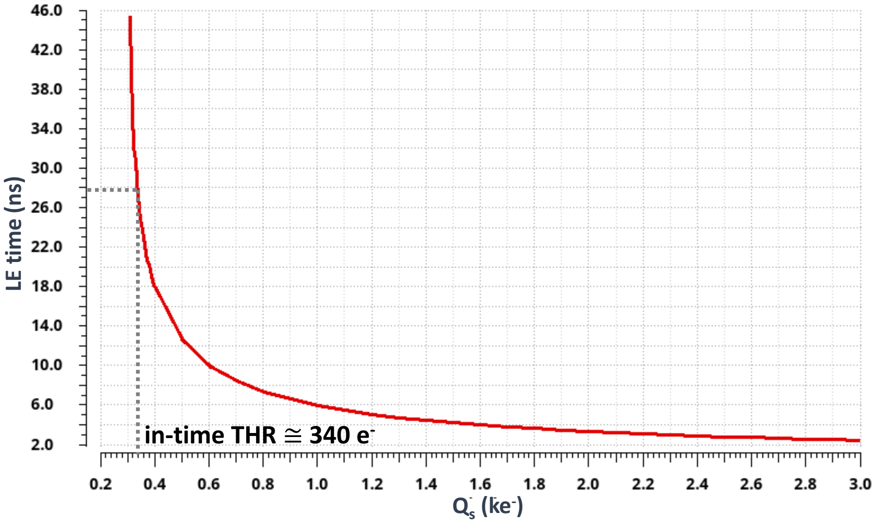


Figure 4.30: Post-layout simulation of the TJ-Monopix1 front-end time-walk curve. The 25 ns in-time threshold is equal to $340\,e^-$ ($40\,e^-$ overdrive charge).

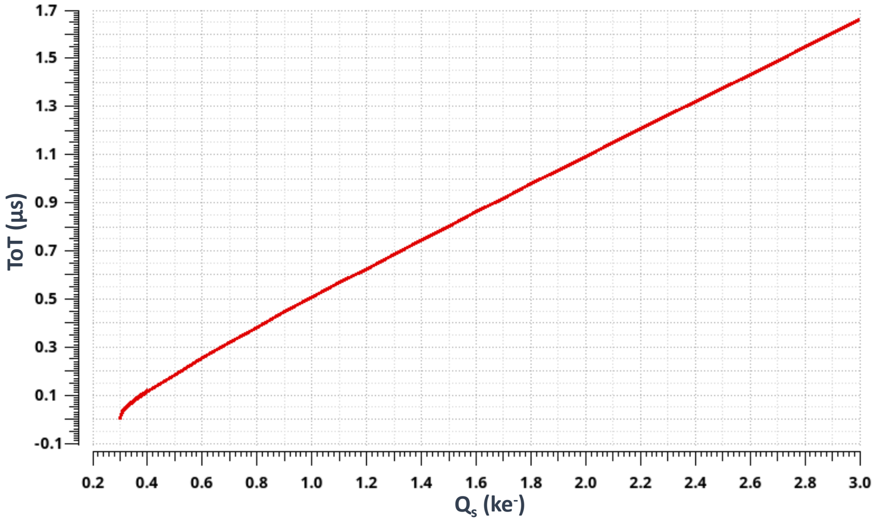


Figure 4.31: Simulated post-layout ToT curve of the TJ-Monopix1 front-end.

The ToT response has been simulated as well and is shown in Fig. 4.31. In contrast to the pre-amplifier output voltage amplitude (v_{OUTA}) and gain, the ToT is highly linear and does not saturate even for high input charge values. As mentioned, this behavior is due to the PMOS input reset scheme and the pre-amplifier output reset mechanism i.e. the time that is required to charge the capacitor C_c by the feedback circuit (through the current I_{THR}) before V_{OUTA} returns to baseline is almost linear with respect to the input charge Q_s. Furthermore, although V_{OUTA} is compressed due to the voltage limit set by the power supply V_{dd} and the transistor drain-to-source saturation voltage (V_{dsat}), there is no such limitation for the ToT.

The ToT duration should normally not be higher than the maximum time allowed by the BCID counter before rollover, which is equal to $2^6 \cdot 25\,\mathrm{ns} = 1.6\,\mu\mathrm{s}$ for 6-bit resolution. Since ToT clipping is not included in TJ-Monopix1, I_{THR} has been selected such that the ToT is approximately equal to 1 µs for the input charge MPV ($1\,600\,e^-$).

Noise performance In order to access the front-end noise performance, the equivalent noise charge (ENC) must be calculated using simulation data. The pre-amplifier noise can be derived by performing an AC (or transient noise) simulation in order to obtain the RMS noise voltage at the output ($\sqrt{\langle v_{\mathrm{OUTA}}^2 \rangle}$) and convert it to ENC by dividing with the gain (G_e) near the threshold (eq. 3.16). The ENC of the full front-end circuit can be determined by the "s-curve" method that makes use of the digital HIT output. For each input charge value, in a range near the threshold, a large number of transient noise simulations are performed and the hit probability ($P(HIT)$) is calculated. The plot of $P(HIT)$ vs Q_s has a sigmoid shape that can be described by the cumulative distribution function (CDF) or the normal distribution (assuming gaussian noise):

$$CDF(Q_s) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{Q_s - \mu}{\sigma\sqrt{2}}\right)\right) \tag{4.30}$$

where erf(x) is the Gauss error function. The threshold Q_{TH} and ENC are obtained by fitting the simulation points with the CDF function and are equal to the mean μ and standard deviation σ respectively. Alternatively, the ENC can be calculated by the fitted s-curve slope at the threshold as described by eq. 3.17.

Before discussing the simulation results, it is useful to analyze the pre-amplifier noise in order to identify the most important contributors and their noise transfer function to the output. The following noise analysis is based on the pre-amplifier small-signal model of Fig. 4.32 that includes the relevant noise sources. The most critical devices, in terms of noise output, are transistors M_1, M_2, M_4 and the input PMOS reset transistor M_{reset}. In contrast, the noise of the current source transistors M_5 and M_7 is smaller because they are designed with a large area WL and low W/L ratio, hence low transconductance $g_m \approx W/L$. Therefore, their thermal and flicker noise current will be low. The noise contribution of M_3 is also small due to the low magnitude of its noise transfer function. Since M_3 is a cascode device, an equivalent voltage noise source at its gate mainly affects the drain-to-source voltage of M_4 ($V_{\mathrm{ds}}^{M_4}$) and does not directly influence the output signal.

At the input the total noise current is the sum of the sensor diode shot noise current $\langle i_{\mathrm{shot}}^2 \rangle$ and the noise current of M_{reset} ($\langle i_{\mathrm{reset}}^2 \rangle$), given by eq. **??**. In the unirradiated case, $\langle i_{\mathrm{shot}}^2 \rangle$ is negligible and can be omitted. The noise current is converted to a noise voltage by the input impedance which is the parallel combination of the total resistance R_{in} and capacitance C_{in} at the input node. R_{in} is the equivalent output resistance at the drain of M_{reset} and depends on the transistor bias I_{RESET}, geometry and the sensor leakage current. C_{in} is dominated by the sensor capacitance C_D. The RMS noise voltage at the input will therefore be:

$$\sqrt{\langle v_{\mathrm{IN}}^2 \rangle (s)} = \sqrt{\langle i_{\mathrm{shot}}^2 + i_{\mathrm{reset}}^2 \rangle (s)}\,\frac{R_{\mathrm{in}}}{sR_{\mathrm{in}}C_{\mathrm{in}} + 1} \tag{4.31}$$

Therefore the noise transfer function of $\langle i^2_{\mathrm{shot}} + i^2_{\mathrm{reset}} \rangle$ to the output will be equal to:

$$\sqrt{\frac{\langle v^2_{\mathrm{OUTA}} \rangle}{\langle i^2_{\mathrm{shot}} + i^2_{\mathrm{reset}} \rangle}(s)} = \frac{R_{\mathrm{in}}}{sR_{\mathrm{in}}C_{\mathrm{in}} + 1} A_v(s) \tag{4.32}$$

where $A_v(s)$ is the pre-amplifier transfer function (eq. 4.26). The input current noise is shaped by the pre-amplifier frequency response with an additional pole at $p_{\mathrm{IN}} = 1/R_{\mathrm{in}}C_{\mathrm{in}}$, as shown in Fig. 4.33. It is important to note that due to the high value of R_{in}, the transfer function gain is large. Thus, although M_{reset} has low transconductance and relatively large area, its noise contribution is still important.

In the case of the input transistor (M_4), it is easier to consider the equivalent noise source at its gate, $\langle v_4^2 \rangle = \langle i_4^2 \rangle / g^2_{m_4}$, that is in series with the input voltage V_{IN}. Therefore the noise transfer function of M_4 is equal to:

$$\sqrt{\frac{\langle v^2_{\mathrm{OUTA}} \rangle}{\langle i_4^2 \rangle}(s)} = \frac{A_v(s)}{g_{m_4}} \tag{4.33}$$

As expected, the noise of M_4 is shaped exactly as the input signal v_{IN}. Because the thermal noise current $\sqrt{\langle i^2_{D_{\mathrm{th}}} \rangle} \approx \sqrt{g_m}$, the output noise due to M_4 scales with $1/\sqrt{g_{m_4}}$, hence g_{m_4} has to be high.

M_1 is also one of the most critical devices (together with M_4), since its equivalent noise voltage is in series with v_{GN} and is directly amplified by $g_{m_1} Z_{\mathrm{OUT}}$. The noise transfer function of M_1 must take into account the effect of the feedback that tries to correct for output voltage fluctuations and maintain the baseline voltage, therefore suppressing noise within its bandwidth:

$$\begin{gathered}
\sqrt{\langle v^2_{\mathrm{OUTA}} \rangle (s)} = \sqrt{\langle i_1^2 \rangle (s)} Z_{\mathrm{OUT}} - \sqrt{\langle v^2_{\mathrm{OUTA}} \rangle (s)} g_{m_2} Z_{\mathrm{GN}} g_{m_1} Z_{\mathrm{OUT}} \implies \\
\sqrt{\frac{\langle v^2_{\mathrm{OUTA}} \rangle}{\langle i_1^2 \rangle}(s)} = \frac{R_{\mathrm{OUT}}\left(sR_{\mathrm{GN}}C_C{+}1\right)}{s^2 R_{\mathrm{OUT}}C_{\mathrm{OUT}}R_{\mathrm{GN}}C_c{+}s\left(R_{\mathrm{OUT}}C_{\mathrm{OUT}}{+}R_{\mathrm{GN}}C_c\right){+}g_{m_1}g_{m_2}R_{\mathrm{OUT}}R_{\mathrm{GN}}{+}1}
\end{gathered} \tag{4.34}$$

Thus, the noise transfer function of M_1 (Fig. 4.33) has the same poles as the pre-amplifier transfer function $A_v(s)$ and a zero at $z_{\mathrm{GN}} = 1/R_{\mathrm{GN}}C_C$. The feedback essentially acts as a high-pass filter with respect to the noise current generated at the output by M_1. At low frequencies, below z_{GN}, the loop gain, and as a result noise suppression by the feedback, is maximum. In this range, noise is suppressed approximately a factor or $1/g_{m_1}g_{m_2}R_{\mathrm{GN}}$. Above z_{GN}, the loop gain starts to decrease and the transfer function magnitude increases with a slope of ≊ 20 dB/*dec*. At the center frequency between p_1 and p_2, the transfer function magnitude becomes maximum since the feedback is not fast enough to respond and for higher frequencies it drops again due to the low-pass effect of C_{OUT}. Similarly to M_4, a high transconductance g_{m_1} improves the signal to noise ratio:

$$\frac{A_v(s)}{\sqrt{\frac{\langle v^2_{\mathrm{OUTA}} \rangle}{\langle i_1^2 \rangle (s)}}\sqrt{\langle i_1^2 \rangle (s)}} = \frac{s g_{m_1} R_{\mathrm{OUT}} R_{\mathrm{GN}} C_C}{R_{\mathrm{OUT}}\left(sR_{\mathrm{GN}}C_C + 1\right)} \frac{1}{\sqrt{\langle i_1^2 \rangle (s)}} \approxeq \frac{g_{m_1}}{\sqrt{\langle i_1^2 \rangle (s)}} \approx \sqrt{g_{m_1}} \tag{4.35}$$

Finally, the noise transfer function of M_2 can be calculated as follows:

$$\begin{gathered}\sqrt{\langle v^2_{\mathrm{OUTA}}\rangle(s)} = \left(\sqrt{\langle v^2_{\mathrm{OUTA}}\rangle(s)}\,g_{m_2} + \sqrt{\langle i^2_2\rangle(s)}\right) Z_{\mathrm{GN}} g_{m_1} Z_{\mathrm{OUTA}} \implies \\ \sqrt{\frac{\langle v^2_{\mathrm{OUTA}}\rangle}{\langle i^2_2\rangle}(s)} = \frac{R_{\mathrm{GN}}R_{\mathrm{OUT}}g_{m_1}}{s^2 R_{\mathrm{OUT}}C_{\mathrm{OUT}}R_{\mathrm{GN}}C_c + s\left(R_{\mathrm{OUT}}C_{\mathrm{OUT}} + R_{\mathrm{GN}}C_c\right) + g_{m_1}g_{m_2}R_{\mathrm{OUT}}R_{\mathrm{GN}} + 1}\end{gathered} \tag{4.36}$$

The noise current of M_2 is directly added at node GN and is low-pass filtered by the two poles (p_1 and p_2) defined by the pre-amplifier transfer function $A_v(s)$. In contrast to the previous cases, no zero exists and therefore the noise transfer function of M_2 (Fig. 4.33) is essentially that of a second-order low-pass filter with low frequency gain equal to $R_{\mathrm{GN}}R_{\mathrm{OUT}}g_{m_1}$.

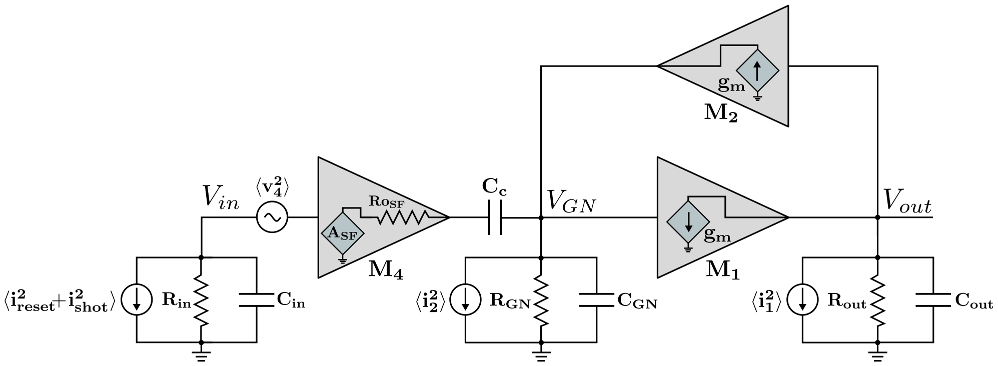


Figure 4.32: Small signal model of the TJ-Monopix front-end pre-amplifier including the main noise sources.

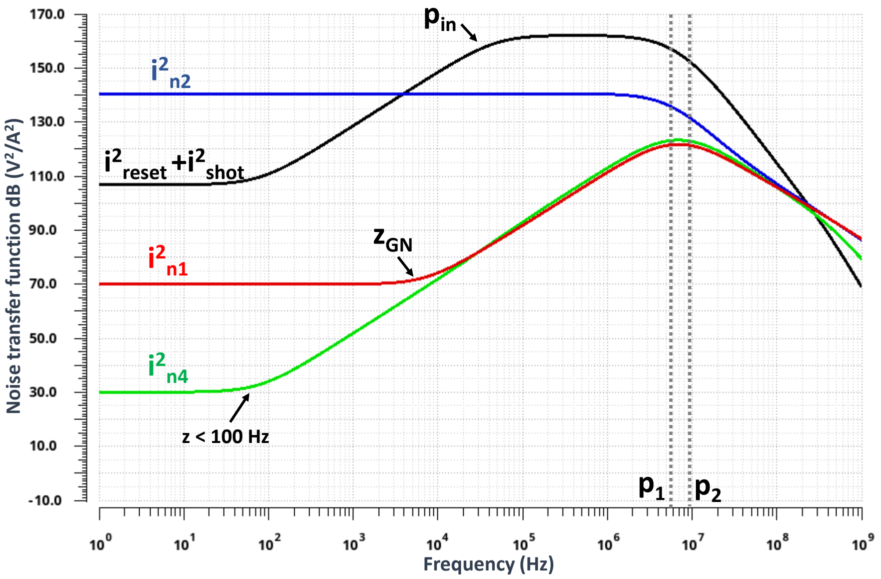


Figure 4.33: Transfer function of the main noise sources ($\langle v^2_{\mathrm{OUTA}}\rangle / \langle i^2_n\rangle$).

The total RMS noise output is equal to the quadratic sum of the power spectral density of each noise current source multiplied by the respective transfer function, integrated within a specified frequency range:

$$\sqrt{\langle v^2_{\text{OUTA}} \rangle} = \sqrt{\sum_{k=1}^{n} \int_{f1}^{f2} \frac{\langle v^2_{\text{OUTA}} \rangle}{\langle i^2_k \rangle}(f)\, \mathrm{d}\langle i^2_k \rangle(f)\; df} \tag{4.37}$$

The simulated output noise voltage has been integrated from 1 Hz to 1 GHz, covering a frequency range much broader that the pre-amplifier pass-band, and has been found equal to ≊ 4 mV RMS. Using this value the pre-amplifier ENC can be calculated as follows:

$$ENC_{\text{OUTA}} = \frac{\sqrt{\langle v^2_{\text{OUTA}} \rangle}}{G_e} \approxeq \frac{4\,\text{mV}}{0.4\,\text{mV}/e^-} = 10\,e^- \tag{4.38}$$

The excellent noise performance of TJ-Monopix, demonstrates the high SNR due to the low sensor capacitance and high Q_s/C_D ratio, combined with an optimized front-end. The five highest noise sources, depending on the integrated noise voltage they contribute to the output, are presented in Table 4.3. In this analysis the thermal and flicker noise parts are considered separately. As expected, all critical devices (M_1,M_2,M_4 and M_{reset}) appear in the list. Their combined noise is approximately 95% of the total which demonstrates the validity of the above theoretical considerations.

Transistors M_1 and M_4 contribute the highest amount of noise, approximately 38% each. Furthermore, all sources are thermal except the flicker noise of M_1 that makes up about 8% of the total noise. S_{fl} of M_1 is comparatively high because the area of M_1 is relatively small (0.18 µm^2) and the finite gain $g_{m_1} g_{m_2} R_{\text{GN}}$ limits noise suppression at low frequencies. In contrast, devices M_2 and M_{reset} have a significantly larger area (≊ 1 µm^2) than M_1, while the noise transfer function magnitude of M_4 is lower at low frequencies due to the high-pass effect of C_c.

A very important noise parameter that if not considered can lead to drastically reduced performance is RTS noise. The unique characteristic of RTS noise is that it is (usually) not included in the models provided by the foundry and therefore it is difficult to estimate its impact during the design phase. Typically, the development procedure is iterative and involves prior fabrication of test structures, in order to access the validity of simulations and the impact of factors such as the RTS noise. Because the drain current fluctuation due to charge trapped or released at the gate interface (Q_T) is proportional to $Q_T/C_{\text{gg}} \approx Q_T/WL$, devices that are particularly affected by RTS noise must be enlarged in order to increase their gate area WL[3].

In the case of the TJ-Monopix front-end, the most critical devices in terms of RTS noise are transistors M_4 and M_1. The gate area of M_4 cannot be very high because it will result in an increased input capacitance C_{in} and therefore reduced signal amplitude. Experience from the ALPIDE front-end has shown that the area of M_4 should not be lower that approximately 0.15 µm^2 because the impact of RTS noise will become significant. On the other hand, the impact of M_1, which is important due to the relatively high noise transfer function magnitude at low frequencies, was underestimated in the design of TJ-Monopix1 as will be observed by the measured ENC distribution long-tail (section 4.4.3). M_1 has been enlarged and correctly sized in the TJ-Monopix2 design significantly improving performance as will be discussed in section 4.5.1.1.

[3] While the Poisson variance of the trap-release process is given by the number of traps that is proportional to the area, the resulting squared voltage fluctuations are proportional to $1/(WL)^2$ resulting in a inverse area dependence.

Table 4.3: The five highest noise sources of the TJ-Monopix1 front-end.

Noise source	Integrated noise (μV^2)	Contribution (%)
M_4 thermal noise (S_{th})	5.5	38%
M_1 thermal noise (S_{th})	4.4	30%
M_{reset} thermal noise (S_{th})	2.2	15%
M_1 flicker noise (S_{fl})	1.15	8%
M_2 thermal noise (S_{th})	0.5	3.6%

The full front-end noise performance, simulated using the s-curve method is shown in Fig. 4.34. Each point represents the hit probability calculated from 100 iterations. The CDF curve fit yields an ENC value equal to approximately $9\,e^{-}$, that is close to the value calculated using the RMS pre-amplifier output noise voltage.

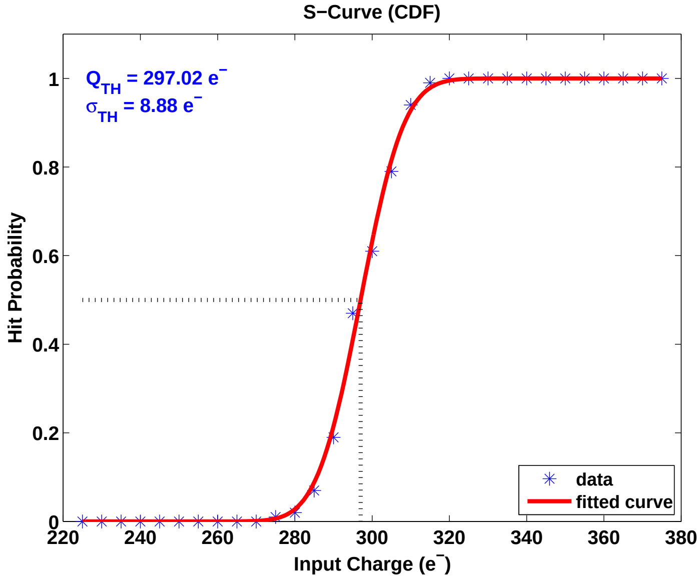


Figure 4.34: Simulated post-layout cumulative distribution function (s-curve) with transient noise applied. The standard deviation of the CDF function fit yields an ENC of ≊ $8.9\,e^{-}$.

Threshold dispersion The threshold dispersion (σ_{TH}) due to mismatch between the transistor parameters of different pixels is crucial because it is essentially added (quadratically) to the ENC and limits the minimum operating threshold (eq. 3.19). The sensitivity of σ_{TH} to mismatch of each transistor parameter is different and depends on the specific device and its role on the circuit topology. Furthermore, the standard deviation of the transistor parameter mismatch is inversely proportional to the square root of its area:

$$\sigma_i = \frac{\sigma_{i_0}}{\sqrt{A_i}} \tag{4.39}$$

where σ_{i_0} is the normalized RMS mismatch value for an area of 1 µm^2 and $A_i \approxeq WL$ is the transistor area. The total RMS threshold dispersion is equal to:

$$\sigma_{\mathrm{TH}} = \sqrt{\sum_{i=0}^{k} \left(\frac{\mathrm{d}\sigma_{\mathrm{TH}_i}}{\mathrm{d}\sigma_i} \right) \frac{\sigma_{i_0}}{\sqrt{A_i}}} \tag{4.40}$$

where σ_i is the transistor parameter mismatch and $\mathrm{d}\sigma_{\mathrm{TH}_i}/\mathrm{d}\sigma_i$ the corresponding sensitivity. It is therefore necessary to identify the critical devices with the highest sensitivity and increase their area in order to reduce mismatch. However, in many cases this requirement contradicts other design aspects and performance metrics that are affected by the increased gate capacitance and an optimum value has to be selected. Furthermore, the total transistor area is constrained by the available pixel area allocated for the front-end circuit.

The influence of transistor mismatch on threshold dispersion is evaluated by monte-carlo simulations that randomly generate a different parameter set for each iteration based on the mismatch distribution included in the foundry model. The top five transistors with the highest contribution to threshold dispersion due to mismatch are presented in Table 4.4. The detection threshold dispersion sensitivity is high for transistors M_{11} and M_2 because the variation of their threshold voltage directly affects the minimum pre-amplifier output voltage amplitude that is detected by the discriminator. Although the influence of M_2 is reduced due to its larger area, M_{11} has to remain relatively small to prevent increasing the output capacitance C_{OUT} that has a negative impact on timing performance. Therefore, M_{11} is the most critical device in terms of mismatch. Its threshold voltage has a standard deviation in the order of $\sigma_{V_{\mathrm{th}}} \approxeq 6.5\,\mathrm{mV}$ yielding a contribution to the total detection threshold variance of approximately 42%. In contrast to M_{11}, a variation in the threshold voltage of M_1 is not as important because it is automatically compensated by the LF-feedback that adjusts the DC level of V_{GN}. In the case of the current source transistors M_5, M_7 and M_{13}, even though the mismatch sensitivity is relatively high, they are designed with a large area in order to reduce their influence to threshold dispersion.

Table 4.4: The five highest threshold dispersion contributors of the TJ-Monopix1 front-end.

Device	Area (µm^2)	Variance contribution (%)
M_{11}	0.36	42%
M_2	1	11%
M_1	0.18	3%
M_{15}	0.075	2%
M_5	4	2%

The total threshold dispersion is evaluated in Fig. 4.35 by employing the s-curve method, as in the case of the ENC simulation. Each point represents the hit probability calculated from 100 iterations with different parameter sets. The RMS threshold dispersion (σ_{TH}) is equal to $\approxeq 25\,e^-$ and is about 2.5 times larger than the ENC. Assuming that the operating threshold has to be higher than 10σ of the combined ENC and threshold dispersion to achieve a very low noise rate, the minimum threshold will be approximately equal to:

$$Q_{\mathrm{TH_{min}}} \approx 10\sqrt{\sigma_{\mathrm{TH}}^2 + \mathrm{ENC}^2} \approxeq 270\,e^- \tag{4.41}$$

Although the minimum threshold is dominantly limited by threshold dispersion instead of noise, a value of approximately $300\,e^-$ was considered sufficient compared to the signal MPV of around $1\,600\,e^-$. Therefore, as mentioned, no in-pixel tuning has been included in TJ-Monopix1 in order to simplify the pixel design. However, it will become evident by measurement results (section 4.4.5) that in order to achieve full efficiency after irradiation, even lower threshold values are necessary which can be accomplished by the improved front-end incorporated in TJ-Monopix2.

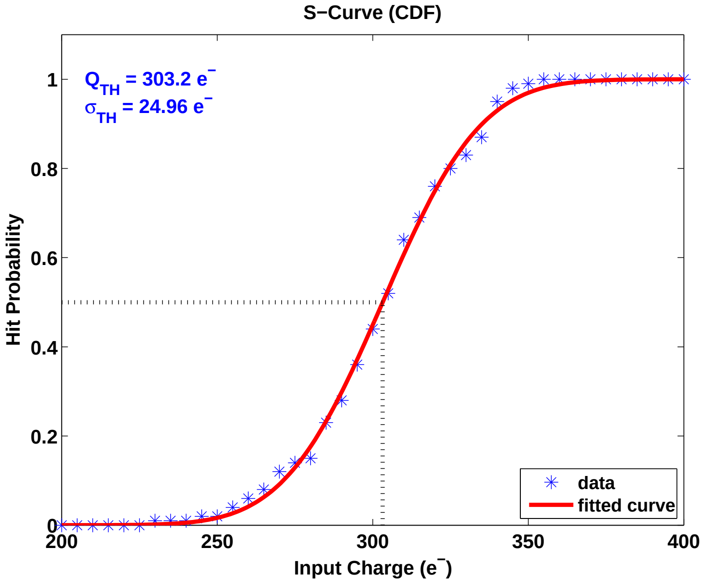


Figure 4.35: Simulated post-layout cumulative distribution function (s-curve) with mismatch enabled. The RMS threshold dispersion is equal to $\approxeq 25\,e^-$ for a nominal threshold of $300\,e^-$.

Baseline adjustment and impact of sensor backbias The NMOS transistors inside the active matrix area are built inside the p-well whose potential is used to reverse bias the sensor and therefore their source-to-bulk voltage V_{SB} depends on the p-well biasing potential that can be as high as $\approxeq 6\,\mathrm{V}$ (for higher V_{SB} junction breakdown occurs). Because the amount of charge in the MOSFET channel and conduction through it is influenced by V_{SB}, the transistor characteristics are affected. The most important consequence of V_{SB} is the modulation of the transistor threshold voltage, called body (bulk)-effect [49]:

$$V_{\mathrm{TH}} = V_{TH_0} + \gamma\left(\sqrt{V_{\mathrm{SB}} + |2\phi_F|} - \sqrt{|2\phi_F|}\right) \tag{4.42}$$

where V_{TH_0} is the threshold voltage with $V_{\mathrm{SB}} = 0$, $\phi_F = (KT/q)\ln\left(N_A/N_i\right)$ is the Fermi potential of the bulk and γ is the body-effect constant and has units of $\sqrt{V}$.

Due to the body-effect, the threshold of the discriminator input transistor M_{11} is increased as the sensor backbias becomes higher. At the same time, the pre-amplifier output baseline voltage will decrease due to the higher threshold and therefore higher gate-to-source voltage (V_{gs}) of transistor

M_2. As a result the front-end detection threshold will significantly increase because a higher output voltage amplitude is required to activate the discriminator. The variation of the gate-to-source voltage of transistors M_{11} and M_2 with the p-well potential is shown in Fig. 4.36. To compensate for the increased transistor threshold, the V_{CASN} voltage has to be increased in order to keep the detection threshold constant. To avoid manually tuning V_{CASN} for different p-well bias voltages, the V_{CASN} generator circuit shown in Fig. 4.37 has been employed to generate the V_{CASN} potential at the column level.

Transistors M_2, M_3 of the generator circuit imitate the feedback branch of the front end circuit (transistors M_2 and M_7) and are biased by the same I_{THR} current. M_1 is a diode-connected NMOS transistor that sinks a current equal to $I_{\text{THR}} + I_{\text{CASN}}$. The generated V_{CASN} voltage will be equal to $V_{gs_1} + V_{gs_2}$ and depends on the value of I_{THR} and I_{CASN}. Because the V_{gs} of the generator M_2 transistor is approximately equal to the V_{gs} of the pre-amplifier M_2 transistor, the output baseline is approximately equal to the gate-to-source voltage of M_1 (V_{gs_1}). Therefore, the current I_{CASN} is used to control the baseline voltage and as a consequence the front-end threshold. Because the threshold shift due to backbias of the front-end transistors M_2 and M_{11} is approximately equal to the threshold shift of the generator transistors M_2 and M_1, the V_{CASN} voltage is automatically adjusted. The generated V_{CASN} voltage compensates the V_{gs} shift of the front-end transistors M_2 and $M_1$1 and is also shown in Fig. 4.36. Furthermore, the V_{CASN} generator circuit helps to reduce the front-end threshold variations due to the horizontal voltage drop gradient across the power distribution lines because the V_{CASN} generator circuit is powered by the local ground of the pixels belonging to the specific column.

It is worth noting that while a negative p-well voltage is necessary to bias the sensor in the case of the DC-coupled front-end pixel variations, the AC-coupled HV version uses front-side sensor biasing at the collection electrode and is not affected by the body-effect. Apart from the threshold variation, other transistor parameters such as the transconductance and output resistance are influenced resulting in a degradation of the pre-amplifier gain up to 40% for a backbias (p-well) voltage of −6 V. Therefore, the pre-amplifier gain is expected to be higher in the case of the HV front-end.

Radiation tolerance considerations Radiation effects that influence the front-end operation and performance include the sensor leakage current increase and TID damage that affects the transistor characteristics and can additionally create parasitic leakage paths. The leakage current (I_{leak}) of an unirradiated sensor is very small, below 1 pA. Therefore, the shot noise (eq. 3.21) contribution is negligible compared to other noise sources. After irradiation to 10^{15} n_{eq}/cm^2 NIEL, I_{leak} in room temperature is significantly increased to values in the order of 100 pA. However, the temperature during normal detector operation is approximately -30 °C. In these conditions, the sensor leakage current is decreased below 10 pA due to its temperature dependence (eq. 3.12). While it still contributes an increased amount of shot noise compared to the unirradiated case, it is not high enough to significantly affect the front-end performance. Nevertheless, the front-end circuit should be able to correctly operate for all possible leakage current values with adequate margin. Therefore, in the case of PMOS input reset, the range of current I_{RESET} is high enough to cope with I_{leak} values even higher than 1 nA. In the case of diode input reset, as explained, the increased leakage current provided through the reset diode will decrease its effective resistance resulting in a faster return to baseline. For high leakage current values, as for example $I_{\text{leak}} = 100$ pA, the short time constant at the input will result to a small signal loss in the order of 5%.

Due to the RINCE and RISCE effects (section 3.1.2.6), the influence of TID damage on MOSFET

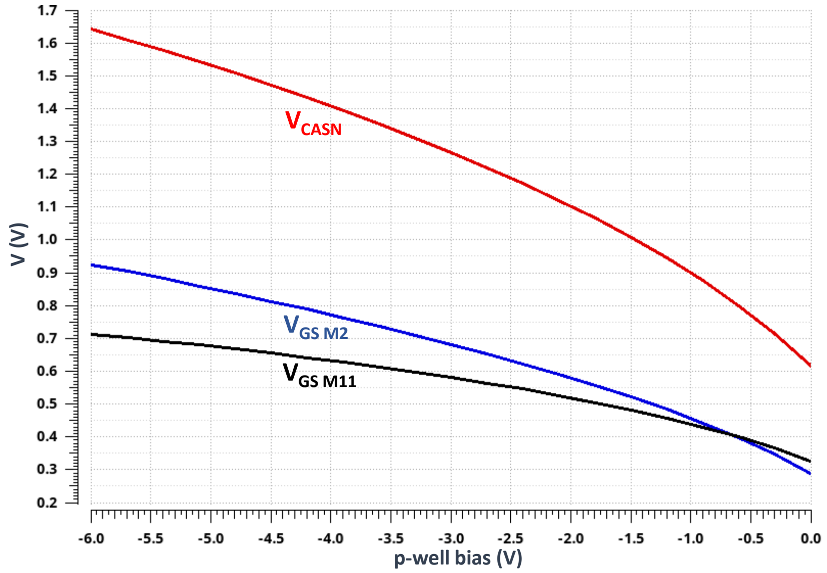


Figure 4.36: Simulation of the gate-to-source V_{gs} voltage increase of transistors M_2 and M_{11} with backbias and compensation by automatically adjusting the V_{CASN} voltage.

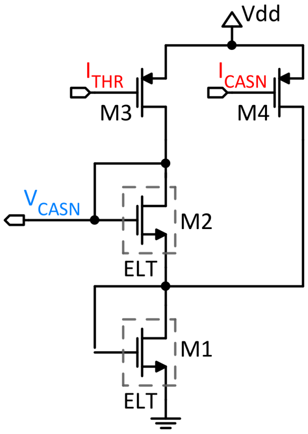


Figure 4.37: V_{CASN} generator circuit used to automatically adjust the pre-amplifier output baseline in order to correct for backbias and ground (V_{ss}) rail voltage drop.

transistors depends on their size and geometry. Fig. 4.38 shows an example of the threshold voltage shift as a function of TID for NMOS transistors in the TowerJazz 180 nm technology with minimum length and different gate widths. As can be observed, small width (and area) transistors are more severely affected with a threshold shift up to ≅ 40 mV above 10 Mrad for a minimum width device. In contrast, the impact on large transistors ($W > 2$ um) is only marginal. Above 10 Mrad the threshold shift starts to gradually recover as negative charge builds-up due to interface traps. It is therefore important to avoid transistors with minimum width and length and enlarge the critical devices as much as possible taking into account other design aspects such as node capacitances and layout constraints.

Increasing the transistor dimensions may be enough to enhance radiation tolerance in current branches of several tens or hundreds of nA, however additional radiation hardening measures are required for transistors conducting small currents, especially in the case of NMOS devices, due to the parasitic leakage current induced by positive charge accumulation in the STI and spacer oxides. For instance, leakage currents in the order of 100 pA have been reported for minimum size NMOS devices after a dose of approximately 20 Mrad TID. The most sensitive part of the front-end pre-amplifier is the feedback branch due to the small values of I_{THR}, which can be even lower than 1 nA. Therefore, in order to increase radiation tolerance up to the required levels of 50 – 80 Mrad, transistor M_2 is designed with an enclosed layout (ELT). Due to the ELT geometry, the transistor size is quite large and its W/L ratio is constrained ($\gtrsim 3$). While the large area of M_2 helps to reduce the threshold shift and further increase its radiation hardness, it results in a higher capacitance. The central diffusion (source), which has a smaller area and lower capacitance that the outer diffusion (drain) is connected to the sensitive output node. Furthermore, M_2 is surrounded by a p$^+$ guard ring that interrupts possible parasitic current paths to neighboring devices. Radiation tolerance of the V_{CASN} generator circuit is also enhanced by implementing transistors M_1 and M_2 with an ELT layout.

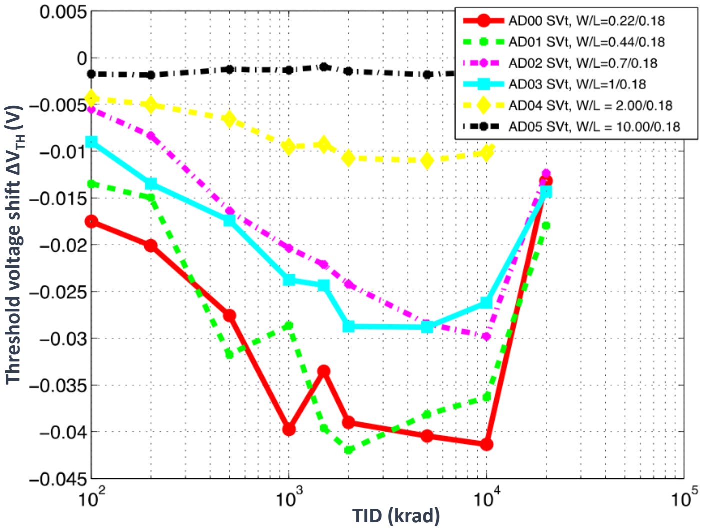


Figure 4.38: Threshold voltage shift as a function of TID for NMOS transistors with varying gate width and minimal gate length of 0.18 µm. Data points at $2 \cdot 10^4$ krad correspond to 24 h of annealing [79]

.

Front-end performance summary A summary of the TJ-Monopix1 front-end characteristics and performance, discussed in the previous sections is presented in Table 4.5.

Table 4.5: TJ-Monopix1 front-end performance summary (simulated)

Metric	Value
Threshold (nominal)	$300\,e^-$
In-time threshold	$340\,e^-$
Overdrive	$40\,e^-$
ENC (AC noise method)	$9.5\,e^-$
ENC (S-curve method)	$8.9\,e^-$
Threshold dispersion	$25\,e^-$
Gain at threshold	$0.4\,\mathrm{mV}/e^-$
Phase margin	90°
ToT at MPV ($1\,600\,e^-$)	1 µs
Power consumption	1 µW

4.3.1.2 In-pixel digital logic implementation

The in-pixel digital readout (R/O) part, shown in Fig. 4.39, contains the control and arbitration logic, the 6-bit LE and TE SRAM memory (12 SRAM cells) and the 9-bit pixel address ROM. The front-end HIT output is connected to an edge-detector that generates the TE and LE pulses. The edge detector works by comparing the HIT pulse with a delayed version of the same signal generated by a buffer chain taking advantage of the gate propagation delay. The LE and TE pulses are used as WRITE enable signals for the storage of the BCID value to the SRAM cells. The width of the LE and TE pulses is determined by the buffer chain delay and is designed to be short (a few ns) but adequate for the SRAM cell to store a new bit. The edge detector can also be disabled in order to avoid new hits overwriting the LE and TE information before the previous hit is read out.

The hit event is registered by a first latch activated by the TE pulse and the edge detector is disabled. If the column is not frozen (the FREEZE signal is not asserted), the HIT flag, stored in a second latch, is set and the readout sequence begins. The HIT flag activates the token signal that propagates through the column and informs the R/O controller that hits are ready to be read out. At the same time, it prohibits lower priority pixels to access the column data bus during the read phase. The token signal is passed on each consecutive pixel through an OR logic between the pixel HIT flag state and the token of all higher priority pixels. Therefore, one OR gate per pixel is required for the token propagation. However, native OR gates cannot be implemented in CMOS logic and are usually constructed by a NOR gate followed by NOT gate (inverter). The drawback of this implementation is that two gate delays are added to the token signal per pixel, significantly increasing its propagation delay. To speed-up the token signal propagation and also reduce the area, a NAND-NOR token pass logic is used that requires only one gate per pixel. Furthermore, because the NAND-NOR scheme inverts the polarity of the token signal at each consecutive pixel, coupling to neighboring column-bus long metal wires is reduced.

After the token is received by the R/O controller, the FREEZE signal which prevents the HIT flag of other pixels to be set due to new hits, is asserted. Therefore, the priority of pixels to be read out within the specific readout sequence is well defined and no arbitration errors can occur due to new hits. While FREEZE is active, hits at idle pixels (not being previously hit) are not rejected and can still be recorded by the first (TE) latch that helps to improve the readout efficiency. Afterwards, a series of read cycles controlled by the READ signal take place, during which pixel data is transferred to the periphery through the column data-bus. At each read cycle, the pixel that has an active HIT flag and the highest priority i.e. the token at its input (Token in) is inactive, is allowed to use the data-bus. At the rising edge of READ, the state of the pixel that fulfills these two requirements is stored in a D-latch. At the same time, the TE and HIT flag latches are reset. If the D-latch is set and while READ is active, an internal read (READINT) signal is produced that controls the switches that enable access of the LE/TE SRAM cells and the address ROM transistors to the data-bus. The D-latch is necessary in order to be able to reset the HIT flag at the start of the read cycle and therefore extend by t_{RD} (read pulse width) the available time for token propagation to the EoC before the end of the data transmission (DTA) phase. As mentioned in section 4.2, if the token signal arrives late, an extra empty read cycle will occur at the end of each readout sequence. Note that even though the TE latch is reset with the rising edge of READ, the edge detector remains disabled until the read phase ends.

A simulation of the readout operation is shown in Fig. 4.40. In this example two neighboring pixels are being hit at the same time and the waveforms correspond to the pixel with lower priority. As can be observed, the READINT signal is produced only at one pixel for each read cycle following the order defined by the pixel priority.

SRAM memory cell and pixel address ROM The LE/TE memory uses a two-port SRAM cell with a buffered "source-follower" readout shown in Fig. 4.41. Transistors $M_1 - M_4$ constitute a bi-stable latch which functions as the memory element. M_5 and M_6 are pass-transistors that are used to write the memory cell. While the WRITE signal (LE/TE pulse) is high, the memory element is connected through M_5 and M_6 to the differential BCID bit lines that set the memory state. Differential BCID transmission, apart from reducing crosstalk and electromagnetic interference, also reduces the timing uncertainty due to jitter and metastability effects by effectively doubling the signal amplitude and suppressing common-mode noise. This is especially important for pixels at the column top, far from the BCID drivers located at the EoC, due to the increased rise/fall time as a result of the line R-C low pass effect.

A similar pass-transistor switch mechanism is not suitable for the readout of the SRAM cell because the memory element transistors ($M_1 - M_4$) would have to drive the large capacitive load of the long data-bus metal wires and risk flipping the memory state thus leading to corrupt data. Instead, buffer transistors M7 and M8 are used to isolate the memory element nodes (MEM and nMEM) from the data-bus bit lines. M_9 and M_{10} are switches that control the access of the SRAM cell to the data-bus and are activated by the READINT signal. The SRAM memory readout also uses differential signal transmission in order to improve the signal to noise ratio (SNR) and achieve faster data transmission. During the read phase, if the memory state equals 1, transistor M_7 will pull-up the bit-line (Bline), while transistor M_8 will be switched off and the complementary bit-line (nBline) will be pulled-down by the EoC column-bus readout circuit. The opposite will happen in the memory state is 0. Essentially, transistors M_7 and M_8 function as two source-followers biased by current sources at the EoC, that in combination with the bit-line pre-charge voltage, control and limit the transient currents during

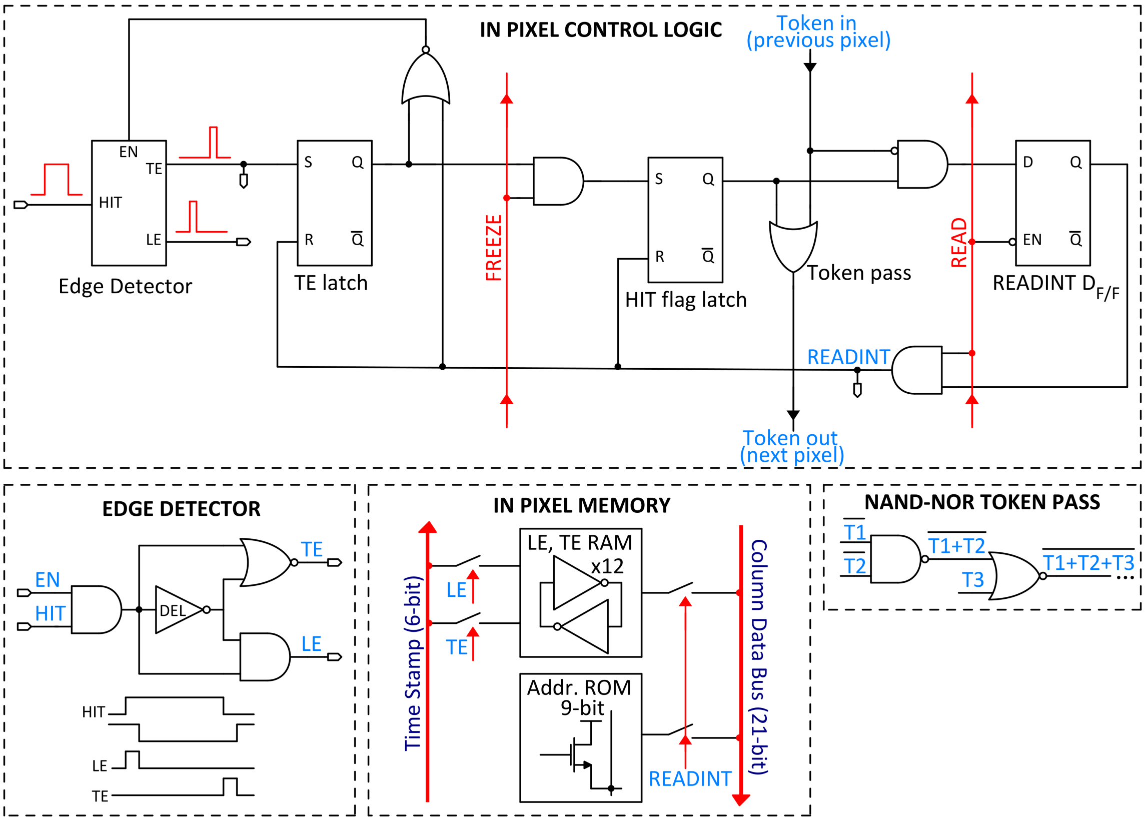


Figure 4.39: Schematic of the TJ-Monopix1 in-pixel R/O architecture digital logic.

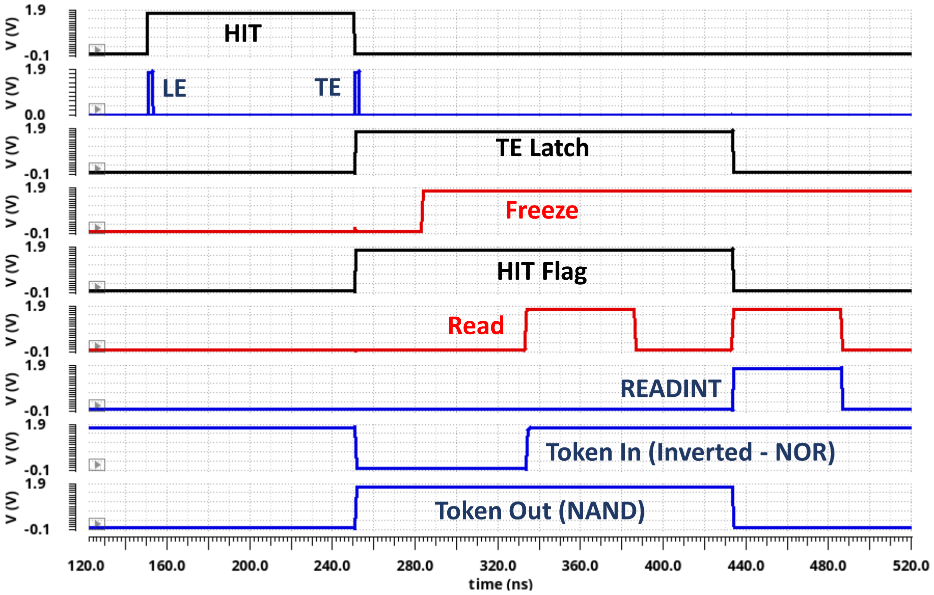


Figure 4.40: Simulation of the in-pixel R/O logic.

the read phase in order to suppress crosstalk interference. The so-called source-follower column-bus readout scheme will be more expensively described in section 4.3.2.1

The pixel address ROM is comprised of nine simple transistor switches, activated by the READINT signal that are hardwired to either the bit-line (Bline) or its complementary (nBline). Each pixel has a unique address bit combination. Depending on the address bit value, either Bline or nBline will be pulled-up and the other will be pulled-down. Similar to the SRAM memory, the address ROM transistors are also connected in a source-follower configuration.

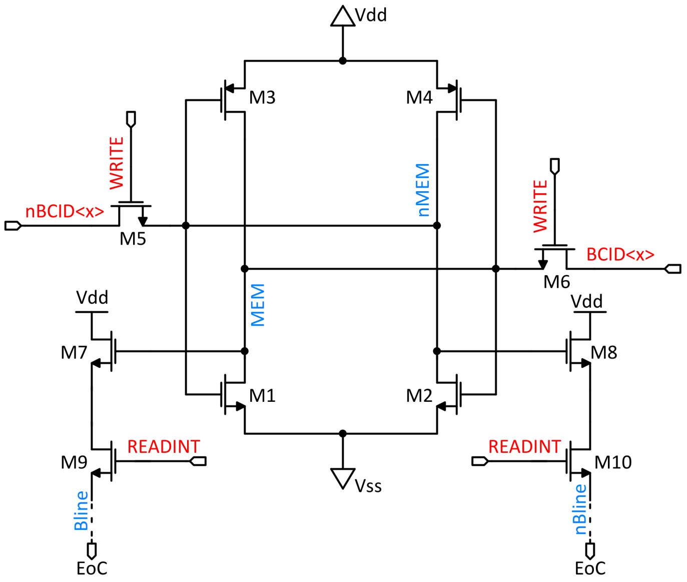


Figure 4.41: Schematic of the LE/TE SRAM cell used in TJ-Monopix1.

4.3.1.3 Test features

Hit injection Artificial hits can be injected for testing purposes through an injection capacitance C_{inj} connected at the collection electrode node. If a negative voltage step with amplitude ΔV_{inj} applied to C_{inj}, the equivalent electron charge injected at the sensor Q_{inj} will be equal to:

$$Q_{\mathrm{inj}} = \frac{C_{\mathrm{inj}} C_D}{C_{\mathrm{inj}} + C_D} \Delta V_{\mathrm{inj}} \approxeq C_{\mathrm{inj}} \Delta V_{\mathrm{inj}} \tag{4.43}$$

Since C_{inj} and the input node capacitance $C_{\mathrm{inj}} \approxeq C_D$ are in series, the injection capacitance has to be much smaller than the detector capacitance C_D such that the injected charge Q_{inj} depends only on the value of C_{inj} and the voltage step amplitude ΔV_{inj}. C_{inj} is implemented by a simple MOM capacitor over the collection electrode and is designed with a value of approximately 230 aF. Therefore the injected charge (in electrons) is equal to:

$$Q_{\mathrm{inj}} = \frac{230\,\mathrm{aF}}{q_{e^-}} \Delta V_{\mathrm{inj}} = 1.4375\ {}^{e^-}\!/_{\mathrm{mV}} \tag{4.44}$$

The full hit injection circuit is shown in Fig. 4.42. Two voltage levels called V_H and V_L are provided by the voltage DAC. The injection capacitance is initially connected to V_H through transistor M_2, while M_1 is switched off. In order to create a negative voltage step and inject charge emulating electrons being collected by the sensor, at the arrival of the injection pulse the gate voltage of transistors M_2 and M_1 is inverted and C_{inj} is connected to V_L. Therefore, the voltage step is equal to $\Delta V_{\text{inj}} = V_H - V_L$. Because the voltage DAC Least Significant Bit (LSB) is equal to 14.06 mV for $V_{\text{DD}_{\text{DAC}}} = 1.8$ V, the injected charge resolution is equal to:

$$Q_{\text{inj}} = 1.4375\ {}^{e^-}\!/\text{mV} \cdot 14.06\,\text{mV} \approxeq 20.2\ {}^{e^-}\!/_{DU} \tag{4.45}$$

where DU is one DAC unit (LSB). An global injection strobe PULSE signal, generated by the DAQ FPGA is applied to time the artificial hit injection. The pixels to be injected are selected using a projection scheme with column and row enable and an internal pulse that controls transistors M_1 and M_2 is generated internally in the specified pixels at the arrival of the PULSE signal.

In the case of the AC-coupled pixels (HV flavor), C_{inj} could not be implemented above the collection electrode because this area is occupied by the coupling MOM capacitor. Therefore, C_{inj} is placed near the front-end input and due to the different layout it has a higher capacitance yielding an injected charge approximately equal to 35 ${}^{e^-}\!/_{DU}$.

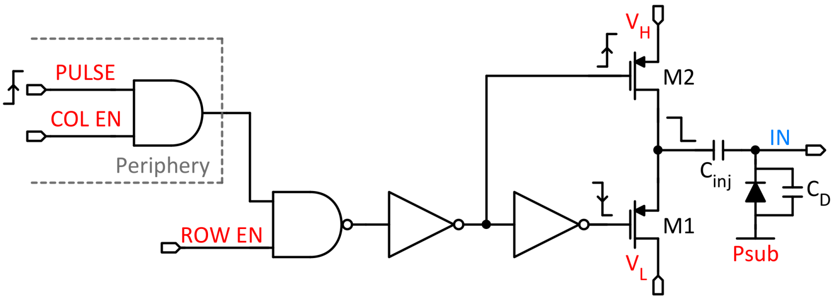
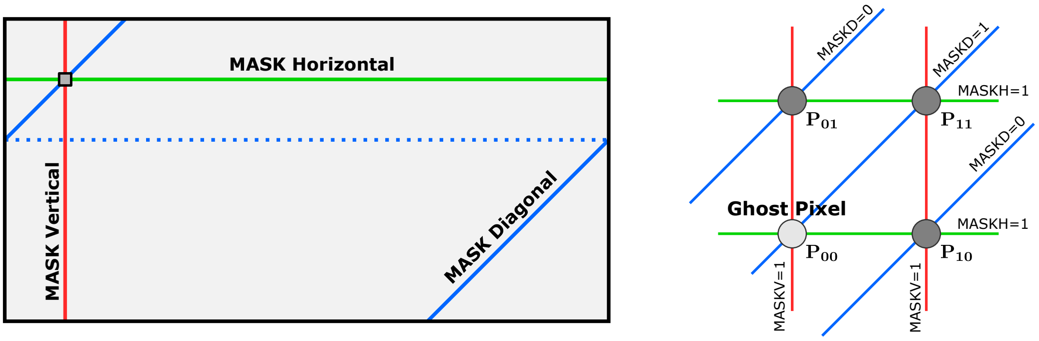


Figure 4.42: Artificial hit injection circuit used in TJ-Monopix1.

Pixel masking A certain number of pixels has to be disabled because their noise rate is very high due to manufacturing defects or low threshold as a result of transistor mismatch. Usually these outlier pixels are lower than 1% of the total pixel number. Masking is also applied in order to achieve a lower operating threshold while keeping the noise rate below the specification by disabling a percentage of the pixels in the lower part of the threshold distribution.

Usually a configurable masking register is included in the pixel such that each pixel can be disabled individually. The drawback of a masking register and configuration logic is its sensitivity to SEU effects that can corrupt the stored masking bit. To reduce SEU sensitivity, a more sophisticated triple modular redundancy (TMR) logic is required that requires a significant amount of area. If only a few pixels need to be masked, an alternative coordinate projection scheme can be implemented instead. The drawback of this approach is that for a given amount of intentionally masked pixels, there is a number of unintentionally masked ("ghost") pixels. Therefore, whether a pixel is disabled cannot be individually controlled as it can also depend on the state of other pixels. In the case of a simple two coordinate projection scheme, the number of ghost pixels scales with N^2, where N is the number of intentionally masked pixels. Therefore it is suitable only if an extremely small number of pixels has to be masked.

To decrease the percentage of ghost pixels, a three coordinate projection masking implementation using horizontal (row), vertical (column) and diagonal vectors, has been employed in TJ-Monopix1 and is illustrated in Fig. 4.43. However, the number of ghost pixels still scales with a power law N^a (where $a < 2$). Therefore, if more than approximately 100 pixels of a flavor are intentionally masked, the number of the masked and ghost pixels quickly grows above 1% of the total pixels of the flavor (25088). As will become evident from chip characterization, this is an important limiting factor that prohibits low threshold operation. Thus, this approach has not been used by TJ-Monopix2.



(a) (b)

Figure 4.43: a) Pixel masking scheme implemented in TJ-Monopix1 based on three coordinate projection, b) Example of a pixel being un-intentionally masked (ghost).

Hit OR In order to provide access to the pixel discriminator output for debugging and precision timing measurements, a HIT-OR logic illustrated in Fig. 4.44, has been implemented. Instead of standard CMOS gates, a wired-OR logic that requires only two transistors per pixel has been used to save area. HIT-OR can be enabled for multiple rows through transistor M_2 and the HIT-OR output of the selected pixels is connected at the column level to a common pull-up resistor located at the EoC. Transistor M_3 acts as a switch that is used for HIT-OR column select. The value of the pull-up resistor should be relatively high to improve the timing precision of the HIT-OR leading edge that is important for the front-end timing characterization. However, a large resistor value distorts the trailing edge timing and increases the ToT by a few tens of ns. Because a leakage current I_{leak} through NMOS transistors M_1 and M_2 due to TID damage would result in a large accumulated DC current through the pull-up resistor equal to $224 \cdot I_{\text{leak}}$ (224 pixels per column), M_2, as well as M_3 at the EoC are designed with an ELT geometry.

Analog monitoring Having the ability to probe the front-end analog output is very useful for the characterization of the front-end circuit, optimization of voltage and current bias settings and identification of possible issues with the chip operation. Although the analog output of regular pixels can not be externally accessed, four special cells are included at each side of the matrix that consist of the front-end pre-amplifier and an analog buffer to drive the output pads. The analog monitoring cell schematic is shown in Fig. 4.45. The analog buffer is composed of two stages. The first stage is a PMOS source follower that is placed as close as possible to the front-end output and drives the

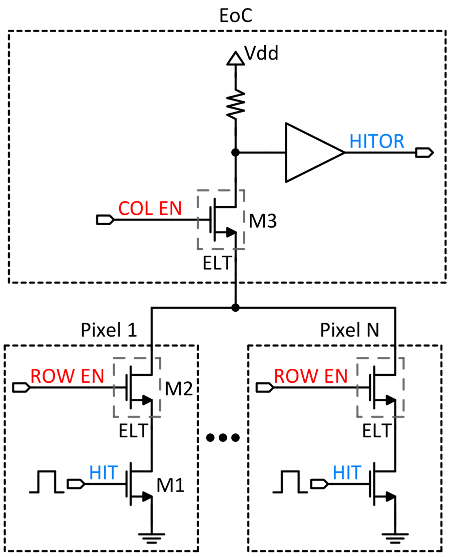


Figure 4.44: Schematic of the HIT-OR circuit used in TJ-Monopix1.

second stage NMOS source follower that is placed outside the matrix area. The bias current of both stages is configurable. The NMOS source follower bias current is selected according to the total capacitance at the output pad C_{PAD}, that is usually dominated the oscilloscope probe. The output resistance of transistor M_{11} should be high enough such that the time constant at the PAD output $\tau_{\mathrm{PAD}} \approxeq C_{\mathrm{PAD}}/g_{m_{11}}$ is significantly higher than the front-end output rise time. The analog buffer dynamic range is limited by the PMOS source follower. The maximum pre-amplifier output voltage amplitude without compression is equal to:

$$v_{\mathrm{OUTA}}^{\mathrm{max}} = V_{\mathrm{dd}} - \left(V_{\mathrm{dsat}}^{M_9} + V_{\mathrm{gs}}^{M_8} + V_{\mathrm{OUTA_{BL}}}\right) \tag{4.46}$$

The dynamic range limitation becomes more significant when p-well backbias is applied because the analog output baseline $V_{\mathrm{OUTA_{BL}}}$ increases as a result of the body-effect and can be as high as 700 mV for the maximum reverse p-well bias voltage of 6 V. The PMOS source follower bias current should be large enough to drive the second stage but not too high in order to reduce $V_{\mathrm{gs}}^{M_8}$ as much as possible and extend the dynamic range.

The list of the implemented analog monitoring cells is presented in Table 4.6. All front-end versions that exist in the pixel matrix are included (PMOS reset, leakage compensation and HV) with full (FDPW) and reduced (RDPW) deep p-well layout variations.

4.3.2 Chip design and architecture

The TJ-Monopix1 chip architecture is shown in Fig. 4.46. The pixel matrix is organized in double columns. Each double column consists of 448 pixels split in two physical columns. While pixel configuration and control is based on the physical row and column, from a digital readout perspective all 448 pixels share the same column bus and are logically regarded as part of a single column with address space as shown in Fig. 4.47.

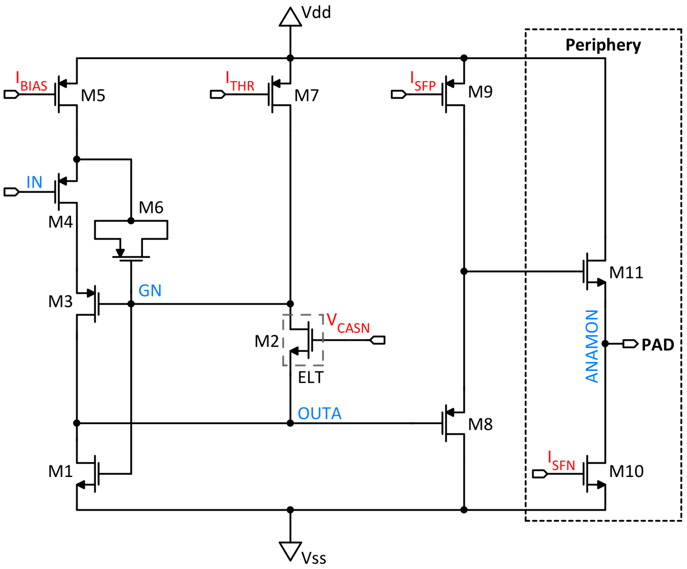


Figure 4.45: Schematic of the special analog monitoring cell that includes the front-end pre-amplifier and an analog buffer.

Table 4.6: Analog monitoring front-end cells included TJ-Monopix1.

Left side	Right side
PMOS Reset FDPW	PMOS Reset FDPW
Leakage Comp. FDPW	Leakage Comp. FDPW
PMOS Reset RDPW	HV FDPW
Leakage Comp. RDPW	HV RDPW

The end of column (EoC) block supports and controls the double column readout and processes the pixel data. It consists of two parts. The first part is a custom block that is placed directly below each double column and contains digital buffers that drive the BCID and control signals and the data-bus readout circuitry (sense amplifiers). The second part is synthesized and implemented using digital tools along with the rest of the digital chip bottom and contains the EoC readout logic. As mentioned, no trigger memory is included and the matrix readout is direct (continuous). Since each flavor has a standalone separate readout, only one pixel of each flavor can be read-out at a time. Therefore, the EoC logic contains a token based priority mechanism, similar to the in-pixel control logic, to arbitrate the double column readout with a direction from left to right. When a token signal arrives at the EoC block, it is passed through an OR logic and transmitted to the off-chip readout (R/O) controller. The global READ and FREEZE signals, issued by the R/O controller, are distributed to across the EoC blocks and local copies are transmitted only to the double column that has been hit and has the highest priority. Data from the pixel that is read-out (6-bit LE time-stamp, 6-bit TE time-stamp, 9-bit pixel address) is received by the sense-amplifiers and is temporarily stored at the EoC. The 6-bit column address is subsequently appended and the complete hit data packet (27-bit) propagates through each lower priority EoC block and arrives at the serializer. The serialized data is transmitted off-chip to the DAQ FGPA by CMOS level drivers.

An external 40 MHz clock, produced by the DAQ FGPA, is used to synchronize and drive a 6-bit gray code counter that generates the BCID time-stamp. The serializer speed is determined by another externally provided clock signal. Although the serializer has been designed for frequencies up to 160 MHz, it is normally operated at 40 MHz to avoid signal integrity issues at the CMOS driver output. The analog bias voltage and current signals are generated by a 7-bit DAC with monitoring and external override capabilities. The chip is controlled and configured using a simple shift-register array that can be programmed using a serial peripheral interface (SPI) protocol. Therefore, the raw register values are transparent to the DAQ and testing system and all high-level calculations (e.g. masking vectors, DAC configuration bit encoding) are handled by the DAQ software.

4.3.2.1 BCID distribution and timing

As explained in section 4.2, the BCID timing dispersion across the whole pixel matrix has to be small (typically below ≊ 4 ns) because it effectively reduces the available front-end time-walk budget and leads to a higher in-time threshold. The two main components of BCID timing variations is the signal propagation delay across the column due to the metal wire low-pass effect and the delay introduced between different columns due to the signal distribution path at the periphery.

Since the periphery digital chip bottom is implemented by automated place and route tools, the timing of specific signals can be constrained and is automatically adjusted by introducing additional gate delays. The success of achieving the desired timing depends on the sophistication of the HDL (hardware description language) code and the layout area and complexity. To further improve BCID timing uniformity at the periphery, an extra re-timing flip-flop has been placed close to each EoC block. As a result the timing dispersion at the input of the BCID drivers at the EoC has been simulated to be approximately only 0.1 – 0.2 ns.

The propagation delay across the column depends on the column length, the wire geometry, distance to neighboring metals, technology dependent parameters (sheet resistance, parasitic capacitance) and the driver strength. The column bus width, and therefore the wire width and spacing is constrained by the small pixel size. Additionally shielding metals have been placed next to the BCID lines to reduce interference to neighboring signals. Therefore, due to the relatively high wire resistance and strong capacitive coupling, the propagation delay can be significant across the 8 mm long column. A simulation of BCID delay between the bottom and top of a column is shown in Fig. 4.48. At the digital threshold of approximately 0.9 V, the delay is equal to 2.5 ns. Therefore the total BCID timing dispersion (≊ 2.7 ns) is well below the 4 ns specification.

4.3.2.2 Column-bus readout

During the read phase, hit data from the pixel has to be transferred to the EoC through the column data-bus wires. Because the pixel has the role of the transmitter, the transient current it provides to charge the data-bus bit lines has to be controlled and limited in order to prevent crosstalk to the sensitive analog part through the substrate and power grid. The long data-bus wires form distributed RC lines due to the metal resistance and parasitic capacitance. The RC bit-line low-pass effect on the transmitted signal causes a delay that depends on the resistance and capacitance values as well as the current supplied by the pixel. Therefore, the data transmission time which is controlled by the READ

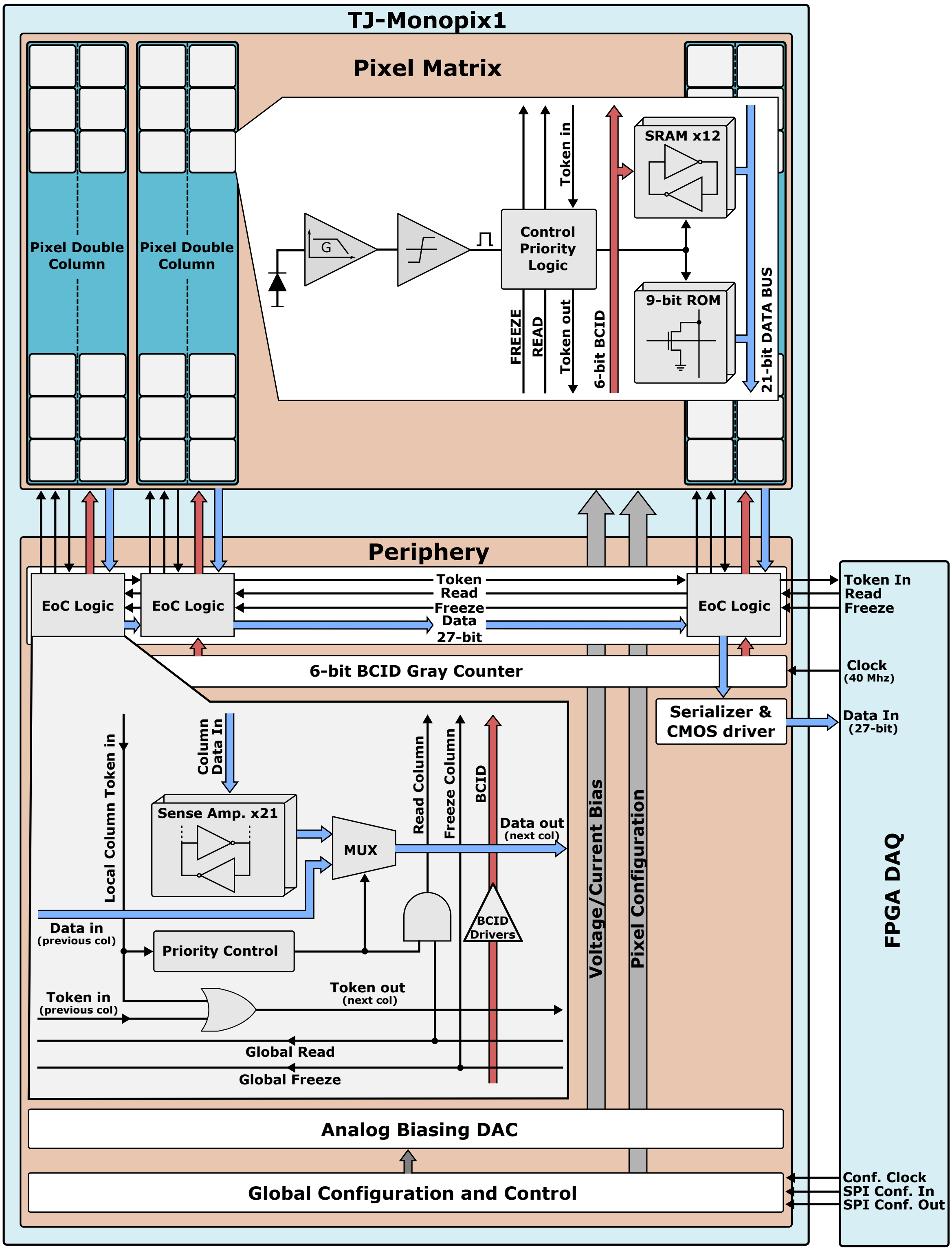


Figure 4.46: Chip Architecture of TJ-Monopix1

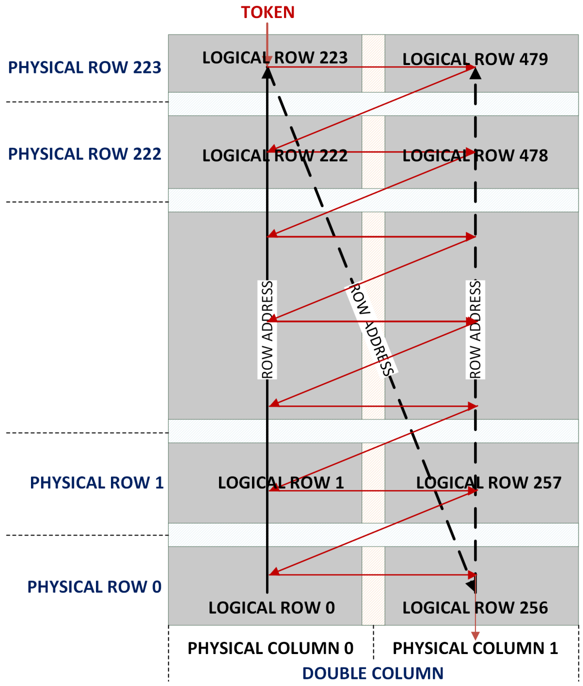


Figure 4.47: Pixel address mapping within a double column of TJ-Monopix1. The token propagation path is also highlighted.

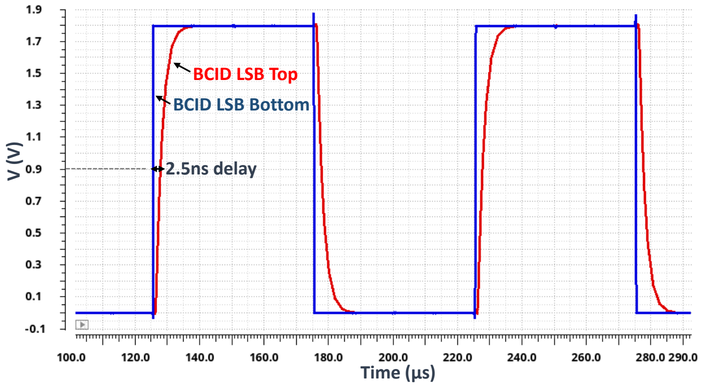


Figure 4.48: Simulated delay of the BCID time-stamp across a pixel column of TJ-Monopix1. The waveforms correspond to one of the two differential components of the BCID LSB bit line at the bottom and top of the column.

pulse width, has to be long enough such that an adequate bit-line voltage amplitude is developed at the EoC end. In other words, the bit line parameters as well as the driving current determine the maximum readout speed.

Each data-bit is transmitted deferentially using two bit-lines (Bline, nBline) in order to increase the signal to noise ratio suppress interference to neighboring lines. A sense-amplifier receiver circuit is used at the EoC to increase the readout speed since the bit-line voltage does not have to reach digital CMOS levels, but only needs to be high enough in order to be reliably detected by the sense-amplifier. Between each read cycle, the bit-lines are reset and pre-charged to an adjustable common-mode voltage (V_{PC}) to prevent additional delays and eliminate the influence of the previous state.

The complete data-bus readout circuit is illustrated in Fig. 4.49. Data transmission is based on a source-follower scheme, derived from LF-Monopix1 [53]. Before the read phase, the transistor switches M_2 and M_4 in the pixel are open, while M_6 and M_8 are switched on. The differential bit-lines are pre-charged by transistors M_5 and M_7 that are connected in a source-follower configuration and are biased by the current sources I_{SF}. The pre-charge voltage (V_{PC}) can be controlled by adjusting the gate voltage potential of M_5 and M_7 ($V_{\mathrm{PC}_{\mathrm{SF}}}$) and is equal to (assuming a simple square-law MOSFET model):

$$V_{\mathrm{PC}} \cong V_{\mathrm{PC}_{\mathrm{SF}}} - V_{\mathrm{gs}}^{M_{5,7}} = V_{\mathrm{PC}_{\mathrm{SF}}} - \sqrt{\left(\frac{L}{W}\right)_{M_{5,7}} \frac{1}{\mu_n C_{\mathrm{ox}}} I_{\mathrm{SF}}} - V_{\mathrm{TH}}^{M_{5,7}} \tag{4.47}$$

where L and W are the transistor dimensions, C_{ox} is the gate capacitance per unit area, μ_n is the electron mobility and V_{TH} is the transistor threshold voltage.

During the read phase M_6 and M_8 are switched off, while M_2 and M_4 are switched on by the READINT signal. Therefore, M_1 and M_2 are connected to the differential bit-lines and operate as source-follower buffers in the place of M_5 and M_7. The input (gate) voltages of M_1 and M_2 are complementary and depending on the in-pixel memory (SRAM or address ROM) state will be either V_{dd} ("1") or V_{ss} ("0"). The source-follower transistor that is connected to V_{dd} will pull the corresponding bit-line up, while the complementary bit-line has no current path to V_{dd} and is pulled down by I_{SF}. Therefore, a differential voltage is developed at the input of the sense-amplifier.

The transient current, drawn from the bit-lines that are pulled up, depends on the source-follower current I_{SF}, the pre-charge voltage V_{PC} and the transistor M_1 or M_2 dimensions. The drain current of transistor M_1 (or M_2 depending on the memory state) is equal to:

$$I_D^{M_{1,2}} \cong \left(\frac{W}{L}\right)_{M_{1,2}} \mu_n C_{\mathrm{ox}} (V_{\mathrm{dd}} - V_{\mathrm{BL}} - V_{\mathrm{TH}}^{M_{1,2}})^2 \tag{4.48}$$

where V_{BL} is the bit-line voltage. The pre-charge voltage is set close to the value at which the bit-line settles ($V_{\mathrm{BL}_{\mathrm{fl}}}$) and is determined by the dimensions of $M_{1,2}$ and the source-follower current I_{SF}:

$$V_{\mathrm{BL}_{\mathrm{fl}}} \cong V_{\mathrm{dd}} - V_{\mathrm{gs}}^{M_{1,2}} = V_{\mathrm{dd}} - \sqrt{\left(\frac{L}{W}\right)_{M_{1,2}} \frac{1}{\mu_n C_{\mathrm{ox}}} I_{\mathrm{SF}}} - V_{\mathrm{TH}}^{M_{1,2}} \tag{4.49}$$

Therefore $I_D^{M_{1,2}}$ will be approximately equal to I_{SF} throughout the read phase. The complementary bit-line is pulled down by I_{SF} and the discharge rate depends on its RC characteristics (mainly the total line capacitance) and the current I_{SF}, which essentially determines the maximum readout speed (minimum READ pulse width). Simulation results indicate that the selected value of $I_{\mathrm{SF}} = 20\,\mu\mathrm{A}$ is adequate to reliably transfer the pixel data within 25 ns. However, in order increase the error margin under different p-well bias voltages and temperature variations, the READ pulse width duration is nominally set to 50 ns.

The sense-amplifier is based on a positive feedback latch (transistors $M_{11}, M_{12}, M_{15}, M_{16}$) with an input tail transistor pair (M_9, M_{10}). At the beginning of the read phase, the latch outputs are reset to V_{dd} by transistors M_{17} and M_{18} and the feedback path is interrupted by transistor switches M_{13} and M_{14}. At the trailing edge of the READ pulse, after the bit-line voltage has been developed, the reset switches are turned off and the feedback path is restored. The new state at which the latch settles is defined by the difference of the drain current of transistors M_9 and M_{10} that depends on the differential bit-line voltage. The latched bit value is stored by an set/reset (SR) latch and is subsequently transferred to the EoC digital logic.

A simulation example of a bit transfer through the data-bus that demonstrates the readout circuit operation is shown in Fig. 4.50. In the first read cycle a pixel at the column bottom accesses the data-bus, while in the second read cycle the pixel that is read out is located at the column top. As can be observed a large differential bit-line voltage, approximately equal to 750 mV is developed within 50 ns (read pulse width).

The drawback of the source-follower (SF) data-bus readout implementation is the high static power consumption because even when no pixel is being read out, $2 \times 20\,\mu\mathrm{A}$ (differential) is continuously drawn for each bit. Since the data bus is 21-bit wide and there are 125 double columns per cm, the total static power consumption is equal to approximately 190 mW/cm, and is higher than the combined analog front-end and BCID distribution components for an area of $1 \times 1\,\mathrm{cm}^2$. Therefore, since one of the main advantages of TJ-Monopix is its high analog performance, and as a consequence low power consumption, the static SF readout power has to be reduced or eliminated. Such a modification is proposed in the next paragraph.

Gated source-follower readout The SF readout static power consumption can be eliminated while maintaining the same operation principle and benefits such as control of the pixel current by a circuit variation called gated source-follower readout that is illustrated in Fig. 4.51. The idea is to provide the pre-charge voltage V_{PC} directly externally (or generate it on-chip) instead of using transistors M_5 and M_7 (Fig. 4.49). Between each read cycle, transistors M_1 and M_2 are switched on forming a low resistance path to V_{PC}, which pre-charges the bit-lines. At the same time, transistors M_3 and M_4 are switched off interrupting the static current path through the current sources I_{SF}. During the read phase, M_3 and M_4 are switched on while M_1 and M_2 are switched off and the in-pixel source-follower transistors are connected to I_{SF} similar to the SF readout case analyzed in the previous paragraph.

The new gated SF readout has been implemented in the first flavor of TJ-Monopix1. In this case, V_{PC} needs to be produced or buffered by a relatively low impedance source since it provides the necessary current to pre-charge the bit-lines. Additionally a large amount of decoupling capacitors have been included on-chip to reduce fluctuations of the V_{PC} potential.

4.3.2.3 Analog bias generation

The 7-bit DAC that is used to generate the analog current and voltage biases is organized in 128 units, placed below the pixel matrix, and can be configured through the SPI interface. The current DAC (IDAC) architecture is shown in Fig. 4.52. The reference current $I_{\mathrm{REF}} = 140\,\mathrm{nA}$ is generated by a simple diode-connected PMOS transistor in series with a $60\,\mathrm{k\Omega}$ resistor. Each IDAC unit contains

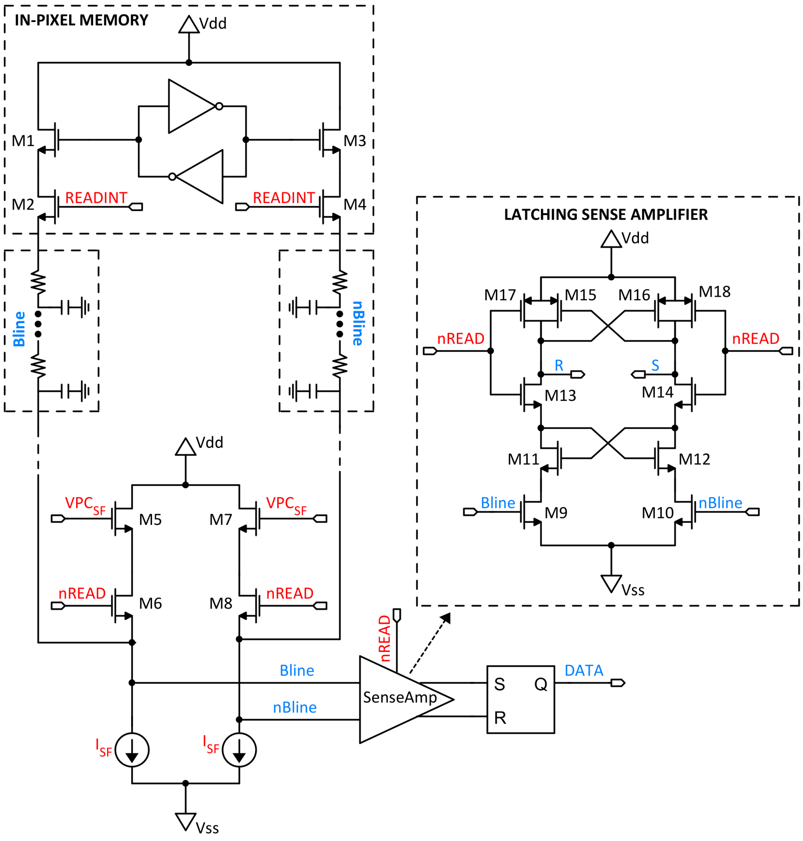


Figure 4.49: The source-follower column data-bus readout scheme used in TJ-Monopix1.

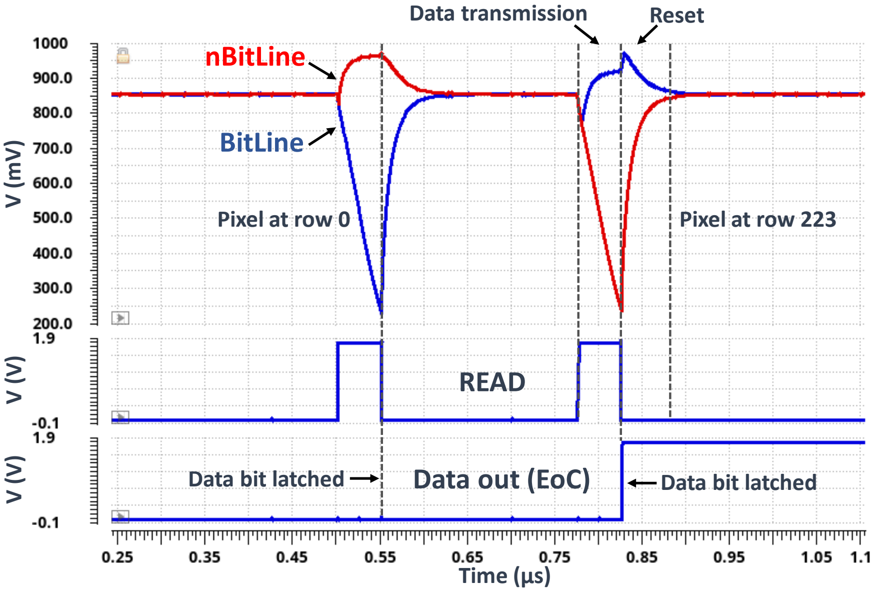


Figure 4.50: Post-layout simulation of a data bit being transferred to the EoC through the data-bus.

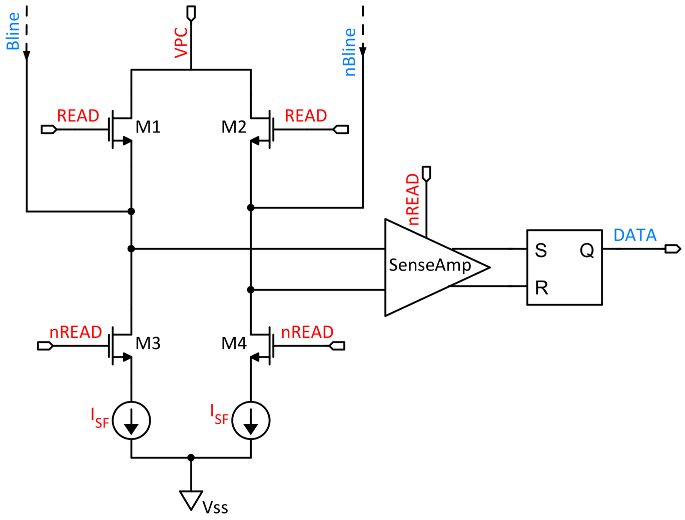


Figure 4.51: Schematic of the gated source-follower data-bus readout variation that eliminates static power consumption.

a current source equal to I_{REF} that can be switched on or off depending on the IDAC code. The IDAC code (SET) ranges from 1 to 128 and is thermometer encoded to an 128-bit array such that the number of activated current sources is equal to the IDAC code value. The total current $I_{\mathrm{REF}} \cdot \mathrm{SET}/128$ is summed at the drain of a diode connected NMOS transistor and is subsequently scaled by two additional mirrors. A diode connected PMOS transistor generates the current source biasing voltage that is distributed to the pixel matrix and forms a final mirror with the front-end PMOS current sources. The ratio of these three mirrors (N,M,K) determines the range and LSB of each current bias. The NMOS transistors of the mirror stages are designed with ELT geometry to enhance tolerance to TID radiation damage.

As previously mentioned, the matrix power pads are distributed along its left and right sides. Therefore, due to the front-end static current (approx. 550 nA), a voltage drop gradient will be developed in the horizontal direction with its maximum in the middle of the matrix. It is estimated that with the available horizontal power grid width, which is limited by the pixel size, the voltage drop will be as high as 5 mV which can lead to significant systematic bias current, and as a result, threshold variations along the horizontal axis. To compensate for the analog power domain (V_{DDA}) voltage drop, the final IDAC diode-connected PMOS transistor is connected to the local analog matrix supply. Therefore, both transistors of the final mirror (IDAC and front-end PMOS) operate at approximately the same supply voltage and bias current uniformity is improved.

Additionally, in order to reduce transistor mismatch effects, the pixel columns are grouped from a biasing point of view in 32 groups consisting of 14 columns each. Each group corresponds to 4 DAC units whose outputs are connected in parallel, effectively increasing the IDAC mirror transistor area by four times. Due to grouping, the voltage drop compensation cannot be applied in a per-column basis, but is common for the whole group. Therefore the group size is a trade-off between accurate voltage drop compensation and mismatch reduction.

The voltage DAC (VDAC), illustrated in Fig. 4.53 is composed of 128 resistors in series between the DAC power supply rails that form a resistor divider. Therefore, 128 potential values (including V_{ss}) are

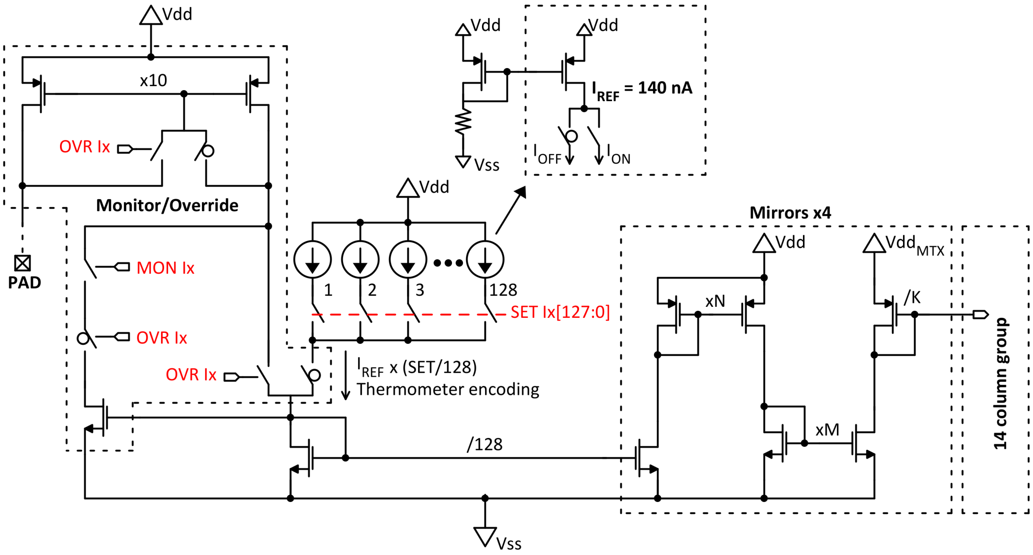


Figure 4.52: Schematic of the analog bias current DAC used in TJ-Monopix1.

available with a resolution of 1.8 V/128 ≅ 14.06 mV (LSB). A set of analog switches (transmission gates) is used select the desired voltage according to the VDAC code value. The VDAC code (SET) is one-hot encoded to a 128-bit array such that only one switch can be active. Since the resistor divider has a high output resistance, source-follower analog buffers are used for bias voltages with relatively high current requirements such as the V_{RESET} line that sources the entirety of the pixel matrix leakage current.

Both the IDAC and VDAC incorporate a monitor and override functionality that allows to measure or provide the DAC currents/voltages externally through a set of pads for testing and debugging purposes.

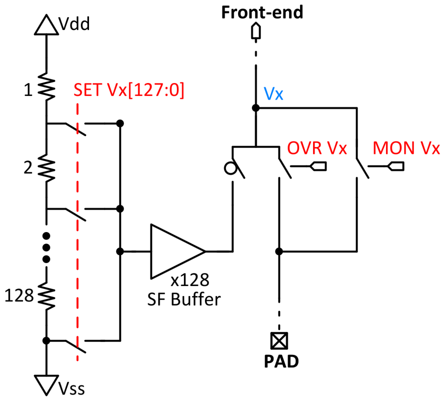


Figure 4.53: Schematic of the analog bias voltage DAC used in TJ-Monopix1.

4.4 TJ-Monopix1 measurement and characterization

In this section, measurement results that demonstrate the performance of TJ-Monopix1 will be presented. Feedback from these results was very valuable in assessing the design choices, the accuracy of simulations and the shortcomings of TJ-Monopix1 that were taken into account by the design of the improved TJ-Monopix2 prototype. However it should be noted that this section does not represent a full and extensive characterization that is not in the scope of this thesis.

4.4.1 Readout system and test setup

The readout system, shown in Fig. 4.54, is based on the MIO3 (multi I/O) readout board [80] and the basil DAQ and testing framework [81]. At the heart of the MIO3 board lies an Enclustra Mercury KX1 FPGA module that is based on a Xilinx Kintex-7 FGPA. Communication with the DAQ software is achieved using a gigabit ethernet adapter included by the KX1 module. Basil is a modular DAQ and testing framework in Python and Verilog. It provides a set of FPGA firmware modules (e.g. SPI, FIFO memory, I/O controllers) and the corresponding software drivers. Additionally it supports the interfacing and control of laboratory instrumentation for automated testing (ATE) purposes.

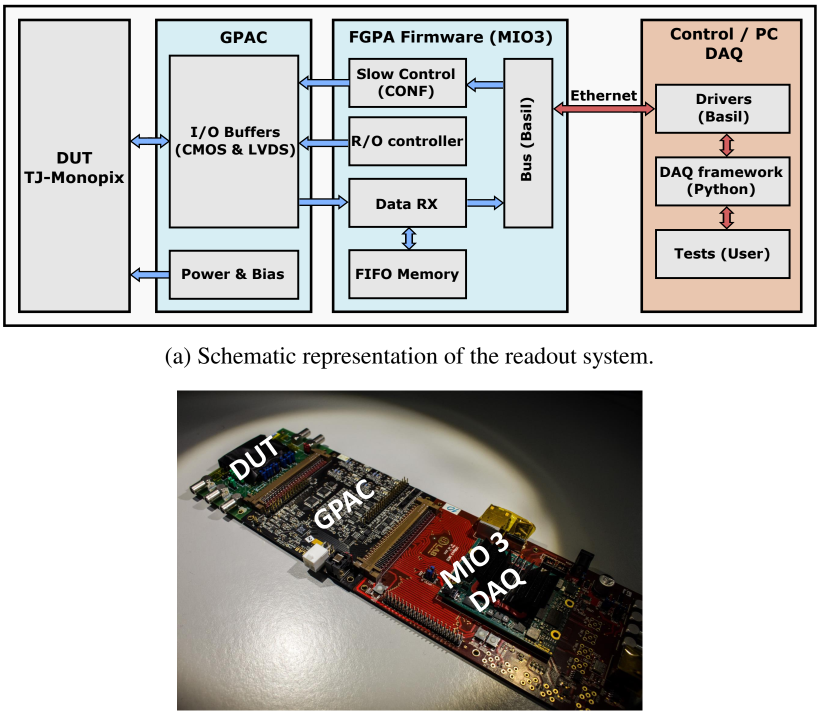
TJ-Monopix1 chips have been wire-bonded on a dedicated carrier board which provides a physical connection interface to the MIO3 board. Between the carrier and MIO3 boards a General Purpose Analog Card (GPAC) is introduced which supports the chip operation and provides, among others, power supply channels, current/voltage bias sources and I/O buffers. A specialized firmware based on basil and adapted to TJ-Monopix1 has been developed for the MIO3 FPGA. In includes an SPI module for the chip configuration, a R/O controller, a data receiver and FIFO memory to store the hit data, clock and pulse signal generators and an ethernet based I/O module to communicate with the DAQ software. A dedicated software library has been developed in Python that enables the configuration and testing of TJ-Monopix1.

Measurement results indicate the correct operation and full functionality of TJ-Monopix1. An example of the chip response to an injection pulse is shown in Fig. 4.55. The signal waveforms are captured at the TJ-Monopix/GPAC interface with the help of an oscilloscope. A single pixel is enabled for injection. The R/O sequence begins after the token signal is received by the FGPA R/O controller and a data packet containing the correct pixel row and column address as well as the hit LE and TE timestamp is received by the DAQ.

4.4.2 Single pixel measurements

4.4.2.1 Analog front-end output

The front-end analog output, provided by the special analog monitoring pixels, has been measured in order to verify the correct pre-amplifier operation, fine tune the bias settings and characterize performance aspects such as the gain and the impact of sensor biasing on the detector capacitance. The front-end pre-amplifier response to an injected charge equal to approximately $750\,e^-$ is shown in Fig. 4.56. The p-well (PW) and p-substrate (PSUB) bias voltage has been set to −6 V, while the HV front-side bias voltage is equal to 10 V. The output waveforms of three front-end variations (PMOS reset, leakage compensation and AC-coupled HV) with FDPW coverage are included in the measurement. As can be observed, the voltage amplitude of the leakage compensation pixel is approximately 2/3 of the PMOS reset pixel amplitude ($\approxeq$ 300 mV), which corresponds to a signal



(a) Schematic representation of the readout system.

(b) Picture of the readout system that includes the DUT (TJ-Monopix1) on the carrier board and the GPAC and MIO3 cards.

Figure 4.54: The readout system used for testing of TJ-Monopix1 chips.

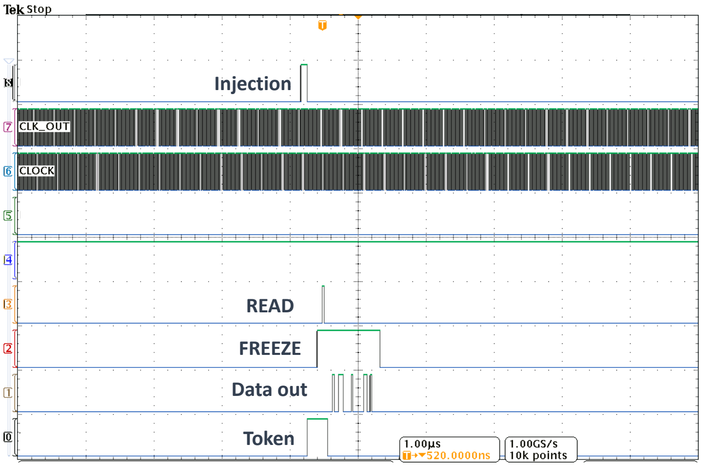


Figure 4.55: Example readout operation of a TJ-Monopix1 chip that is connected to the DAQ system. Hit data are being read-out as a response to an injection pulse.

loss of about 30% due to the extra capacitance at the front-end input node introduced by the leakage compensation circuit. Because in the case of the HV front-end the optimal sensor biasing, as well as its injection capacitance, are different compared to the other front-end variations, a useful comparison of its voltage amplitude cannot be made from this measurement, but will be studied in the following paragraphs.

In order to verify the leakage compensation functionality, the TJ-Monopix chip has been illuminated with a light source that artificially increased the sensor leakage current to approximately 500 nA. Initially, the input reset current I_{RESET} has been set to 100 pA. When the light source was turned on, the PMOS reset analog monitoring pixel stopped responding while the leakage compensation pixel continued to operate normally. To restore the PMOS reset front-end operation, I_{RESET} had to be set above 500 nA as expected.

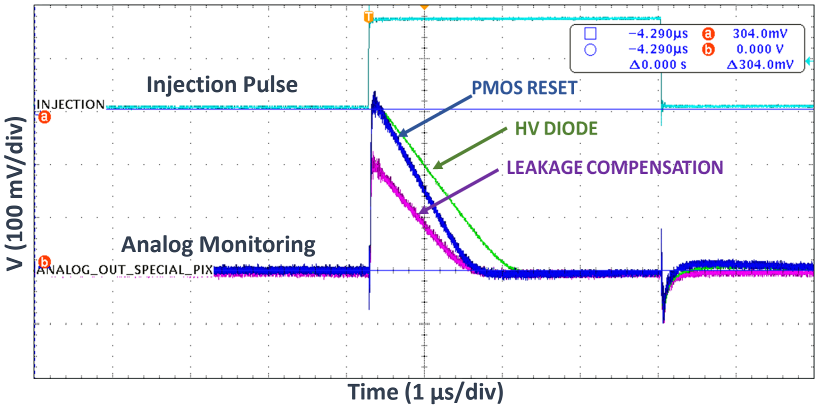


Figure 4.56: Front-end analog output waveforms, measured using the analog monitoring pixels. The amplitude corresponds to charge injection of 700 e^{-} and the measurement includes all three front-end variations with full deep p-well.

The peak output voltage amplitude and gain of the PMOS reset pre-amplifier has been measured for different input charge (Q_s) values up to 1 650 e^{-}. The measurement results as well as a comparison with the simulated values are presented in Fig. 4.57. As expected, the gain increases (due to the non-linear pre-amplifier behavior) up to approximately 0.45 mV/e^{-} and then starts to decrease as the output voltage reaches its saturation point. There is good agreement between the measured and simulated values and the difference is mainly attributed to the accuracy of the transistor models in the extreme case of −6 V backbias.

A systematic comparison of the front-end gain with respect to the sensor biasing potentials is required in order to identify the optimum biasing conditions that minimize the sensor capacitance and maximize performance in terms of signal to noise ratio. In order to accurately convert the output voltage amplitude to gain (mV/e^{-}), the injection capacitance for each front-end type has to be calibrated. Calibration has been performed by measuring the front-end output response for a deposited charge amount, generated by the absorption of known energy photons, and subsequent comparison to the injection pulse voltage that induces the same response. The calibration measurement can either use the analog voltage amplitude of the monitoring pixels or the ToT of regular pixels using the digital readout. Although the ToT resolution is not as high compared to the analog voltage amplitude method, a large number of measurements can be simultaneously obtained from the pixel matrix that

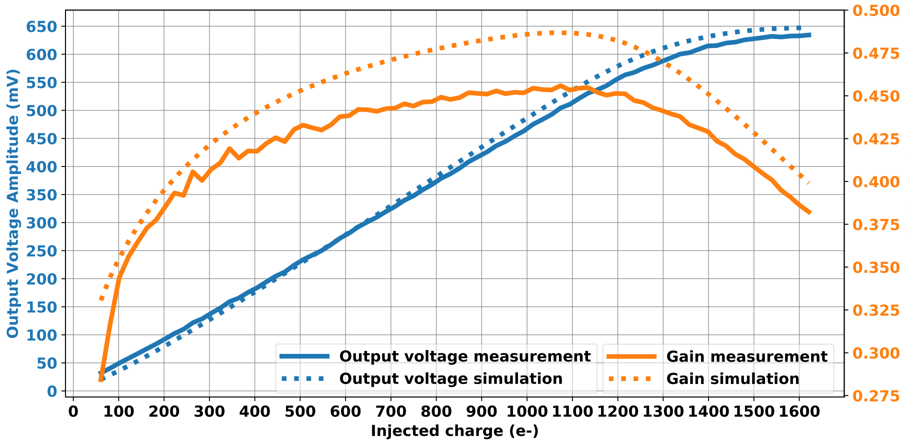


Figure 4.57: Measurement of the PMOS reset pre-amplifier output and gain with respect to the injected charge Q_s. The dotted lines correspond to the simulation results with the same parameters.

significantly increases confidence in terms of statistics. The injection capacitance (C_{inj}) has been calibrated using the 5.9 keV (K_α) X-ray photons that are emitted from an ^{55}Fe radioactive source and generate an average charge amount (at 300 K) equal to approximately 1 616 e^-. In the case of the DC-coupled front-end flavors (PMOS reset and leakage compensation), C_{inj} has been calibrated to approximately 230 aF, which is equal to the design value. In contrast, the injection capacitance of the AC-coupled HV front-end has been calibrated to approximately 400 aF. The value of C_{inj} is higher in this case due to layout differences as a result of the AC-coupling MOM capacitor placed in the area above the collection electrode.

A 2D map of the measured output voltage amplitude and resulting gain in the case of the PMOS reset front-end for an injected charge equal to 700 e^- and varying p-well and p-substrate bias voltages is shown in Fig. 4.58. As mentioned in section 4.1.1, the minimum p-well bias voltage value is limited by the in-pixel NMOS transistor source/drain junction breakdown and is equal to −6 V. Since the p-epitaxial layer is fully depleted, the p-substrate can be biased separately. For p-well voltage values below −0.6 V. the minimum PSUB voltage is approximately −20 V and is limited by the punchthrough effect between the p-well and p-substrate through the n$^-$ layer. If the p-well voltage is higher than −0.6 V, the minimum PSUB voltage is approximately −16 V instead. High currents due to punchtrough are also observed if the p-well bias voltage is higher than the p-substrate bias voltage and therefore these combinations are not allowed.

As expected, the voltage amplitude and gain increase with the sensor reverse bias due to the resulting reduction in the detector capacitance C_D. The influence of the p-well reverse bias to C_D is dominant since it laterally depletes the region between the p-well and the collection electrode n-well. The p-substrate reverse bias mainly controls the vertical field in the sensor bulk area, which is already depleted, and has a less significant effect on the sensor capacitance. The measurement results confirm that the sensor bias combination which results in the highest signal amplitude in the DC-coupled front-end case is: $V_{\text{PW}} = -6$ V and $V_{\text{PSUB}} = -20$ V. However, from a charge collection perspective a higher V_{PSUB} (e.g. −12 V) may be preferred in order to avoid the "flattening" effect it has on the potential landscape since the penalty on C_D is only marginal.

A similar measurement has been performed using the AC-coupled HV front-end for varying

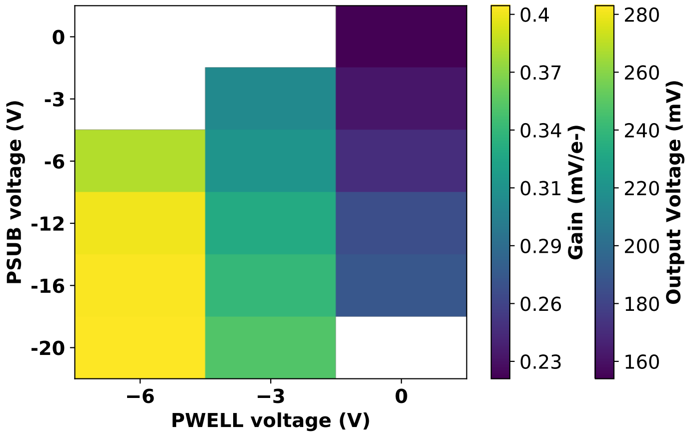


Figure 4.58: 2D map of the measured output voltage amplitude and gain with respect to the p-well and p-substrate bias voltages in the case of the PMOS reset front-end.

p-well/p-substrate and front-side HV bias voltages. In this case, the p-well and p-substrate are connected together for simplicity. Due to the higher injection capacitance, the injected charge is approximately 1 225 e^{-} (for the same injection pulse height). The measurement results are shown in Fig. 4.59. As expected, a high collection electrode bias voltage (HV) fully depletes the area around the collection electrode and results in a low detector capacitance and high gain. The HV voltage can be as high as 50 V before the junction breakdown begins. However, for values higher than 20 V, the output voltage amplitude remains approximately constant (full depletion) and these points are not included in the plot.

By decreasing the p-well and p-substrate voltage, oddly enough, the gain begins to also decrease. The main reason is the influence of backbias (p-well voltage) on the front-end transistors, as discussed in section 4.3.1.1, that can result in a pre-amplifier voltage gain degradation up to 35-40% for $V_{\mathrm{PW}} = -6V$. Apart from the pre-amplifier voltage gain, an effect discussed in [82] that can influence the depletion around the collection electrode may additionally contribute to the gain degradation, but has not yet been conclusively confirmed. According to this effect, by increasing the p-well reverse bias with respect to the collection electrode (which is biased by the HV voltage), the potential valley moves deeper in the substrate and a larger number of field lines end up to the collection electrode. Hence, the local field strength and therefore the potential drop underneath the collection electrode are increased resulting in a displacement of the depletion boundary in this area.

Although in the case of the HV front-end a significant percentage of the signal (up to 50%) is lost due to the AC-coupling MOM structure (section 4.3.1.1), it can be observed that the gain for the optimum bias conditions ($V_{\mathrm{PW}} = V_{\mathrm{PSUB}} = 0\,\mathrm{V}$, $HV > 20\,\mathrm{V}$) is comparable to the DC-coupled PMOS reset front-end. The reason is twofold: First it is related to the higher pre-amplifier voltage gain due to the absence of backbias. It should be noted that this gain increase does not translate to a higher the signal to noise ratio because noise is also amplified by approximately the same amount, as will be confirmed by ENC measurements. Second, the pre-amplifier coupling capacitance C_c, has been enlarged by approximately 40% in the case of the HV front-end in order to compensate to some extent for the reduced input signal. A higher C_c not only increases the gain but also results in a higher SNR because the pre-amplifier high pass filter bandwidth is increased and there is a stronger coupling of the input signal to the gain node GN (see section 4.3.1.1).

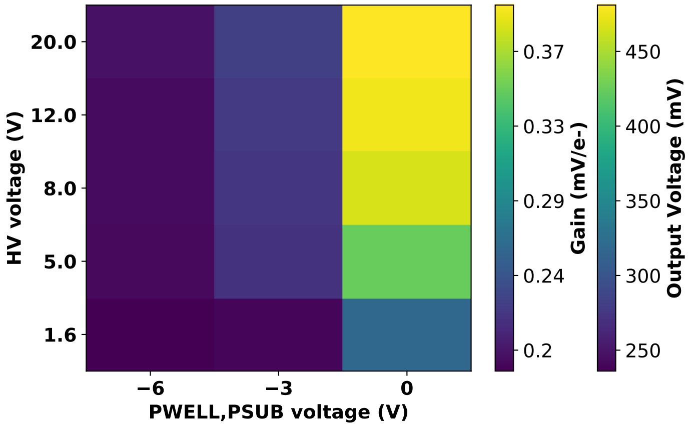


Figure 4.59: 2D map of the measured output voltage amplitude and gain with respect to the p-well/p-substrate and collection electrode (HV) bias voltages in the case of the HV front-end.

4.4.2.2 Timing response and ToT

Precision measurements of the front-end timing response and ToT can be conducted with the help of the HIT-OR feature that provides direct access to the discriminator hit pulse output. The HIT-OR output pulse is sampled by a digital oscilloscope (DSO) and the timing of its leading edge (LE) and trailing edge (TE) are recorded. The measured time-walk curve for injected charge values up to 1 650 e^- is shown in Fig. 4.60. The threshold is set to approximately 300 e^-. The LE time (y-axis) represents the time duration between the injection strobe pulse produced by the FPGA and the arrival of the HIT-OR output pulse and therefore includes the propagation time of the injection and HIT-OR pulse through the chip periphery and pixel column. However only the relative LE time difference for high and low input charge values is important in the calculation of the front-end time-walk and in-time threshold. As can be observed from the plot, the in-time threshold is equal to approximately 340 e^- and therefore the overdrive charge is equal to 40 e^- in agreement with simulation results.

The ToT curve, presented in Fig. 4.61, is calculated by measuring the HIT-OR pulse width and is highly linear due to the PMOS input reset circuit (in contrast to a diode reset) and the pre-amplifier output reset mechanism, discussed in section 4.3.1.1. The higher ToT slope compared to simulation is due to the different front-end settings (I_{THR}) used during this measurement.

4.4.2.3 Radioactive source measurements

The response to an ^{55}Fe radioactive source has been measured both by using the analog monitoring pixels, and the regular pixels with full digital readout capability. During the electron capture decay of ^{55}Fe to ^{55}Mn, the ^{55}Fe source produces two X-ray photos with energies equal to 5.9 keV (K_α) and 6.5 keV (K_α) that result in an average generated amount of charge equal to approx. 1 616 e^- and 1 781 e^- respectively. Apart from the being useful for the calibration of the injection capacitance, the ^{55}Fe spectrum measurement provides valuable information about the detector performance in terms of gain, energy resolution (SNR) and charge sharing.

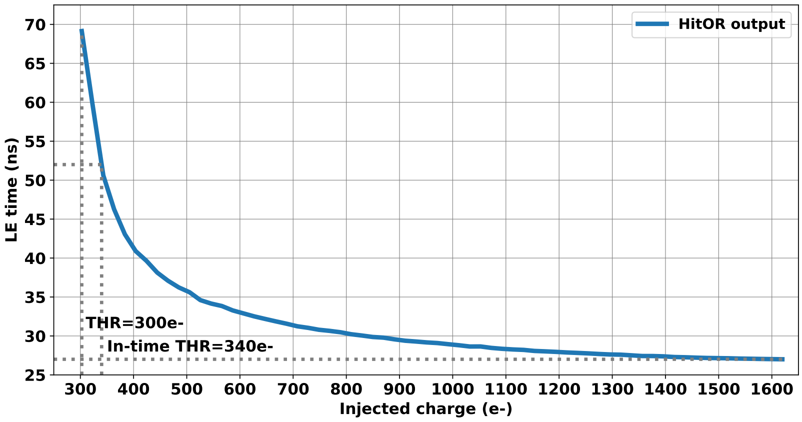


Figure 4.60: Time-walk curve measurement of a PMOS-reset front-end using the HIT-OR output.

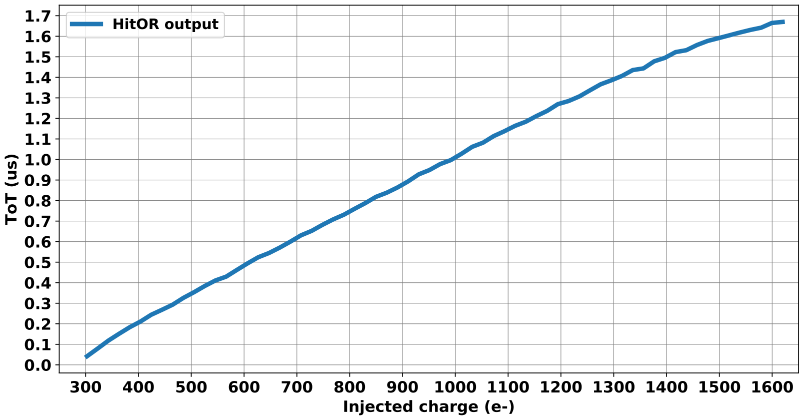


Figure 4.61: ToT curve measurement of a PMOS-reset front-end using the HIT-OR output.

The ^{55}Fe spectrum, measured using the output voltage of a PMOS reset analog monitoring pixel, is shown in Fig. 4.62. During the measurement, the p-well (and p-substrate) bias voltage has been set to -3 V. Both K_α and K_β peaks are visible. The events captured in region below the K_α peak are a result of charge sharing, which as expected is small due to the modified process sensor fast collection by drift. A higher number of events at low voltage amplitudes are produced due to noise.

By measuring the voltage amplitude of the K_α peak, the front-end gain in the corresponding input charge region is calculated equal to approximately 0.27 mV/e^- and is lower compared to Fig. 4.57 due to the lower backbias voltage (-3 V instead of -6 V). The ability to resolve the K_β peak is an indication of high signal to noise ratio. The ENC can be estimated from the standard deviation of the K_α voltage distribution. The voltage fluctuation around the peak is caused by the number fluctuation of generated carriers (Fano noise) and the noise introduced by the detector (sensor and front-end pre-amplifier). By fitting the curve with a Gaussian function[4], and converting to electrons using the gain value (0.27 mV/e^-), the standard deviation is measured equal to approx. $\sigma_{\text{meas}} = 23.3\,e^-$. The

[4] Alternatively, the standard deviation σ can be calculated from the full width at half maximum (FWHM), $FWHM = 55\,e^- \approx 2.355\sigma$ which leads to the same result.

ENC is estimated by subtracting the Fano noise (eq. 3.3) as follows:

$$
\begin{aligned}
\sigma_{\mathrm{e/h}} &= \sqrt{N_{\mathrm{e/h}} \cdot F} = \sqrt{1\,616\,e^{-} \cdot 0.115} \approx 13.63 \implies \\
\mathrm{ENC} &= \sqrt{\sigma_{\mathrm{meas}}^{2} - \sigma_{\mathrm{e/h}}^{2}} \approxeq 18.9\,e^{-}
\end{aligned}
\tag{4.50}
$$

The ENC is in this case higher than the simulated vale of approx. 9 e^{-} for two reasons. First, the detector capacitance is higher than 3 fF due to the moderate reverse bias (3 V) and as a result the SNR is decreased. Second, extra noise is added from the system (test setup) at the analog monitoring pixel output.

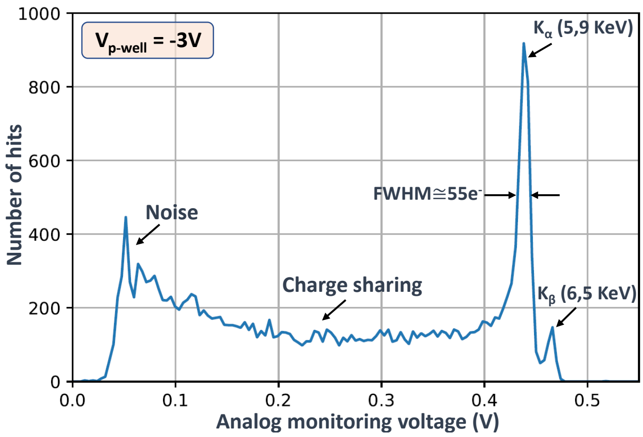
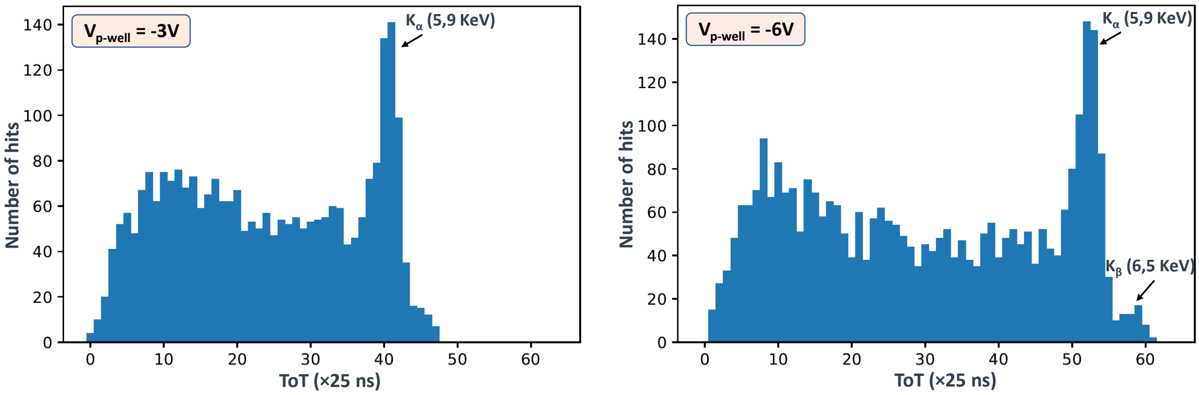


Figure 4.62: Measurement of an ^{55}Fe radioactive source spectrum using the analog output of a PMOS reset front-end.

The ^{55}Fe spectrum measurement using the digital readout is shown in Fig. 4.63 for two different backbias voltages: −3 V and −6 V. The captured events (statistics) correspond to about 10 m of run-time. Instead of voltage amplitude, the input charge is in this case measured as a function of ToT. Because of the limited ToT resolution (compared to the analog voltage measurement), the K_{β} peak is not resolved for a backbias voltage of −3 V. When the reverse bias voltage is increased to −6 V, the gain is increased by about 25% and the ENC is decreased. As a result the K_{β} peak becomes clearly visible.

4.4.3 Injection scan characterization: Noise and threshold dispersion

The noise and threshold dispersion performance is characterized with the help of an injection scan measurement. A specified number of hit injections is performed for each pixel and varying input charge value (injection pulse voltage) in an adequately large range around the threshold. The hit probability for each input charge value Q_s is calculated as the ratio of the number of injections and the number of hits recorded by the DAQ. Therefore, an s-curve is constructed for each pixel and the process continues until all pixels have been measured. For each pixel, the s-curve is fit with a CDF



(a) (b)

Figure 4.63: Measurement of an ^{55}Fe radioactive source spectrum by a PMOS reset pixel using the digital readout (ToT). The p-well bias voltage is set to a) −3 V and b) −6 V.

function (eq. 4.30) in order to calculate the threshold and ENC. From these values, the threshold and ENC distribution of the pixel matrix can be obtained. The mean threshold and ENC values, as well as their dispersion (standard deviation) is subsequently calculated by fitting each distribution with a Gaussian function.

The analysis is performed separately for each flavor because the pixels are dissimilar and their response and characteristics are different. Furthermore, the pixels with full deep p-well coverage (bottom half of each column) are also analyzed separately from the pixels with reduced deep p-well coverage (top half of each column) because the deep p-well geometry near the collection electrode affects the sensor capacitance and as a result the pixel performance.

Fig. 4.64 shows an s-curve histogram, constructed by an injection scan measurement of the full PMOS reset flavor. It can be viewed as a superposition of the s-curves of all pixels of the specific sector and is used to identify possible issues (e.g. more hits than the number of injections caused by cross-coupling) while giving an indication about the threshold mean, dispersion and ENC performance. During this measurement, the p-well and p-substrate have been biased at −6 V and −12 V respectively in order to maximize performance. A very low noise hit rate of about $4 \cdot 10^{-8}$ /25ns per pixel has been achieved by intentionally masking only 34 pixels of the whole flavor. Due to the projection masking scheme, 156 additional "ghost" pixels are also masked. Nevertheless, the number of masked pixels (190) is smaller than 1% of the total 25088 pixels of the PMOS reset sector.

The corresponding threshold distribution, derived from fitting each s-curve, is shown in Fig. 4.65. The mean threshold of the FDPW sector is approx. 255 e^- and the threshold dispersion is equal to approx. 30 e^-. In the case of the RDPW sector, the threshold is higher by approx. 45 e^-. The reason is that the deep p-well removal causes a slight reduction of the depletion around the collection electrode and results in an increase of the sensor capacitance C_D. Therefore, the charge to voltage conversion gain at the front-end input (Q_s/C_D) is lower and as a result, for the same input charge a lower voltage amplitude is induced at the front-end input. The threshold dispersion is also slightly higher (by approx. 4 e^-) due to the overall lower gain (eq. 4.28). The measured threshold dispersion is close to the simulated value of 25 e^-. The difference (about 5 e^-) is due to other contributions such as the analog power rail voltage drop and mismatch effects of the biasing DAC that have not been

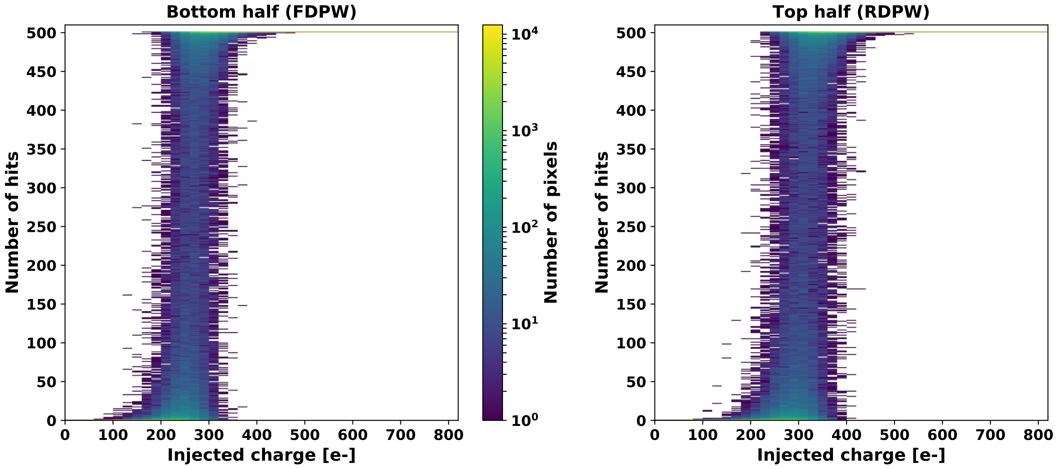


Figure 4.64: S-curve histogram of the full PMOS reset flavor. The FDPW and RDPW parts are plotted separately.

included in the simulation.

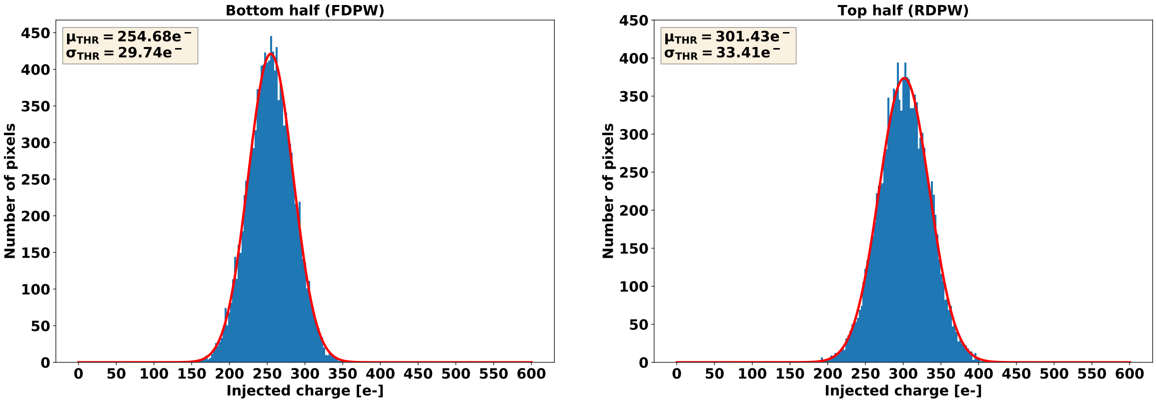


Figure 4.65: Threshold distribution of the bottom (FDPW) and top (RDPW) part of the PMOS reset flavor. A gaussian fit function is applied to calculate the threshold mean (μ) and dispersion (σ).

Th ENC distribution is presented in Fig. 4.66. The mean ENC value is approximately equal to 8 e^- (1 e^- higher for the RDPW pixels due to the higher C_D) with a dispersion of approx. 0.9 e^- and is in close agreement with the simulated value (8.9 e^-). However, the ENC distribution is not purely Gaussian and there is a long-tail of pixels with higher noise than described by a normal distribution. The long-tail is formed as a result of RTS noise from the pre-amplifier transistors and especially the critical gain stage input transistor M_1 due to its small area as described in section 4.3.1.1. It is important to reduce this effect since the ENC long-tail increases the noise rate and limits the operating threshold, especially after irradiation. Furthermore, because of the RTS noise, an increased number of pixels with high noise is required to be masked, which can lead to a high percentage of total masked pixels due to the exponential "ghost" pixel number as a result of the projection masking scheme. It has been proven from measurements of the mini-Malta chip [67] that RTS noise and therefore the ENC long-tail is significantly reduced by enlarging the area of transistor M_1.

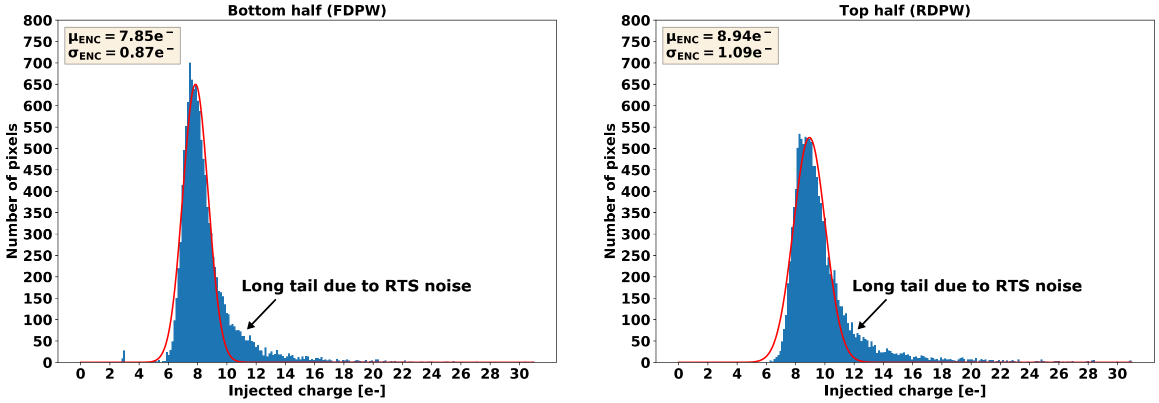
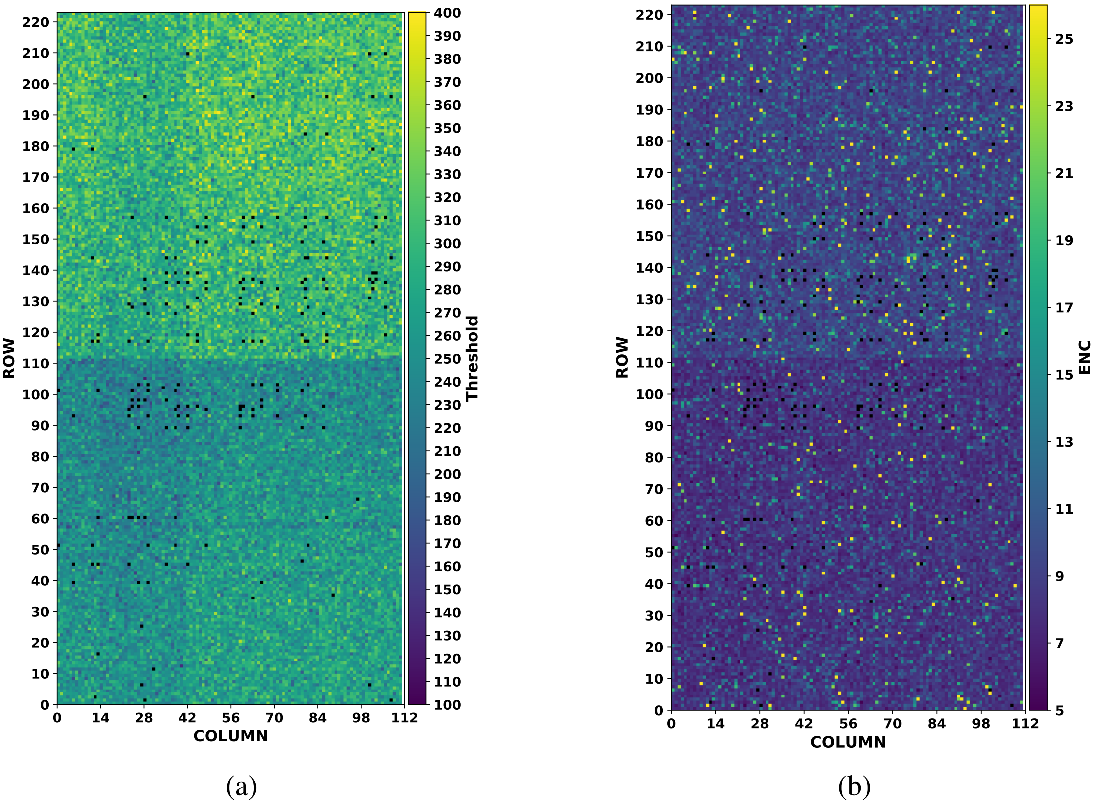


Figure 4.66: ENC distribution of the bottom (FDPW) and top (RDPW) part of the PMOS reset flavor. The long-tail in the distribution is attributed to pre-amplifier RTS noise.

In order to identify possible systematic variations across the matrix area, caused for example due to the biasing scheme or voltage drop gradients, a 2D map of the threshold and noise of the PMOS reset flavor pixels has been created and is shown in Fig. 4.67. The difference between the bottom half (FDPW) and hop half (RDPW) pixels is clearly visible. No strong systematic effects are observed apart from small threshold variations that correlate with the biasing groups (14 columns per group, section 4.3.2.3). To study this effect more carefully, a plot of the average threshold per column is presented in Fig. 4.68. Although the threshold is lower for the second and third group (columns 14 to 42), there is not enough evidence for a conclusive confirmation. Therefore, another chip sample has been measured, as shown in Fig. 4.69. In this case there is a clear indication of systematic threshold variation due to the transistor mismatch of the biasing DAC mirrors and the V_{CASN} generator circuit. A solution, implemented in TJ-Monopix2, is to include more columns in each biasing group and thus increase the effective transistor area (trading-off with the voltage drop compensation accuracy).

The injection scan characterization results in the case of the HV flavor are shown in Fig. 4.70 (threshold distribution) and Fig. 4.71 (ENC distribution). The optimum sensor biasing combination has been applied by setting the p-well and p-substrate voltage to 0 V and the HV front-side bias to 50 V. Due to the high collection electrode front-side bias voltage, the area around the collection electrode is fully depleted for both FDPW and RDPW pixels and therefore there is practically no difference in their performance in terms of noise and threshold dispersion. The main drawback of the HV front-end is, as mentioned, the high signal loss of approx. 50% at the input due to the parasitic capacitance introduced by the MOM AC-coupling structure. Therefore the ENC ($\approxeq 15-16\,e^-$) is approximately doubled compared to the PMOS reset flavor. As a result, in order to maintain the same noise hit rate, the threshold had to be increased to approx. 400 e^-. In contrast to noise, the threshold dispersion is only about 23% higher than the PMOS reset flavor because of the higher pre-amplifier gain (due to the absence of backbias and larger C_c) that compensates for the signal loss at the input.

The measurement results in the case of the leakage compensation flavor are shown in Fig. 4.72 (threshold distribution) and Fig. 4.73 (ENC distribution). The ENC ($\approxeq 12\,e^-$) is approx. 30% higher compared to the PMOS reset flavor due the signal loss (by the same amount) at the input as a result of the additional capacitance introduced by the leakage compensation circuit. The main drawback of this implementation is the significantly higher threshold dispersion as a result of the reduced gain and the



(a)

(b)

Figure 4.67: 2D map of the PMOS reset flavor a) threshold and b) ENC. The black dots represent masked pixels.

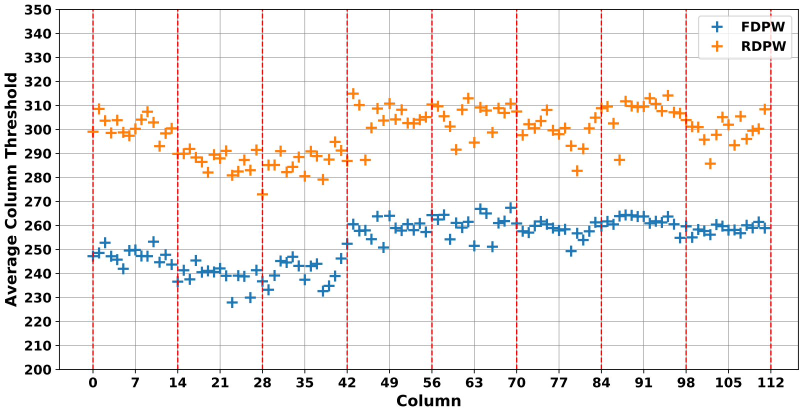


Figure 4.68: Plot of the average threshold per pixel column of the PMOS reset flavor.

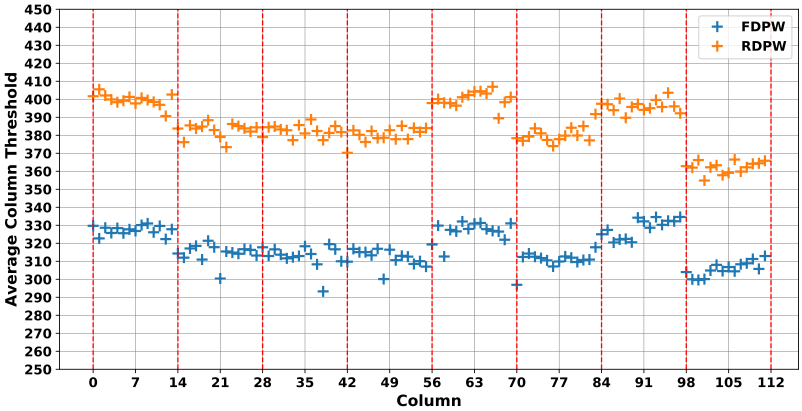


Figure 4.69: Plot of the average threshold per pixel column for a different chip sample. The threshold is higher due to different front-end settings. Systematic variations for different biasing groups are visible.

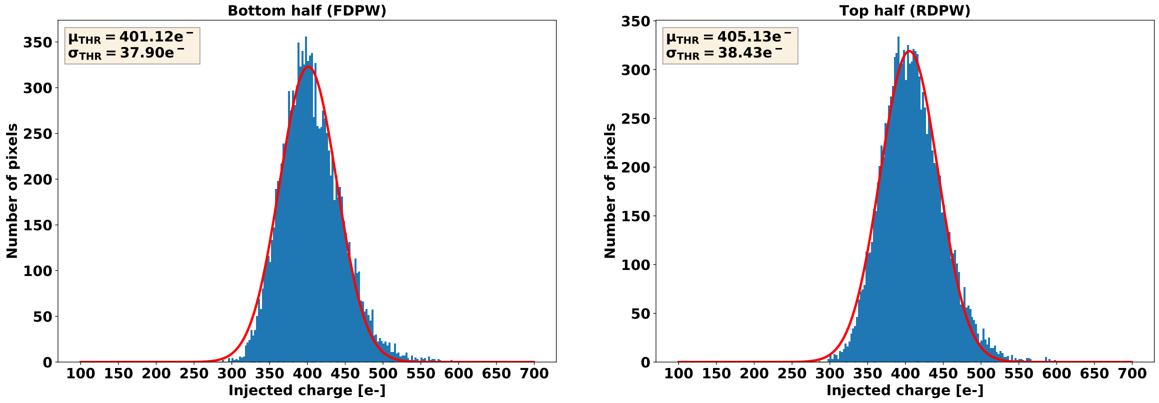


Figure 4.70: Threshold distribution of the HV flavor. The bottom (FDPW) and top (RDPW) part are analyzed separately.

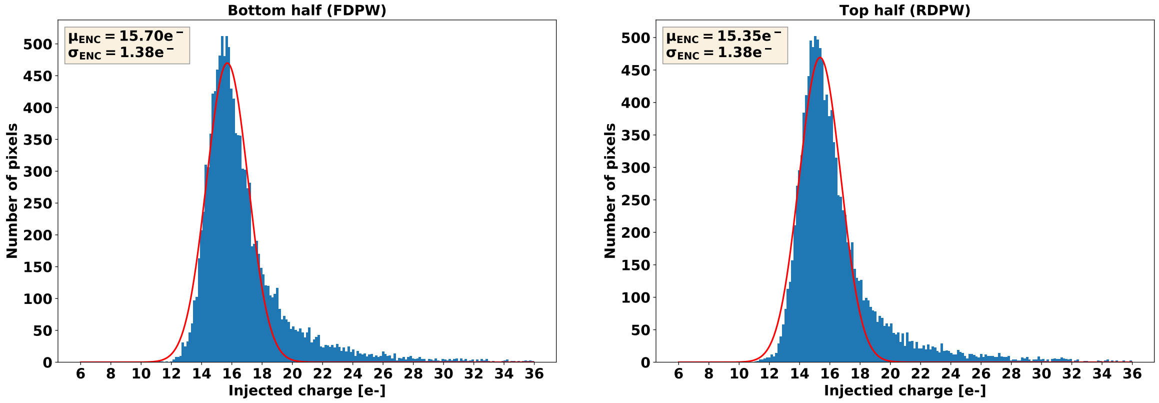


Figure 4.71: ENC distribution of the HV flavor. The bottom (FDPW) and top (RDPW) part are analyzed separately.

additional leakage compensation transistor mismatch. Therefore, the threshold had to be set higher than 400 e^- in order to maintain the same noise hit rate and percentage of masked pixels.

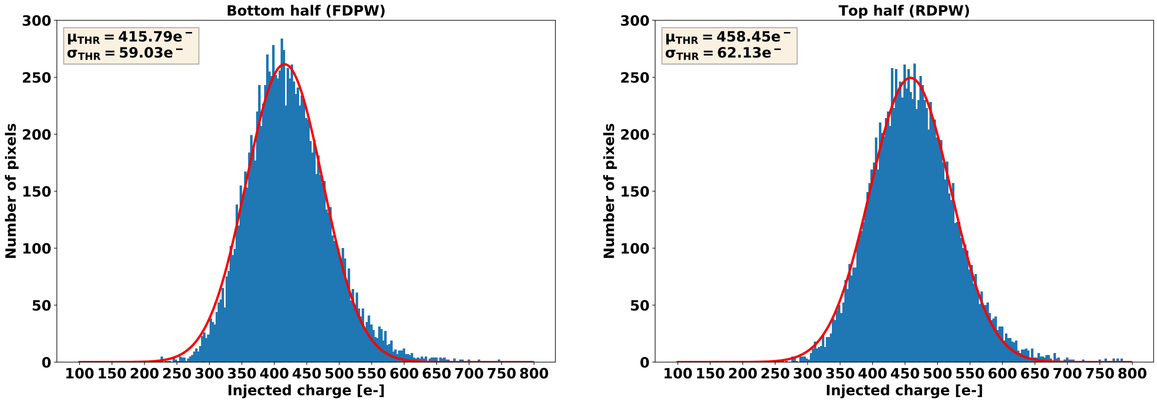


Figure 4.72: Threshold distribution of the leakage compensation flavor. The bottom (FDPW) and top (RDPW) part are analyzed separately.

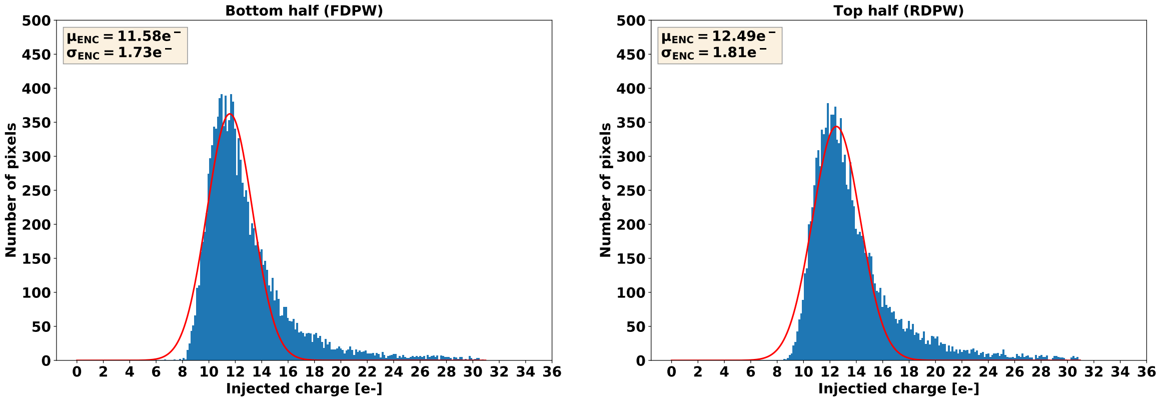


Figure 4.73: ENC distribution of the leakage compensation flavor. The bottom (FDPW) and top (RDPW) part are analyzed separately.

4.4.4 Performance after irradiation

In order to measure the detection efficiency after irradiation to high NIEL fluences, in the order of the ATLAS ITk L4 specification (Table. 2.1), a number of TJ-Monopix1 chip samples have been irradiated with neutrons at the TRIGA Mark II research reactor [69] up to 10^{15} n_{eq}/cm^2 NIEL. It is estimated that due to background radiation, a TID dose of approx. 1 Mrad has been additionally received.

Prior to the detection efficiency measurement at a test-beam facility, characterization of the irradiated chip samples by injection scan and radioactive source measurements is necessary in order to verify the chip functionality and evaluate the irradiation impact on performance aspects such as the gain, noise and threshold. During measurement, TJ-Monopix1 has been cooled down to the normal detector operating temperature of about -30 °C that helps to reduce the sensor leakage current from over 100 pA (room temperature) to less than 10 pA (see section 4.3.1.1).

No issues have been observed with the irradiated chip samples, which remain fully functional. The response to an ^{55}Fe radioactive source before and after irradiation is shown in Fig. 4.74. The measurement has been performed using regular pixels (digital readout) of the HV flavor[5]. In the irradiated case, the K_β peak is no longer visible, which is an indication of a reduced signal to noise ratio. Furthermore, the K_α peak is shifted to lower ToT values. The reason is twofold: first, the gain is reduced due to damage to the pre-amplifier transistors from the received 1 Mrad TID and second the front-end settings (including the feedback current) have been modified after irradiation to increase the threshold as a result of increased noise. Nevertheless the ^{55}Fe spectrum response remains adequate after irradiation.

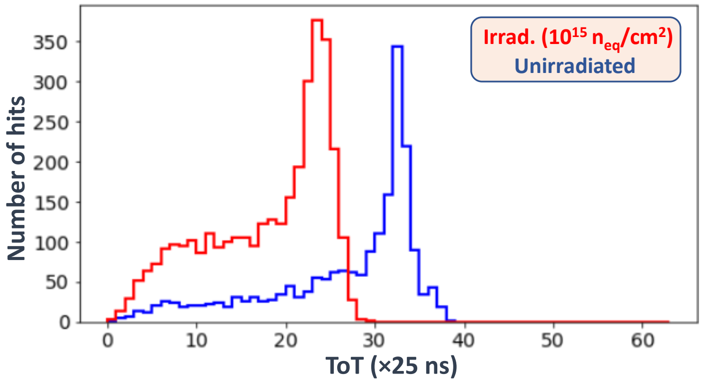


Figure 4.74: Comparison of the response to an ^{55}Fe radioactive source (spectrum) in the case of an unirradiated and an irradiated TJ-Monopix chip sample. The measurement has been performed using the HV flavor.

The injection scan characterization results are shown in Fig. 4.75 (threshold distribution) and Fig. 4.76 (ENC distribution). The mean threshold is higher compared to the un-irradiated case by about 100 e^-, mainly due to the pre-amplifier voltage gain reduction as a result of TID. Furthermore, there is a significant increase of the threshold dispersion in the order of 50%, mainly due to TID damage. Although the mean ENC value (to the extent that the Gaussian fit is representative) is only slightly higher (by approx. 1 e^-) in comparison to the un-irradiated case, the ENC dispersion and RTS noise (long-tail) are considerably increased. The high threshold dispersion and ENC long-tail are important because they prohibit low threshold operation $< 300\,e^-$ and as a result affect the detection efficiency, especially after irradiation when a portion of the generated charge is lost due to trapping.

In order to measure the radiation tolerance to TID effects, TJ-Monopix chip samples have been irradiated with X-rays at the University of Bonn up to a total dose of 100 Mrad [34]. The X-ray tube that contains a tungsten cathode was biased at 40 kV and a thin aluminum filter was used to harden its spectrum. The chip temperature was kept constant at approx. −3 °C and the dose rate was equal to 0.6 Mrad/h. The pre-amplifier analog output as a response to a constant charge injection has been measured at specific intervals directly after the accumulation of the desired TID dose and no annealing was performed.

The normalized gain as a function of TID in the case of the PMOS reset front-end is shown in Fig. 4.77. A large gain degradation, up to 80% can be observed in the region between approx. 0.5 – 10 Mrad. For doses higher than 10 Mrad the gain starts to gradually recover and at 100 Mrad it is about

[5] The irradiated chip characterization focuses on the HV flavor because it has been used in the detection efficiency measurement (see section 4.4.5).

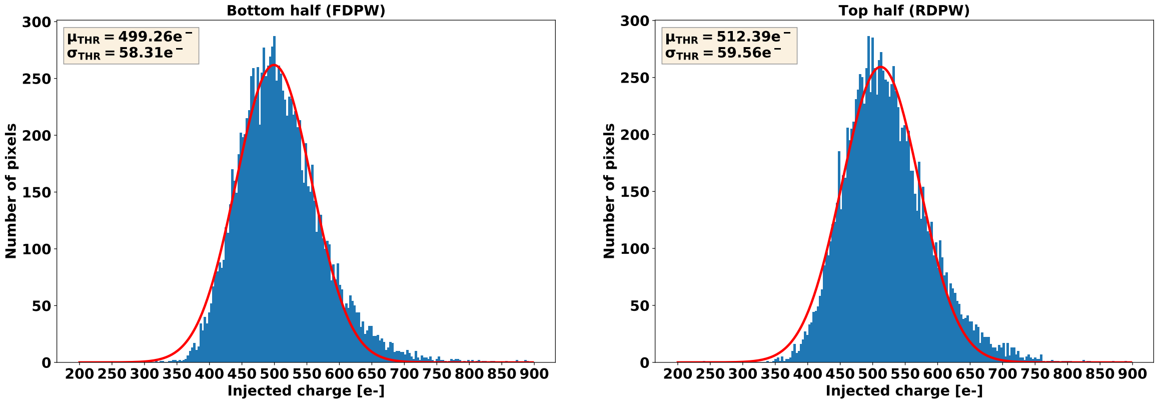


Figure 4.75: Threshold distribution of the HV flavor after irradiation to $10^{15}\,n_{eq}/\mathrm{cm}^2$ NIEL and 1 Mrad background TID. The bottom (FDPW) and top (RDPW) part are analyzed separately.

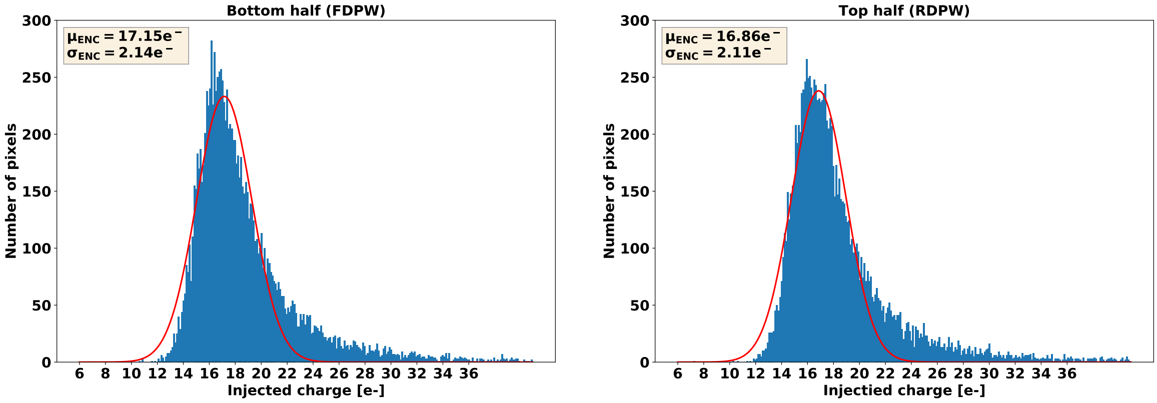


Figure 4.76: ENC distribution of the HV flavor after irradiation to $10^{15}\,n_{eq}/\mathrm{cm}^2$ NIEL and 1 Mrad background TID. The bottom (FDPW) and top (RDPW) part are analyzed separately.

20% lower compared to 0 Mrad. The recovery is due to negative charge accumulation at interface traps and correlates with the NMOS transistor threshold shift measurement (Fig. 4.38). The gain degradation is thought to be mainly caused by the PMOS reset transistor M_{reset} and the pre-amplifier gain stage input transistor M_1. M_{reset} can be sensitive to TID due to RINCE (section 3.1.2.6) because it is narrow (is designed with minimum gate width to reduce its capacitance) and its current (during reset) is very small, usually in the order of tens of pA. This hypothesis is supported by the HV flavor (which uses a diode reset) measurement results, shown in Fig. 4.78. As observed, no significant gain degradation is observed for doses up to 0.8 Mrad. However, due to time constraints a measurement for higher doses was not possible. M_1 is an NMOS transistor that has been designed with minimum length (short channel), and is therefore sensitive to TID due to RISCE (section 3.1.2.6). X-ray irradiation measurements of the mini-Malta chip, presented in [72] demonstrate the improvement in radiation hardness with respect to TID by enlarging transistor M_1.

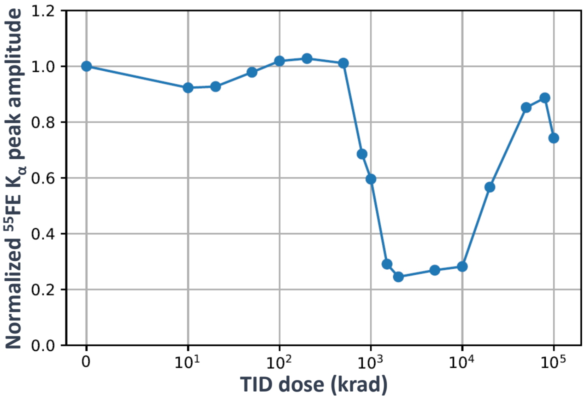


Figure 4.77: Normalized gain of the PMOS reset front-end as a function of TID measured using the pre-amplifier output voltage [34].

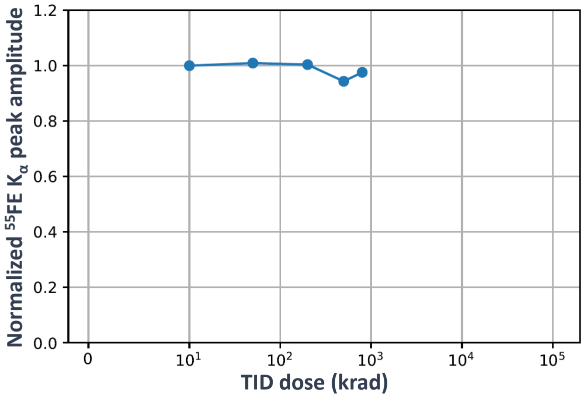


Figure 4.78: Normalized gain of the HV front-end as a function of TID measured using the pre-amplifier output voltage [34].

4.4.5 Detection efficiency measurement

The detection efficiency of unirradiated as well as irradiated (to $10^{15}\, n_{eq}/\text{cm}^2$) TJ-Monopix1 chip samples was measured with a 2.5 GeV electron beam extracted from the electron stretcher accelerator (ELSA) [83, 84] at the University of Bonn. The test setup includes the TJ-Monopix1 readout system (GPAC+MIO3) and an EUDAQ-type [85] reference telescope array that is composed of 6 MIMOSA26 [86] planes, used to measure the incident particle tracks, and 1 FE-I4 [28] plane for timing measurements. Irradiated chip samples have been placed in a styrofoam cooling box capable of reaching temperatures down to −30 °C with the use of solid CO_2 (dry ice). The analysis of test beam data in order to calculate the detection efficiency of TJ-Monopix1 was carried out using a Python based software framework [87].

The measured efficiency map in the case of an unirradiated chip with 25 µm p-epitaxial layer and continuous n$^-$ implantation (without process modification enhancements) is shown in Fig. 4.79. These results correspond to the HV flavor and a threshold of approx. $350\, e^-$. The mean efficiency of the RDPW part is equal to 97.1%, while in the case of the FDPW part it drops to 93.7% due to the weaker lateral field (see section 4.1.3). In the case of the PMOS reset flavor, the corresponding efficiency values are equal to 95.9% for the RDPW part and 93.5% for the FDPW part.

In order to investigate the reason behind the efficiency drop below 100% and gain more insight on how it is influenced by the pixel layout, the in-pixel efficiency map of a 2 × 2 pixel group is shown in Fig. 4.80. The minimum spatial resolution depends on the telescope resolution ($\approxeq$ 5 µm) and the amount of particle scattering, which is inversely proportional to the beam energy. It can be observed that there are low efficiency regions localized in the highlighted corners of every double column that correlate with areas with high active area density due to the placement of large area transistors that operate as decoupling capacitors (MOSCAP). The active area mask specifies the removal of the thick Shallow Trench Isolation (STI) oxide. Because this process is performed before the n$^-$ layer implantation, the differences in the oxide thickness cause local variations of the implantation depth that reduce the lateral electric field in these regions. Additionally, in the case of the RDPW pixels, the asymmetry of the deep p-well removal geometry results in a stronger filed at the double column center (digital area) compared to the left and right sides (analog area).

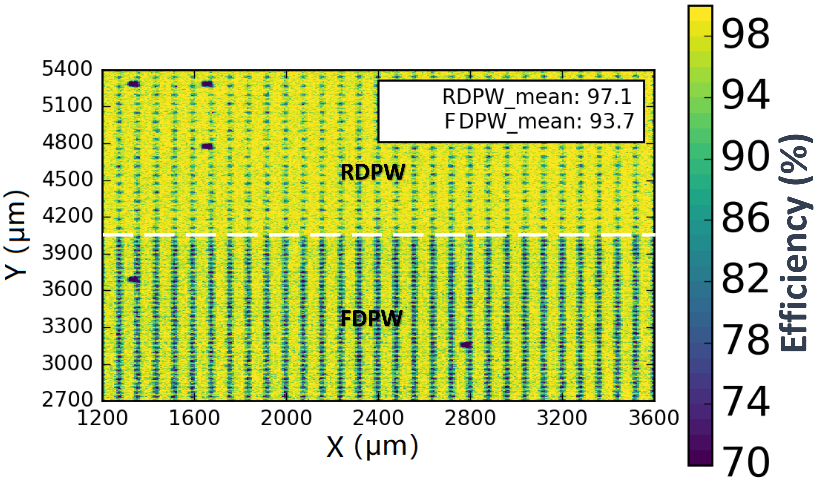


Figure 4.79: Detection efficiency map of an unirradiated TJ-Monopix1 chip with 25 µm p-epitaxial layer and continuous n$^-$ implantation. The measurement was carried out using the HV flavor pixels [33, 34, 88].

The efficiency measurement of irradiated samples was performed using the HV flavor due to the possibility to bias the sensor at a higher voltage (up to 50 V) compared to the PMOS reset flavor. The minimum threshold that could be achieved in this case was equal to approx. 570 e^-. The measured efficiency map of an irradiated chip sample with 25 µm p-epitaxial layer and continuous n$^-$ implantation is shown in Fig. 4.81. In the case of the RDPW part the efficiency drops to 69.4% while in the case of the FDPW part it is reduced to 50.7%. There are two main reasons for the significant decrease in efficiency after irradiation. The first has to do with the sensor charge collection properties. As explained in section 4.1.3, charge that flows in the region under the deep p-well between two pixels is trapped due to the low lateral field. Second, the front-end threshold is too high compared to the generated charge MPV (approx. 1 600 e^-), especially after taking into account the signal reduction due to trapping.

These results indicate that to obtain full efficiency after irradiation, both the sensor and the pixel design has to be improved. The deep p-well removal was a step in the right direction since the higher efficiency of the RDPW pixels helped to identify the low lateral field issue and led to the

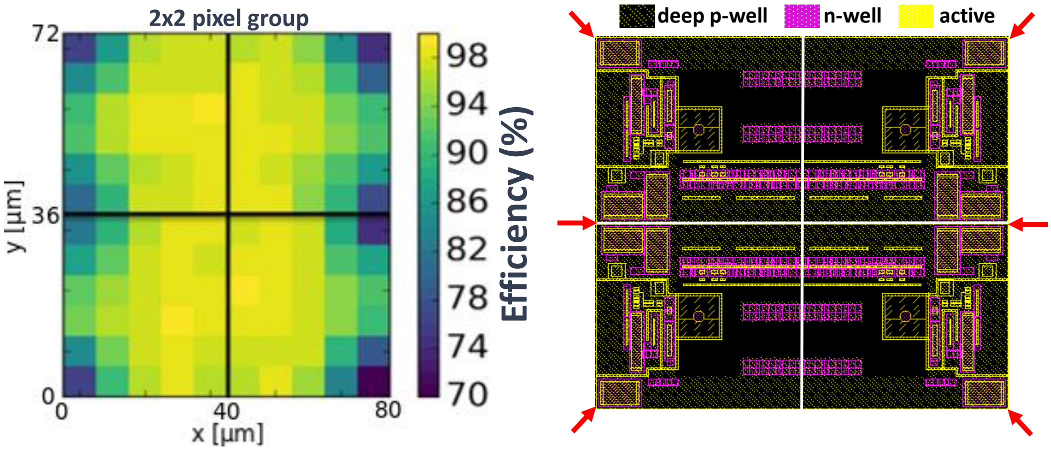


Figure 4.80: In-pixel efficiency map of a 2 × 2 pixel group. The low efficiency regions correlate with high active area density [33, 34, 88].

development of the process modification enhancements. Since they require only a simple mask change, TJ-Monopix1 chips have been produced with the n¯ layer gap and extra deep p-well modifications and were irradiated to 10^{15} n_{eq}/cm^2 NIEL. The detection efficiency of those samples, measured with a 5 GeV electron beam at the DESY II test-beam facility [89], is increased to 85% and is in agreement with the efficiency results of TJ-Malta1 chips in the same process [72].

Although the effectiveness of the process modification enhancements is demonstrated, the ratio of the generated charge to threshold is still not high enough to achieve full efficiency. In order to increase the generated charge (and signal to noise ratio), TJ-Monopix1 chips have been manufactured on a 300 µm thick high-resistivity Czochralski (Cz) substrate (see section 4.1.4). Because after irradiation the p-substrate bias voltage can be as low as −20 V a depletion depth higher than 25 µm (which is the epitaxial layer thickness) is achieved. Furthermore, in the case of the HV flavor, a total reverse bias of more that 50 V can be applied to the sensor which results in an even larger depleted volume and therefore higher signal. The detection efficiency of TJ-Monopix1 chip samples with process modification enhancements, produced on a Cz substrate has been measured equal to approx. 97%. In order to achieve an even higher performance, improvements to the front-end and pixel design have been implemented in the TJ-Monopix2 chip presented in the following section.

4.5 Design of the TJ-Monopix2 next generation full scale prototype

In order to achieve full efficiency after irradiation to $1 \cdot 10^{15} - 2 \cdot 10^{15}$ n_{eq}/cm^2 for both Cz and epitaxial substrate materials, improvement of the TJ-Monopix1 design in two key areas is required. First, a lower operating threshold, below approx. 200 e^- is necessary in order to detect events with low charge due to trapping or charge sharing, especially in the case of a thin (25 – 30 µm) p-epitaxial sensor material. Second, in addition to the modified process enhancements, the pixel layout has to be optimized in order to increase charge collection efficiency, especially in the pixel corners.

These requirements have led to the design of the next generation small collection electrode DMAPS prototype in TowerJazz 180 nm, called TJ-Monopix2. Apart from addressing the shortcomings of its predecessor, TJ-Monopix2 aims to demonstrate the feasibility of a full size chip (2 × 2 cm^2), as

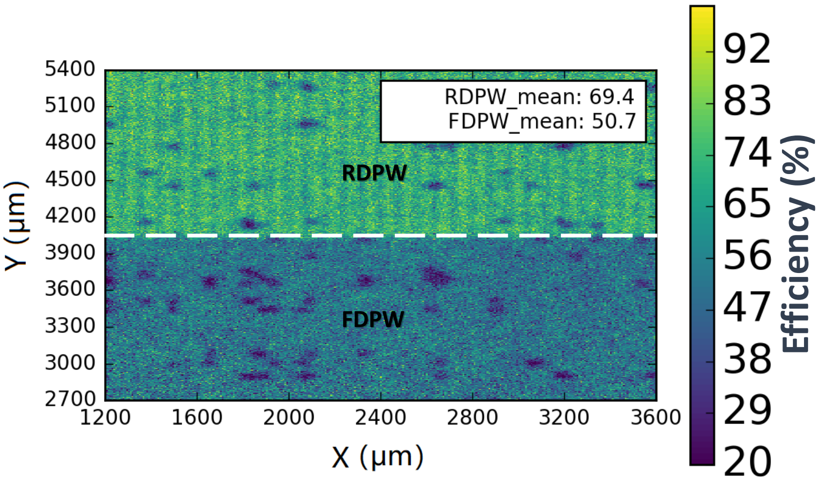


Figure 4.81: Detection efficiency map of a TJ-Monopix1 chip with 25 µm p-epitaxial layer and continuous n⁻ implantation that has been irradiated to 10^{15} n_{eq}/cm^2 NIEL. The measurement was carried out using the HV flavor pixels [33, 34, 88].

required by future detectors such as the ATLAS ITk. TJ-Monopix2 incorporates an improved front-end circuit that reduces ENC by approx. 40% and threshold dispersion by approx. 80-90% (depending on the circuit variation). Furthermore, in-pixel threshold tuning has been integrated in order to achieve a more uniform threshold distribution across the pixel matrix, particularly after irradiation. Pixel masking is also improved by employing individual in-pixel configuration memory that eliminates the issue of unintentionally masked ghost pixels and allows for a more efficient configuration of the pixel matrix and reduced noise hit rate. As a result of these improvements, the operating threshold of TJ-Monopix2 is expected to be $\leq 100\,e^-$, three times lower compared to TJ-Monopix1 (before irradiation).

In addition to the modifications of the front-end circuit, the design effort was focused on optimizing the pixel layout in order to reduce its size. As discussed in [66], a small pixel size is crucial in order to achieve a high lateral electric field and short drift path, hence fast charge collection across the full sensitive area. A pixel size in the order of $30 \times 30\,\mu\text{m}^2$ or less is estimated to be sufficient in order to maintain high charge collection efficiency after irradiation, while for even smaller pixel sizes below $20 \times 20\,\mu\text{m}^2$ sub-ns current pulse peaking times have been simulated. Reducing the pixel size while incorporating the full digital readout and even more functionality, such as the threshold tuning circuit, has been a challenging task that pushed the electronics design to the technology density limit and required further modifications such as single ended data transmission in order to reduce the column-bus width. As a result of these efforts, the TJ-Monopix2 pixel size has been decreased to $33.04 \times 33.04\,\mu\text{m}^2$, which is approx. 25% smaller compared to TJ-Monopix1.

The layout and floorplan of TJ-Monopix2 is illustrated in Fig. 4.82. The pixel matrix consists of 512×512 pixels and has an active area of approx. $286\,\text{mm}^2$. It is split in four sectors (flavors) that contain different variations of the front-end circuit. The first two flavors are composed of DC-coupled pixels with diode input reset and their difference lies in the pre-amplifier circuit design. The second flavor (cascode FE) features an extra cascode transistor that increases the pre-amplifier gain and results in 50% reduction of the threshold dispersion compared to the first (normal FE) flavor. The remaining two sectors contain AC-coupled pixels with front-side HV biasing and also incorporate

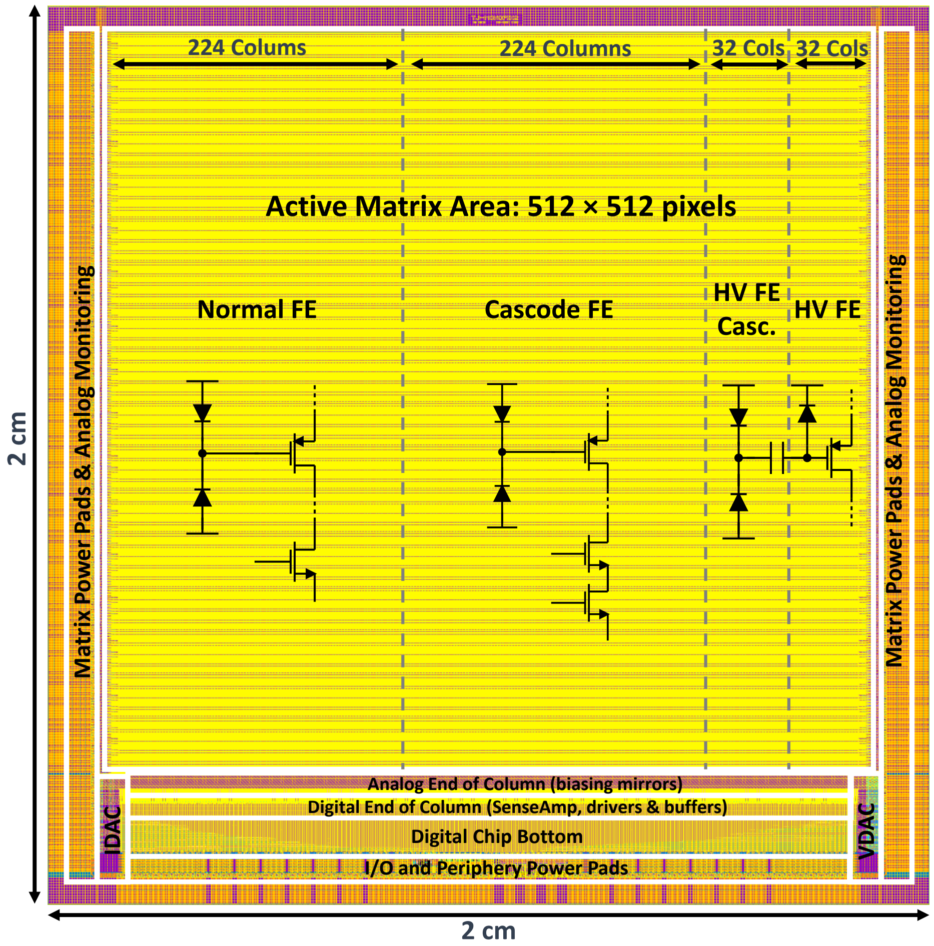

224 Colums
224 Columns
32 Cols
32 Cols
Active Matrix Area: 512 × 512 pixels
Normal FE
Cascode FE
HV FE Casc.
HV FE
Matrix Power Pads & Analog Monitoring
Matrix Power Pads & Analog Monitoring
Analog End of Column (biasing mirrors)
Digital End of Column (SenseAmp, drivers & buffers)
Digital Chip Bottom
I/O and Periphery Power Pads
IDAC
VDAC
2 cm
2 cm

Figure 4.82: Layout of the TJ-Monopix2 prototype. The chip size is equal to $2 \times 2\,\text{cm}^2$ and the active matrix is composed of 512×512 pixels split in four sectors.

the aforementioned pre-amplifier variations. All pixels are designed with a reduced deep p-well geometry (RDPW) due to its superior charge collection properties compared to full deep p-well coverage (FDPW), as demonstrated by the measurement results of TJ-Monopix1. The BCID bus width has been increased to 7-bits due to the higher gain and ToT slope of the improved front-end design. Therefore, because the BCID counter rollover time is increased to $128 \cdot 25\,\text{ns} = 3.2\,\mu\text{s}$, the ToT can be as high as approximately $2\,\mu\text{s}$ without affecting the readout architecture efficiency (in contrast to $400\,\mu\text{s}$ for 6-bit BCID).

Due to the large matrix area and small pixel size, additional design challenges arise that are related to signal transmission across the column and the robustness of power delivery. The large column height ($\cong 17\,\text{mm}$) and aggressive column-bus routing (minimum line width and spacing) as a means to reduce the pixel size, results in a significant signal transmission delay due to the RC low pass filtering effect of the long metal wires. Therefore, in order to compensate for the BCID propagation delay, a special circuit has been designed that adds a variable delay to the hit pulse across the column that matches that of the BCID signal. Additionally, a fast-token logic has been implemented that takes advantage of pixel grouping in a 2 × 2 core and reduces the token propagation delay by approximately four times. The column data-bus bandwidth and signal integrity is particularly affected due to single-ended transmission that results in two times lower signal amplitude and increased crosstalk interference from neighboring lines. A current-mode data-bus readout scheme has been devised in an effort to increase the data transmission speed and reduce the error probability due to noise and mismatch effects. The resulting maximum column readout bandwidth of 10 MHz is still sufficient for hit rates up to approximately 250 MHz/cm^2 as discussed in section 4.2.1.

The robustness of power delivery to the pixels is influenced by the resistance of the power grid and the activity of the matrix (amount of pixels firing). If the power grid resistance is relatively high and large transient currents are generated by significant activity, the resulting voltage drop can cause spatial and temporal threshold variations or even induce extra hits that are possible to set off a chain reaction. Therefore, in order to increase the front-end immunity to power supply noise, the discriminator and pre-amplifier ground have been separated. Additionally, elaborate post-layout simulations of a full double column that include a power grid model have been performed to identify potential issues.

The chip periphery has been designed to allow easier system integration by essentially requiring two sets of wires (clock and data) for the communication uplink and downlink in order to operate the chip. Configuration and control is performed via a command decoder operating at 160 MBps, while the R/O controller is integrated on-chip. The readout is continuous (trigerless) and is common for all sectors of the matrix. Communication to the DAQ is achieved through a Low Voltage Differential Signaling interface (LVDS) by using high-speed LVDS transceivers and output data is transmitted with a speed up to 320 MBps. Analog voltage and current biases are generated by a modular 8-bit DAC. The total power consumption is approximately equal to 650 mW and is analyzed in Table 4.7. The matrix power consumption, which is equal to the sum of the analog, BCID distribution and dynamic readout contributions is equal to approximately 170 mW/cm^2 and is slightly higher compared to TJ-Monopix1 due to the smaller pixel size.

Table 4.7: TJ-Monopix2 power consumption summary

Contributor	Power consumption
Matrix analog	90 mW/cm^2
BCID distribution	80 mW/cm^2
Matrix readout dynamic (ATLAS ITk L4 hit rate)	< 5 mW/cm^2
Periphery (w/o LVDS)	60 mW
LVDS I/O	< 120 mW

4.5.1 Pixel design and layout

The 2 × 2 pixel core layout, shown in Fig. 4.83, is fully optimized and as dense as possible according to the technology limitations. The analog region includes the front-end circuit, the 3-bit threshold tuning DAC and the pixel configuration registers. The digital region is comprised of the 7-bit LE and TE memory (14 SRAM cells per pixel), the 10-bit address ROM, the readout control logic and the hit delay circuit that is used to correct for the BCID propagation delay. In order to save area, the 2 × 2 core is designed to share as much functionality as possible between the 4 pixels. The pixel position inside the 2×2 core is encoded in 2 bits, such that only one set of ROM transistors (2 for the internal address and 8 for the group address) are required for each core. A fast token signal that propagates across the double column is used to set the priority of each core, while pixels within the 2×2 core are arbitrated with the help of a local token signal.

The silicon well structure of the 2×2 pixel core is illustrated in Fig. 4.84. The removed deep p-well area is smaller compared to TJ-Monopix1 due to the higher density of the in-pixel electronics and therefore n-wells that house PMOS devices. Furthermore, large regions of continuous active area (STI) were avoided in order to prevent an uneven distribution of the n^- implant layer depth that negatively impacts charge collection.

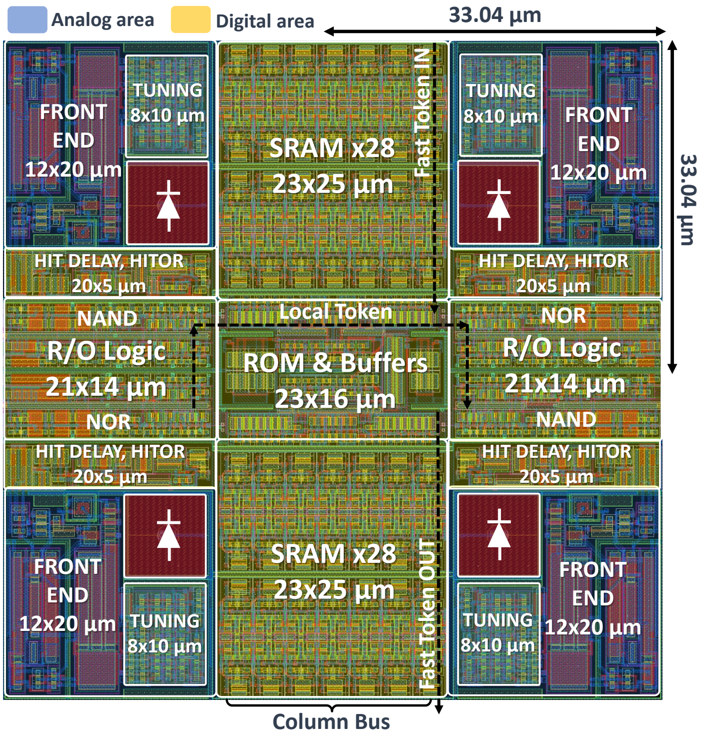


Figure 4.83: Layout of a TJ-Monopix2 2×2 pixel core. The separate analog (blue) and analog (yellow) areas are highlighted.

4.5.1.1 Improved front-end circuit design

The schematic of the improved analog front-end[6] used in TJ-Monopix2 is shown in Fig. 4.85. There are two variations of the front-end circuit, called "normal" and "cascode", depending on the inclusion of the cascode transistor M_{1C} which can improve performance by increasing the pre-amplifier gain (see below). Instead of a PMOS transistor, a simple diode is used to reset the input node, in order to increase tolerance to TID radiation effects. Additionally, the input capacitance is reduced by approx.

[6] The TJ-Monopix2 front-end is an evolution of the TJ-Monopix1 front-end circuit which is analyzed in detail in section 4.3.1.1.

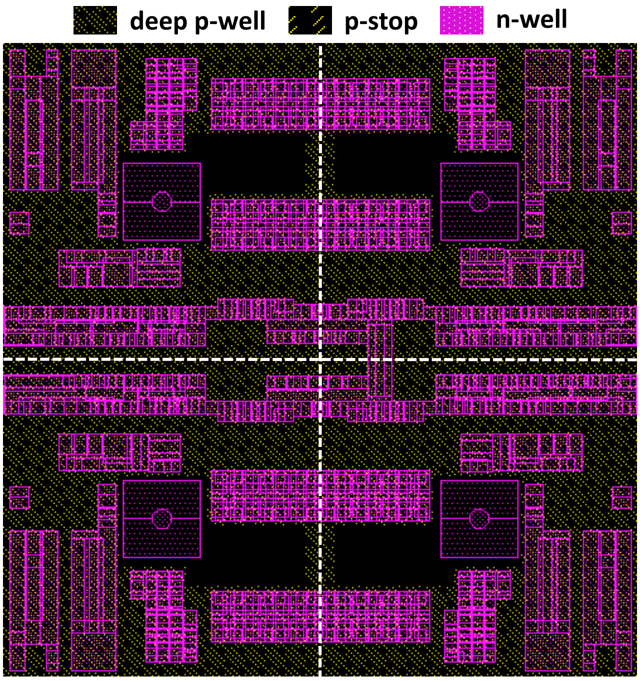


Figure 4.84: Top-view illustration of the TJ-Monopix2 2 × 2 pixel group silicon well layout. A portion of the deep p-well has been removed (RDPW) to enhance the lateral electric drift field.

0.2 fF, resulting in a slightly higher input signal amplitude. As mentioned, the separate pre-amplifier and discriminator ground helps to reduce crosstalk from the switching noise that is generated when the discriminator is activated which can become significant if there is high activity of the pixel matrix (large number of pixels firing).

The gate of M_3 is no longer connected to node GN and its potential is set by the adjustable bias voltage V_{CASP} generated by the DAC. This modification gives more flexibility and an extra handle to tune the front-end in order to maximize the pre-amplifier output dynamic range and gain for different backbias voltages, temperature variations and process corners.

Transistor M_{16} is introduced to provide clipping of the ToT, as required by the readout architecture in order to avoid data loss by late copy (see section 4.2). Clipping works by creating a conductive path between the pre-amplifier output (OUTA) and GN node through transistor M_{16} when the output voltage amplitude V_{OUTA} becomes higher than $V_{\mathrm{CLIP}} + V_{\mathrm{TH}}^{M_{16}}$, where $V_{\mathrm{TH}}^{M_{16}}$ is the threshold voltage of M_{16}. Therefore, the voltage at node GN that controls the gate of M_1 is pushed back up by V_{OUTA} and the output baseline is quickly restored. The clipping value is adjustable by the control voltage V_{CLIP} that is generated by the biasing DAC. This feature can be disabled by setting $V_{\mathrm{CLIP}} = V_{\mathrm{dd}}$.

Performance in terms of noise and threshold dispersion has been improved by addressing the following shortcomings of the TJ-Monopix1 pre-amplifier design:

- Due to the relatively small size of the coupling capacitor C_c, implemented by transistor M_6, the time constant ($\tau = R_{\mathrm{GN}}C_c$) of the high pass filter formed by C_c and the input resistance at node GN was short and a significant percentage of the input signal was lost due to inefficient coupling of the source follower output to the gain stage input node (GN). Therefore, in the case

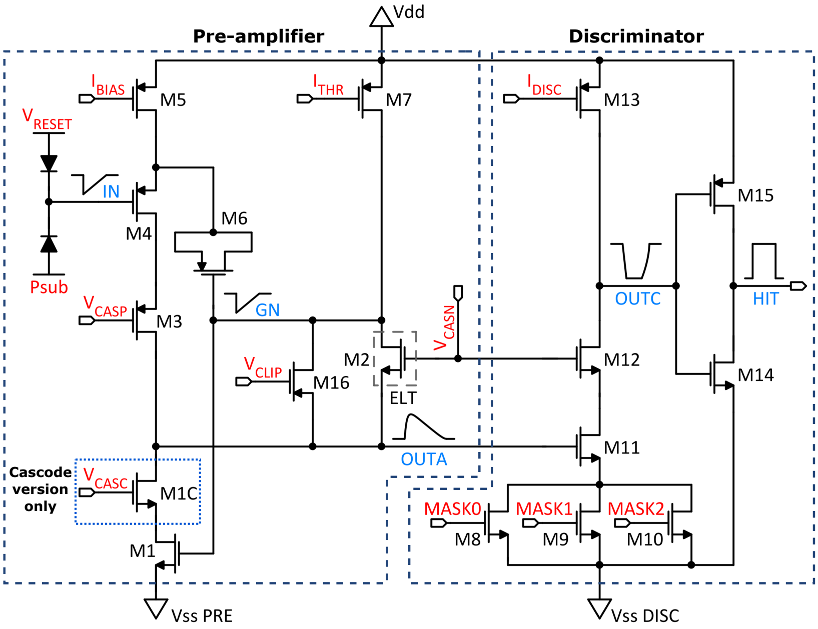


Figure 4.85: Schematic of the improved TJ-Monopix2 front-end circuit. The cascode variant includes transistor M_{1C} that is used to further increase the gain.

of TJ-Monopix2, the area of transistor M_6, hence the capacitance C_c, has been increased by 7.5 times resulting in approximately 2 times higher gain and signal to noise ratio (100% increase).

- The gain stage input transistor M_1 had been designed with minimum length ($1/0.18\,\mu$m). As a result, its output resistance ($r_{\mathrm{ds}_1} \approx 1/L$) was smaller than the input resistance at the source of the feedback transistor M_2 ($r_i^{M_2} \approxeq 2/g_{m_2}$) and dominated the total resistance at the pre-amplifier output node R_{OUT} (eq. 4.3.1.1). Therefore, due to the low output resistance R_{OUT}, the pre-amplifier voltage gain was relatively small (eq. 4.23,4.26). In the case of TJ-Monopix2, transistor M_1 has been enlarged to $1.25/0.8\,\mu$m, eventuating in approximately 50% higher gain at the threshold.

The combination of these modifications (enlarging transistors M_1 and M_6) results in an expected reduction of the ENC by approximately 2 times (due to the higher SNR) and a total increase of the gain at threshold from $0.4\,\mathrm{mV}/e^-$ to $1.2\,\mathrm{mV}/e^-$. A large pre-amplifier gain is the key to reduce threshold dispersion because it results in a high the output signal amplitude v_{OUTA} compared to the detection threshold voltage variation due to transistor mismatch. Increasing the gain is a particularly effective measure because it directly reduces the influence of the discriminator input transistor mismatch (see eq. 4.28), which was responsible for approx. 42% of the total threshold variance of the TJ-Monopix1 front-end.

Although increasing the length of transistor M_1 has lead to a significant improvement of the gain, its output resistance (r_{ds_1}) is still smaller that the output resistance at the drain of M_3 ($r_o^{M_3}$, eq. 4.19), which is high due to the cascode topology of M_3, and the input resistance at the source of M_2 ($r_i^{M_2}$, eq. 4.3.1.1), which is high due to the very low transconductance of M_2. Therefore, the total resistance at the output node (R_{OUT}), which is equal to parallel combination of r_{ds_1}, $r_o^{M_3}$ and $r_i^{M_2}$, is still limited by r_{ds_1}. Furthermore, because of the smaller W/L ratio of transistor M_1 (compared to the previous sizing in TJ-Monopix1), its transconductance (g_{m_1}) has been reduced by approximately 3.5 times. Increasing the width of M_1 as a means to achieve a higher transconductance g_{m_1} is impractical because it would result in an excessive increase of its area, hence a high capacitance at the output node (C_{OUT}, eq. 4.22) which impacts the timing performance.

In order to increase both the value of R_{OUT} and g_{m_1}, while avoiding a high capacitance penalty at the output, a cascode transistor (M_{1C}) can be used in series with M_1. The output resistance at the drain of M_{1C} is higher than r_{ds_1} by a factor of $g_{m_{1C}} r_{\mathrm{ds}_{1C}}$:

$$r_o^{M_{1C}} = g_{m_{1C}} r_{\mathrm{ds}_{1C}} r_{\mathrm{ds}_1} \tag{4.51}$$

The total small-signal resistance at the output node OUTA is equal to:

$$R_{\mathrm{OUT}} = r_i^{M_2} /\!/ r_o^{M_3} /\!/ r_o^{M_{1C}} \approxeq \frac{2}{g_{m_2}} /\!/ g_{m_3} r_{\mathrm{ds}_3} r_{\mathrm{ds}_4} /\!/ g_{m_{1C}} r_{\mathrm{ds}_{1C}} r_{\mathrm{ds}_1} \approxeq \frac{2}{g_{m_2}} \tag{4.52}$$

Therefore, due to the high resistance of the cascode configuration, R_{OUT} is significantly increased and is now dominated by $r_i^{M_2}$ (within the validity of the linearized small signal analysis). Furthermore, in the case of the cascode front-end, the W/L ratio of transistor M_1 has been increased to 2.06/0.36 µm. Hence, its transconductance (g_{m_1}) is higher by a factor of approximately 3.5 compared to the normal front-end (and similar to the TJ-Monopix1 front-end case). As a result of the higher R_{OUT} and g_{m_1}, the gain at threshold is increased to approx. 1.9 mV/e^-. Despite its large area, M_1 does not influence the capacitance C_{OUT}. In contrast, the size of M_{1C} has been kept relatively small (0.42/0.8 µm) in order to improve the timing response.

A challenge of the cascode implementation is to ensure that both transistors M_1 and M_{1C} operate in the saturation region due to the small available voltage headroom. Because the sum of the drain-to-source voltage V_{ds} of M_1 and M_{1C} is equal to the pre-amplifier output voltage V_{OUTA}, which is approximately equal to the gate-to-source voltage V_{gs} of the discriminator input transistor M_{11}, the following condition must be fulfilled:

$$V_{\mathrm{gs}}^{M_{11}} \geq V_{\mathrm{dsat}}^{M_1} + V_{\mathrm{dsat}}^{M_{1C}} \tag{4.53}$$

where V_{dsat} is the transistor drain-to-source saturation voltage. In the standby state (DC analysis), M_{11} operates in the sub-threshold regime due to the small discriminator current, which is in the order of 50 nA. Therefore, $V_{\mathrm{gs}}^{M_{11}}$ is only marginally higher than $V_{\mathrm{dsat}}^{M_1} + V_{\mathrm{dsat}}^{M_{1C}}$. The bias voltage at the gate of M_{1C} (V_{CASC}) must be set such that the available voltage headroom ($V_{\mathrm{gs}}^{M_{11}}$) is distributed between M_1 and M_{1C} in order to keep both of them in saturation. The valid range of V_{CASC} is equal to:

$$V_{\mathrm{gs}}^{M_{11}} + V_{\mathrm{gs}}^{M_{1C}} - V_{\mathrm{dsat}}^{M_{1C}} \geq V_{\mathrm{CASC}} \geq V_{\mathrm{dsat}}^{M_1} + V_{\mathrm{gs}}^{M_{1C}} \tag{4.54}$$

Because the threshold voltage (and as a result the V_{gs}) of M_{11} and M_{1C} increase with the p-well reverse bias voltage (backbias) due to the body effect (section 4.3.1.1), V_{CASC} has to be manually adjusted accordingly.

Although the front-end circuit is not complex, manual fine-tuning of the transistor sizing (W/L) and bias currents and voltages is a tedious task and can result in sub-optimal performance. Therefore, in the design process of TJ-Monopix2 a BFGS [90] numerical optimization algorithm included in the Cadence® Virtuoso® IC design suite [91] has been employed to fine-tune the front-end parameters as a means to maximize performance and achieve the desired threshold and ToT slope. Optimization is performed by iterating over a specified range of the free parameters in order to minimize a cost function that is generated using the given specifications, constraints and optimization goals. Good results can be achieved only if the input specifications, parameter range and starting point are carefully selected, otherwise the algorithm may be unable to reach the local minimum as a result of a specific constraint and lead to worse performance compared to manual tuning. Therefore, a good understanding and intuition on the design topology and trade-offs is required.

The specifications can be open or closed depending on the optimization goal. A closed specification essentially works as a constraint whose optimization will stop as long as the desired value is met. In contrast, if a specification is open, the algorithm will still try to optimize it in the desired direction (minimize or maximize). A list of the most important specifications that were used during the optimization run are given in Table 4.8. Additionally, the gate area of critical transistors has been constrained in order to avoid high RTS noise and minimize mismatch.

Table 4.8: Basic specifications used during the TJ-Monopix2 front-end optimization run

Specification	Value	Type
Threshold	$\geq 100\,e^{-}$	closed
Time-walk	$\leq 25\,\mathrm{ns}$	open, minimize
Gain at threshold	$\geq 1\,\mathrm{mV}/e^{-}$	open, maximize
ENC	$\leq 8\,e^{-}$	open, minimize
Phase margin	$\geq 75°$	closed
ToT at MPV (no clipping)	$\leq 2\,\mu\mathrm{s}$	closed
Total power	$\leq 1\,\mu\mathrm{W}$	closed
Total gate area	$\leq 150\,\mu\mathrm{m}^2$	closed

Each iteration of the optimization run consists of three simulations. A first simulation is performed using an input charge value equal to the desired threshold with the additional constraint that a hit should be registered. A second simulation is performed at a slightly lower input charge, for which no hit pulse should be generated. A third simulation is used to calculate the ToT slope (ToT at the MPV). The optimization run ends when no better solution within the specified parameter range can be achieved after a large number of iterations.

Pre-amplifier transfer function and gain A better understanding of the impact that the TJ-Monopix2 front-end improvements have on its frequency response can be obtained by observing the pre-amplifier small-signal voltage transfer function plot shown in Fig. 4.86. In order to keep the comparison with the TJ-Monopix1 pre-amplifier transfer function (Fig 4.28) consistent, the feedback current has been set to 2.5 nA. The first major modification is the increase of the pre-amplifier coupling capacitor C_c from approximately 40 fF to 300 fF that results a significantly longer time constant at node GN ($\tau_{\mathrm{HP}} = C_c R_{\mathrm{GN}}$, where R_{GN} is the total input resistance at node GN). Therefore, the cutoff frequency of the high-pass filter at GN, which is given by the open-loop pole $p_{\mathrm{GN}} = 1/C_c R_{\mathrm{GN}}$ (see eq. 4.23), is reduced by a factor of approximately 7.5 and becomes lower than 1 MHz. The resulting extension of the pre-amplifier bandwidth towards the low frequency range is crucial in order to achieve a higher gain and SNR because a significant portion of the input signal spectral power is contained at low frequencies[7]. Hence, the signal loss due to AC-coupling of the source-follower output to the gain stage input transistor M_1 through C_c is reduced by approximately a factor of 2.

The second major modification is the increase of the voltage amplification gain $A_v = g_{m_1} R_{\mathrm{OUT}}$ (see eq. 4.14,4.23) by extending the length of M_1 and introducing the cascode transistor M_{1C} (in the case of the cascode front-end variant). The higher voltage gain results in an increase of the transfer function amplitude at the pass-band by approx. 9 dB in the case of the normal front-end and 16 dB in the case of the cascode front-end. Additionally, according to the closed loop transfer function equation (eq. 4.26), the position of the closed loop poles p_1 and p_2 is shifted due to the combined effect of the higher loop gain (as described by Fig. 4.24) and lower frequency of the open-loop pole at the output ($p_{\mathrm{OUT}} = 1/C_{\mathrm{OUT}} R_{\mathrm{OUT}}$). Therefore, p_1 and p_2 become a complex conjugate pair and are additionally moved closer to the imaginary axis. As a result, the transfer function becomes steeper around the center frequency or, in the language of filter theory, the quality (Q) factor is increased. This effect is more pronounced in the case of the cascode front-end, as expected due to the higher gain.

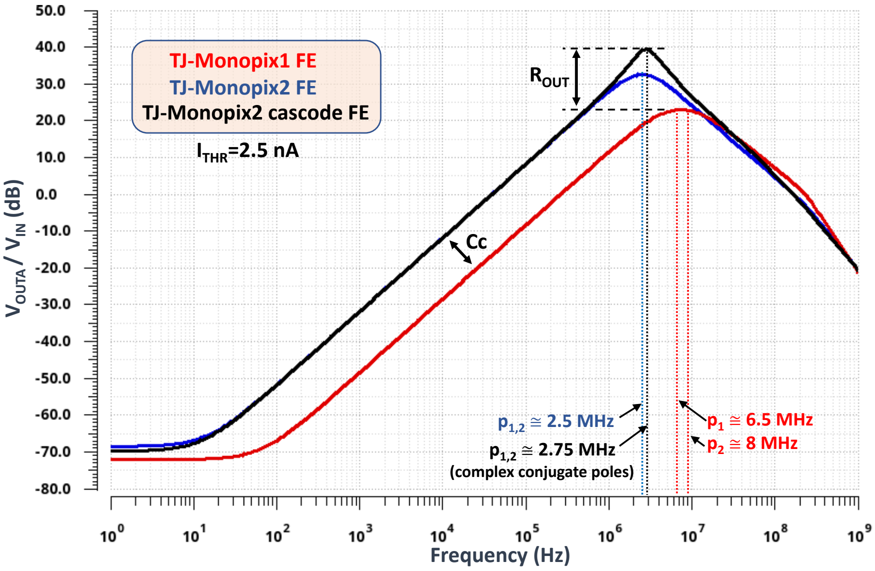


Figure 4.86: Comparison of the TJ-Monopix2 (both variants) and TJ-Monopix1 pre-amplifier transfer function for the same feedback current. The effects of the higher coupling capacitor (C_c) and output resistance (R_{OUT}) are highlighted.

It is important to note that due to the resulting pole shift, the increase of R_{OUT} leads to a significant reduction of the phase margin. In contrast, a larger coupling capacitor C_c increases the phase margin by shifting the dominant pole of the loop gain equation (eq.4.25) to lower frequencies. Therefore, in order to maintain a high phase margin and ensure the stability of the pre-amplifier circuit, it is essential to simultaneously enlarge the capacitance C_c when the voltage amplification gain is increased. As a result of the aforementioned opposite effects, the phase margin of the TJ-Monopix2 front-end

[7] The input signal waveform can be approximated for a long reset time constant by a Heaviside unit step function, whose Fourier transform is given by $\mathcal{F}(H(x))(\omega) = 1/j\omega + \pi\delta(\omega)$

(non-cascode variant) remains high and equal to the value of the TJ-Monopix1 front-end (90°) despite the total gain increase by a factor of 3. In the case of the cascode variant, the phase margin drops to 79° due to the higher voltage gain but is still more than adequate.

The simulated transient response of the TJ-Monopix2 (both normal and cascode variants) and TJ-Monopix1 pre-amplifier output for two input charge values is shown in Fig. 4.87. The improvement of the TJ-Monopix2 front-end gain can be clearly observed. For an input charge of only 300 e^-, the output voltage amplitude of the non-cascode pre-amplifier variant is approximately equal to 600 mV, while in the case of the cascode pre-amplifier the output voltage amplitude already approaches the dynamic range and starts to compress. Due to the TJ-Monopix2 front-end high gain and large C_c capacitance (that takes longer to charge up during the output baseline reset) the feedback current (I_{THR}) has to be increased to approx. 8 nA in order to avoid an excessively high ToT slope (that limits the ToT resolution) and keep the ToT duration below 2 µs at the input charge MPV. It should be noted that the TJ-Monopix1 front-end is normally operated with a lower feedback current due to the 7.5 times lower C_c capacitance value compared to TJ-Monopix2. Therefore, in the case of the TJ-Monopix1 pre-amplifier response simulation of Fig. 4.87, I_{THR} has been set to 1 nA.

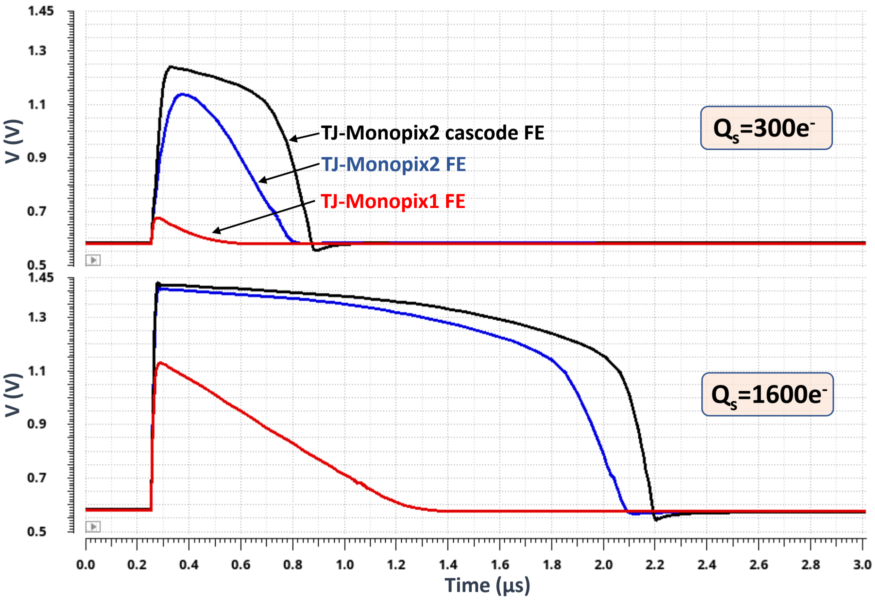


Figure 4.87: Post-layout simulation of the TJ-Monopix2 pre-amplifier output voltage for two input charge values ($Q_s = 300\,e^-, 1\,600\,e^-$). The response of TJ-Monopix1 is also included for comparison. The feedback current (I_{THR}) is set to 8 nA in the case of TJ-Monopix2 and 1 nA in the case of TJ-Monopix1 in order to maintain a similar feedback ratio after the increase of the coupling capacitance C_c.

The gain (in mV/e^-) of the TJ-Monopix2 front-end pre-amplifier as a function of the input charge is shown in Fig. 4.88. For small input charges and therefore small output voltage amplitudes, the pre-amplifier voltage gain is minimum because R_{OUT} (see eq. 4.52,4.3.1.1) is dominated by the input resistance at the source of M_2 ($r_i^{M_2} \approxeq 2/g_{m_2}$). The higher gain of the cascode variant in this region is mainly attributed to the higher transconductance of transistor M_1 (g_{m_1}). As the output amplitude increases, the transconductance of transistor M_2 (g_{m_2}) decreases due to the pre-amplifier non-linear behavior, as discussed in section 4.3.1.1, until M_2 is completely switched off. Therefore, $r_i^{M_2}$ increases with the output amplitude and quickly becomes very high. As a result, R_{OUT} becomes approximately equal to $r_o^{M_3} /\!/ r_{\text{ds}}^{M_1}$ in the case of the normal front-end variant and $r_o^{M_3} /\!/ r_o^{M_{1C}}$ in the case of the cascode front-end variant. Due to the higher R_{OUT}, the gain increases with the input charge until it reaches a maximum when the output voltage amplitude becomes equal to the output dynamic range and therefore cannot increase any further. As expected, there is a steep decrease of the gain after the maximum since the output voltage amplitude remains approximately constant. The maximum in the case of the normal front-end is approximately 1.8 mV/e^-, while in the case of the cascode front-end the maximum is much higher, at around 3.7 mV/e^- because $r_o^{M_{1C}} > r_{\text{ds}}^{M_1}$ (see eq. 4.51).

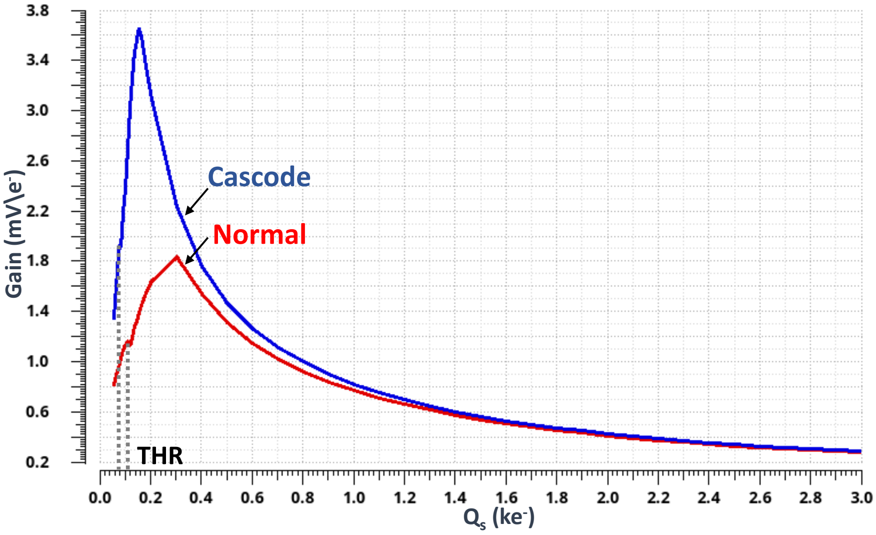


Figure 4.88: Post-layout simulation of the TJ-Monopix2 front-end charge to voltage conversion gain G_e (mV/e^-). Both pre-amplifier variations, normal and cascode, are included.

Timing response and ToT As discussed in section 4.3.1.1, the front-end timing response is mainly determined by the pre-amplifier output rise time which depends on its bandwidth and more specifically the low-pass behavior at high frequencies. Typically, the bandwidth of a voltage amplifier with feedback (e.g. resistive divider) is given by $GBW \cdot \beta$, where GBW is the amplifier gain-bandwidth product and β is the feedback ratio. However, because the TJ-Monopix pre-amplifier feedback is frequency dependent and designed to be effective only at relatively low frequencies, the gain at high frequencies, and therefore the pre-amplifier bandwidth, depends on the resistance at the output (R_{OUT}). For a constant gain-bandwidth product, $GBW = g_{m_1}/C_{\text{OUT}}$, an increase of R_{OUT} will result in higher voltage amplification gain ($g_{m_1} R_{\text{OUT}}$) and lower cutoff frequency ($1/2\pi R_{\text{OUT}} C_{\text{OUT}}$).

The timing response of the TJ-Monopix2 front-end (non-cascode) is influenced by the circuit modifications in two main ways. First, the pre-amplifier GBW is reduced due the decrease of g_{m_1} by approximately 3.5 times and the slightly higher output capacitance C_{OUT} as a result of the resizing of transistor M_1. Second, because the output resistance (R_{OUT}) is increased, the bandwidth is reduced by being traded for gain as can be confirmed by the lower cutoff frequency of the transfer function shown in Fig 4.86.

The simulated time-walk curve is shown in Fig. 4.89. Due to the slower pre-amplifier response, the overdrive charge is increased compared to TJ-Monopix1 and becomes equal to 65 e^-. However, despite the higher time-walk, because of the expected threshold reduction from approx. 300 e^- to 100 e^-, the in-time threshold of TJ-Monopix is equal to only 165 e^- (compared to 340 e^- in the case of TJ-Monopix1). Therefore, the in-time efficiency (which is ultimately the goal in terms of timing) is expected to increase as a result of the front-end improvements.

In the case of the cascode front-end, the timing response is improved, in spite of the higher gain, due to the increase of g_{m_1} by approx. 3.5 times and the lower contribution to the output capacitance C_{OUT} by the relatively small cascode transistor M_{1C}. Therefore, the time-walk is smaller and the over-drive charge is reduced to approx. 40 e^-. Furthermore, because of the smaller threshold dispersion as a result of the high pre-amplifier gain, it may be possible to achieve an even lower threshold using the cascode front-end. Hence, the in-time threshold can theoretically be even lower than 140 e^-.

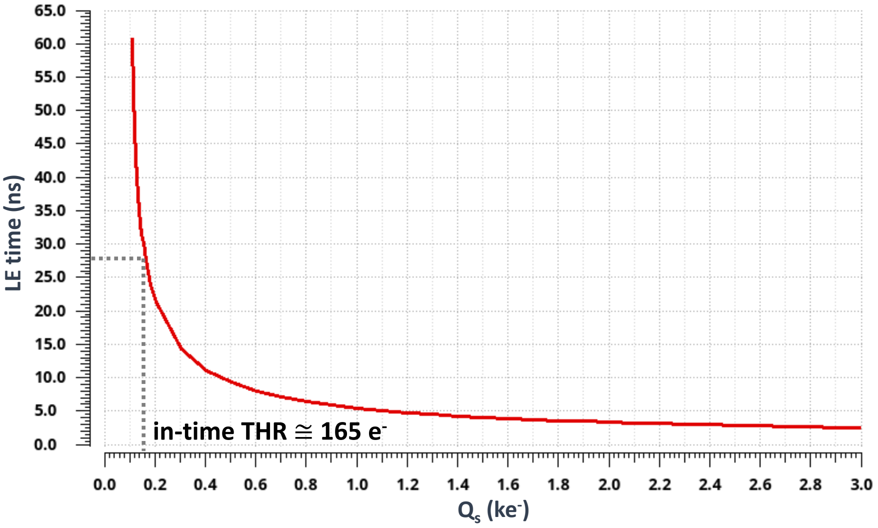


Figure 4.89: Post-layout simulation of the TJ-Monopix2 front-end (non-cascode) time-walk curve. The 25 ns in-time threshold is equal to 165 e^- (65 e^- overdrive charge).

The ToT response of the TJ-Monopix2 front-end, shown in Fig 4.90, is not as linear as in the case of TJ-Monopix1 due to the diode input reset that is used instead of a PMOS transistor. As discussed, the main drawback of a diode reset is the dependence of the input reset rate on the sensor leakage current that can influence the ToT slope. However, for the expected leakage current of an irradiated sensor at −30° and the purpose of distinguishing small from large input charges, that is of interest for the ATLAS experiment, the ToT resolution is adequate. The ToT clipping functionality is also demonstrated by setting V_{CLIP} for a maximum ToT value of approximately 1.6 µs.

Noise and threshold dispersion The ENC and threshold dispersion simulation of the normal (non-cascode) TJ-Monopix2 using the s-curve method is given in Fig. 4.91. As expected, the ENC is reduced from approx. 8.9 e^- to 5.3 e^- due to the higher SNR as a result of the increased coupling capacitance (C_c). In addition, the RTS noise (that is not included in the simulation models) is expected to be significantly reduced due to the approx. 5.5 times larger area of transistor M_1. It should be noted that the size of M_1 in the case of TJ-Monopix2 is about 2 times higher compared to mini-Malta [67], which already demonstrated a considerable improvement of the ENC long-tail due to RTS by enlarging M_1 2.5 times compared to TJ-Monopix1 and TJ-Malta1.

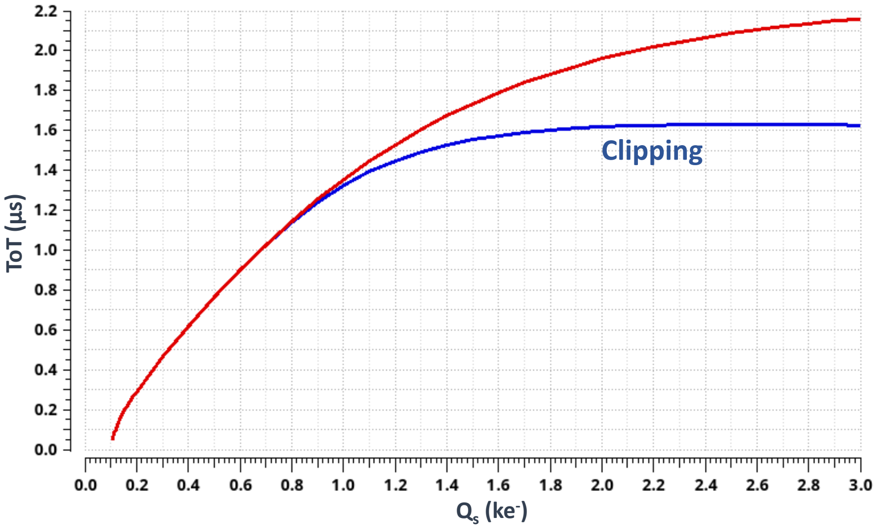
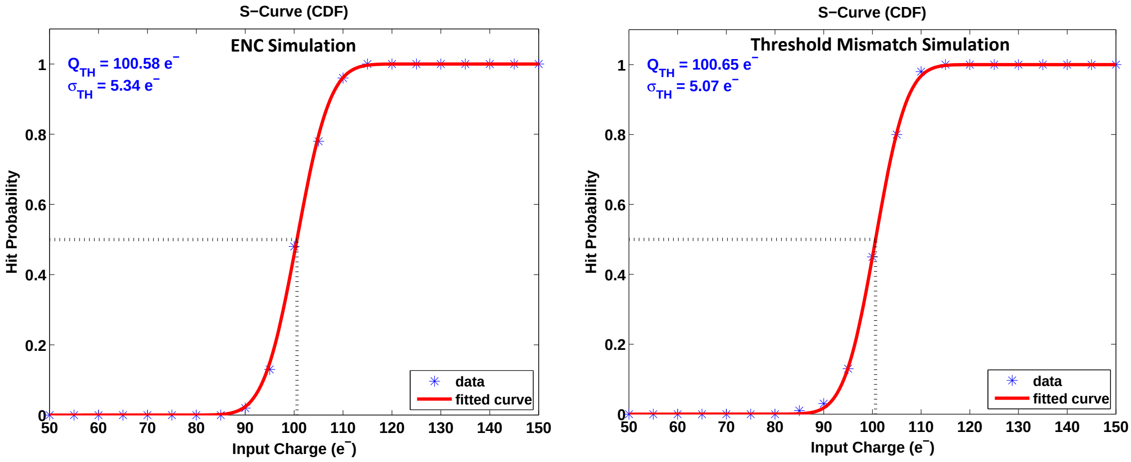


Figure 4.90: Simulated post-layout ToT curve of the TJ-Monopix2 front-end (non-cascode variant).

The threshold dispersion is also significantly reduced by a factor of 5, from approx. $25\,e^{-}$ to $5\,e^{-}$, due to the higher gain (by approx. 3 times) and the optimization of mismatch by increasing the area of transistors (e.g. M_1). The combined low ENC and threshold dispersion yield a minimum operating threshold of:

$$Q_{\mathrm{TH_{min}}} \approx 10\sqrt{\sigma_{\mathrm{TH}}^2 + \mathrm{ENC}^2} \approxeq 75\,e^{-} \tag{4.55}$$

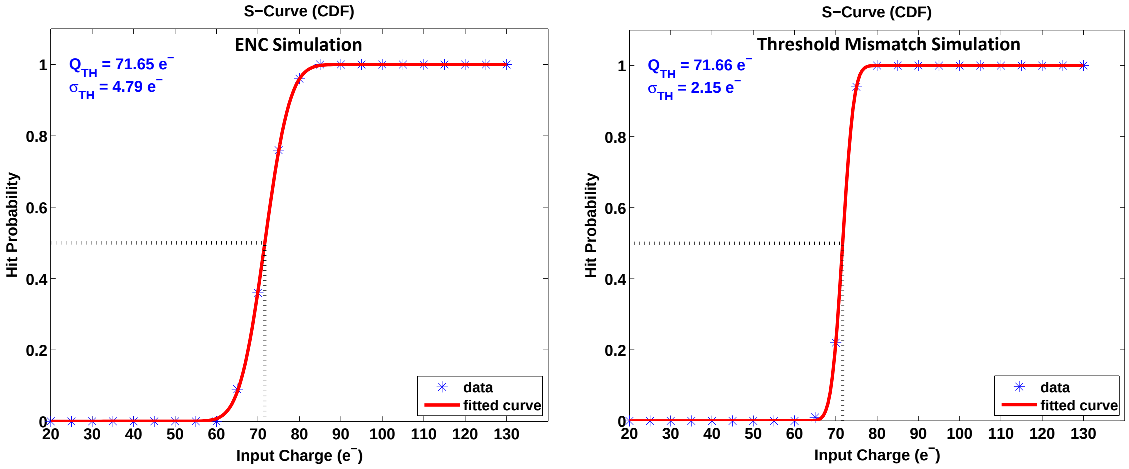
where 10σ of the distribution has been taken into account to ensure low noise hit rate. The nominal threshold is set a bit higher, to $100\,e^{-}$, in order to account for other effects such as crosstalk and systematic threshold variations.



(a) (b)

Figure 4.91: Post layout simulation of the a) ENC and b) threshold dispersion of the TJ-Monopix2 normal front-end.

In the case of the cascode front-end, the simulated ENC and threshold performance is shown in Fig. 4.92. The higher voltage gain does not increase the SNR (in first order approximation) because the signal and noise are equally amplified. Therefore, the ENC is only slightly reduced to approx. $4.8\,e^{-}$. In contrast, due to the even higher gain of the cascode implementation, the threshold dispersion is reduced to only $2.15\,e^{-}$. As a result, the minimum threshold (10σ) is approximately equal to $50\,e^{-}$.



(a) (b)

Figure 4.92: Post layout simulation of the a) ENC and b) threshold dispersion of the TJ-Monopix2 cascode front-end.

Suppression of crosstalk due to ground bounce During the measurement of TJ-Monopix1 it has been observed that when a large number of pixels fire within the same time scale (ToT duration), due to for example a low threshold setting, new hits can be induced due to crosstalk and a chain reaction can be initiated causing continuous hit generation across the whole matrix. In order to restore the stability of the pixel matrix, the user has to intervene by masking every pixel for a short time.

It is therefore crucial to identify the source of the induced crosstalk noise and improve the design in order to suppress it. To this end, a simulation has been performed for the extreme case of n-1 (511) pixels of the same row, firing at the same time and therefore injecting noise to the n$^{\text{th}}$ pixel in the middle. The simulation results are shown in Fig. 4.93. It has been observed that the main source of crosstalk is the voltage drop of the pre-amplifier ground (ground bounce) as a result of the accumulated current that is drawn from the discriminator of the n-1 pixels when a hit pulse is produced. Sensitivity to the pre-amplifier ground fluctuations is very high because it directly shifts the gate-to-source voltage (V_{gs}) of the gain stage input transistor M_1. Therefore, noise at the pre-amplifier ground rail is amplified with the same gain as the signal at node GN, essentially acting as a "second input" to the front-end.

A row has been used in the simulation (instead of e.g. a column) because it represents the worst case in terms of power supply voltage drop due to the resistance of the power grid being maximum at the middle of the matrix in the horizontal direction (as a result of the power pad placement). As mentioned, power supply fluctuations can be significant due to the large matrix area and small pixel size that constrains the width of the power grid metal wires. In the case of TJ-Monopix2, each main analog power horizontal top-metal wire (one per pixel) is approximately 7 µm wide, yielding a resistance per pixel of about 0.3 Ω. Therefore, considering that power is applied from both sides, the static voltage drop due to the front-end DC current (≅ 550 nA) in the middle of each row is approximately equal to 5 mV.

When the pre-amplifier output voltage crosses the threshold, the discriminator input transistor M_{11} is activated and initially a large current is drawn for a short time until the discriminator output node OUTC is discharged. As a result, a high amplitude voltage drop transient with short duration (in the order of a few ns) is induced at the ground rail. Afterwards, and as long as the hit pulse is active (ToT duration), the discriminator draws a current equal to I_{DISC}, which is typically similar to the main branch bias current $I_{\text{BIAS}} = 500\,\text{nA}$. Therefore, during this time the front-end power consumption is doubled and the voltage drop of the ground rail in the middle of each row becomes equal to approximately 10 mV. The pre-amplifier of the pixel which is used as the noise "victim", responds to the ground voltage fluctuation by producing an output of about 185 mV which is enough to trigger its discriminator and generate an extra hit pulse.

In order to suppress noise due to ground bounce, a different set of metal wires has been used in the case of TJ-Monopix2 for the pre-amplifier and discriminator ground across the matrix power grid. Therefore, as can be observed from the simulation of Fig. 4.93, the pre-amplifier ground is more stable and no extra hits are generated due to crosstalk.

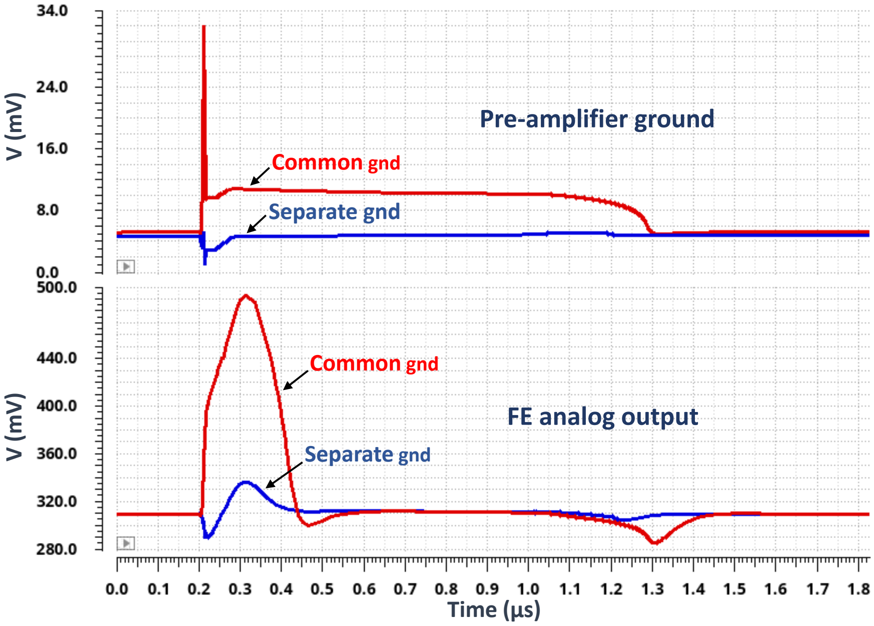


Figure 4.93: Simulation of crosstalk through the power grid due to switching of the discriminator in the extreme case of n-1 (511) pixels of the same row firing at the same time as a response to an input charge of $1\,000\,e^{-}$. The waveforms show the impact of the injected noise (voltage drop) on the n$^{\text{th}}$ pixel in the middle of the matrix. Crosstalk is significantly suppressed by separating the pre-amplifier and discriminator ground.

HV AC coupling capacitor optimization In order to reduce the high signal loss due to AC-coupling in the case of the HV flavor, there has been an effort to improve the layout of the MOM structure and optimize the coupling capacitor value. The idea is to take advantage of the capacitance between

VIAs that are used to connect different metal layers as a means to maximize the ratio of the coupling capacitance C_{ac} to the parasitic capacitance C_{p_1} and C_{p_2} (see section 4.3.1.1). The concept of the new layout is shown in Fig. 4.94. As in the case of TJ-Monopix1, the MOM structure is composed of multiple metal layers in an inverse pyramid shape (more fingers for each layer). However, metal-1 is not used anymore because its parasitic capacitance is high due to being close to the substrate. The coupling capacitance C_{ac} is the sum of individual capacitances between fingers of the same layer ($C_{m_i-m_i}$), between metal VIAs ($C_{v_i-v_i}$) and between different layers diagonally ($C_{m_i-m_j}$). Although $C_{m_i-m_j}$ is much smaller compared to TJ-Monopix1 due to the higher distance and shielding effect of other metal segments, the overall capacitance density is increased by approximately 13% as calculated by parasitic extraction using electromagnetic field solving.

The next step is to size the MOM capacitor such that the signal amplitude at the front-end input (eq. 4.13) is maximized. In order to do so, a mathematical expression of the relationship between the coupling capacitance (C_{ac}) and the parasitic capacitance at the sensor and the front-end input (C_{p_1} and C_{p_2}) has been extracted by fitting the calculated values for various MOM structure sizes with a polynomial function as shown in Fig. 4.95. The capacitance optimization is performed by a parametric simulation of the HV front-end that makes use of the extracted C_{ac}-$C_{p_{1,2}}$ relationship. Simulation results are shown in Fig. 4.96. The optimum point is defined by the trade-off between the higher coupling ratio (eq. 4.11) and the higher parasitic capacitance as C_{ac} is increased. The signal loss for the optimum C_{ac} capacitance value of 7 fF is 41.5% which is an improvement compared to TJ-Monopix1 (50% signal loss).

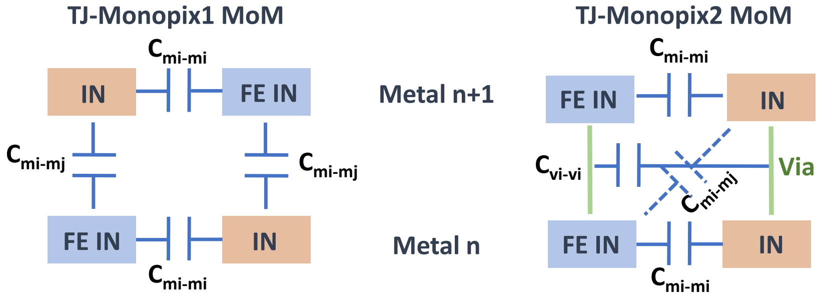


Figure 4.94: Cross section of the optimized MOM capacitor layout implemented in the HV front-end of TJ-Monopix2. The density is increased by additionally taking advantage of the capacitance between metal VIAs.

Baseline adjustment The V_{CASN} generator circuit (see chapter 4.3.1.1) that is used to automatically adjust the pre-amplifier output baseline and compensate for p-well bias (backbias) and ground voltage drop variations has been also improved in order to allow for more accurate control the discriminator standby current. The schematic of the new design that is used in TJ-Monopix2 is illustrated in Fig. 4.97. It is based on a replica biasing concept that uses a copy of the front-end circuit (with some modifications) in order to achieve an almost identical operating point of the V_{CASN} generator and corresponding front-end transistors through symmetry and therefore improve the accuracy of the generated V_{CASN} bias voltage. The V_{CASN} generator circuit is essentially constructed by removing the front-end input source-follower transistor (M_4) and coupling capacitor (M_6) and modifying the discriminator branch by adding a diode-connected transistor (M_7) in the place of the front-end

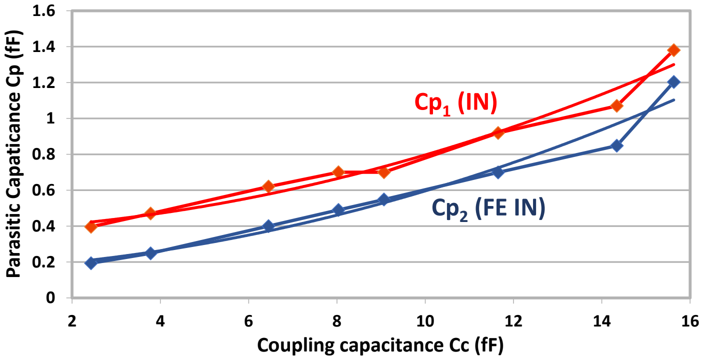


Figure 4.95: Graph of the MOM structure coupling (C_{ac}) and parasitic (C_{p_1},C_{p_2}) capacitances. Each point corresponds to the values calculated using a parasitic extraction software for different MOM capacitor sizes. The continuous line represents a 3rd order polynomial fit.

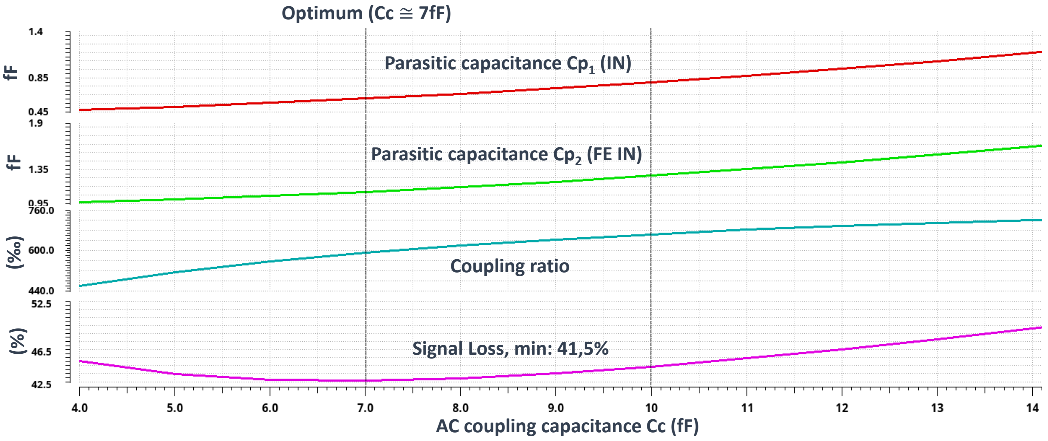


Figure 4.96: Optimization of the AC-coupling capacitance in order to maximize the signal amplitude at the front-end input. The signal loss is approximately 8.5‰ lower compared to TJ-Monopix1 for the optimum value of $C_{\mathrm{ac}} = 7$ fF.

discriminator NMOS cascode transistor (M_{12}).

Automatic adjustment of V_{CASN} is effectively performed by two nested feedback loops. The first is equivalent to the front-end feedback that makes use of the common-gate transistor M_2 in order to adjust the gate voltage of M_1 such that the gate voltage of M_6 is equal to $V_{\mathrm{CASN}} - V_{\mathrm{gs}}^{M_2}$. The second makes uses of transistor M_7 in order to adjust V_{CASN}, and therefore the gate voltage of M_6, such that the current through M_6 is equal to the bias current I_{CASN}. Because the front-end discriminator standby current is equal to the drain current of M_6, it can be accurately set by adjusting I_{CASN} and is independent of the p-well reverse biasing voltage (backbias). Due to the nested feedback loops, that tend to be unstable, a large capacitance has been added to the V_{CASN} output to ensure stability and additionally filter high-frequency noise.

Radiation tolerance to TID effects Measurement results of TJ-Monopix1 have indicated that the front-end radiation tolerance to TID must be improved. The main shortcomings of the TJ-Monopix1

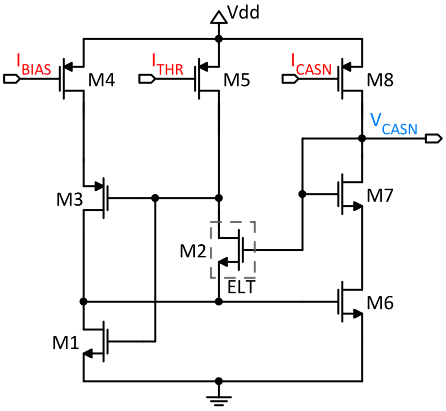


Figure 4.97: Schematic of the improved V_{CASN} generator circuit used in TJ-Monopix2 to automatically adjust the pre-amplifier output baseline. It is cased on a replica biasing concept in order to accurately set the discriminator standby current.

front-end design in terms of radiation hardness, discussed in section 4.4.4, have been addressed by TJ-Monopix2. First, the input PMOS reset transistor that can be sensitive due to its small width and current has been replaced by a diode that is implemented by a p$^+$ diffusion within the n-well collection electrode (see section 4.3.1.1). Second, the length of the pre-amplifier gain stage input transistor (M_1) which was sensitive to RISCE (section 3.1.2.6) due to its short channel (1/0.18 µm) has been enlarged by a factor of approx. 4.4 times (1.25/0.8 µm), while its total area is increased by approx. 5.5 times.

M_1 is one of the most critical transistors because it directly affects the pre-amplifier gain. Therefore, radiation tolerance to TID can be significantly enhanced by increasing its size, as confirmed by X-ray irradiation measurements of the mini-Malta chip [72]. Mini-Malta contains two front-end versions, one similar to TJ-Monopix1 and TJ-Malta1 and one where the size of transistor M_1 has been enlarged to 1.22/0.38 µm. Measurement results of the pre-amplifier output voltage as a function of TID up to 80 Mrad for both front-end versions that are included in mini-Malta are shown in Fig. 4.98. In the case of the original front-end version with standard M_1 size, a large degradation of the gain is observed around a dose of approx. 10 Mrad that is similar to the corresponding measurement of TJ-Monopix1 shown in Fig. 4.77. In contrast, in the case of the enlarged M_1 size, there is no significant reduction of the gain and the front-end behavior is more consistent across the received dose range. A similar or even larger improvement of radiation tolerance to TID is expected in the case of TJ-Monopix2 because the length of M_1 is increased by a factor of 2.5 compared to the enlarged version of mini-Malta.

Front-end performance summary and comparison The performance and characteristics of the TJ-Monopix2 (normal and cascode variants) and TJ-Monopix1 front-end are summarized and compared in Table 4.9.

4.5.1.2 Threshold tuning and masking

Even though the simulated threshold dispersion due to mismatch of the front-end transistors is low in the case of TJ-Monopix2 (approx. $< 5\,e^-$) as a result of the circuit improvements, the uniformity of the threshold across the pixel matrix can be reduced due to systematic effects (e.g. related to biasing),

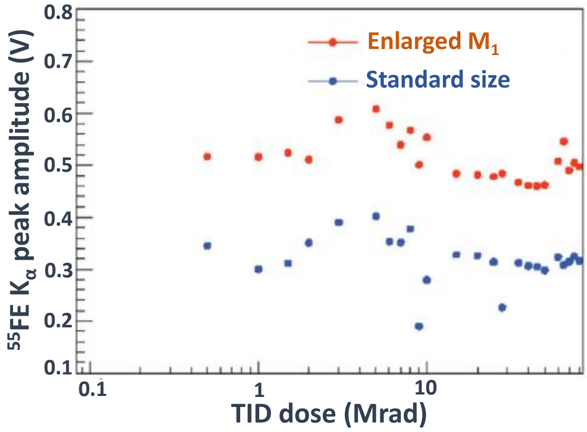


Figure 4.98: Measurement of the mini-Malta ^{55}Fe K_α fluorescence peak amplitude as a function of TID for two different front-end versions. Radiation tolerance is significantly improved by enlarging the pre-amplifier gain stage input transistor (M_1) by 2.5 times [72].

Table 4.9: TJ-Monopix front-end performance summary (simulated)

	TJ-Monopix1 FE	TJ-Monopix2 FE	TJ-Monopix2 cascode FE
Threshold (nominal)	300 e^-	100 e^-	70 e^-
In-time threshold	340 e^-	165 e^-	130 e^-
Overdrive	40 e^-	65 e^-	50 e^-
ENC (AC noise method)	9.5 e^-	3.8 e^-	2.8 e^-
ENC (S-curve method)	8.9 e^-	5.3 e^-	4.8 e^-
Threshold dispersion	25 e^-	5.1 e^-	2.2 e^-
Gain at threshold	0.4 mV/e^-	1.2 mV/e^-	1.9 mV/e^-
Phase margin	90°	90°	79°
ToT at MPV (no clipping)	1 µs	1.75 µs	1.9 µs
Power consumption	1 µW	1 µW	1 µW

process and temperature variations and radiation damage. Therefore, in-pixel threshold tuning has been implemented in order to ensure that a low threshold can be achieved, especially after irradiation. Its resolution is limited to 3-bits due to the small pixel size, but is considered adequate due to the reduced threshold dispersion of TJ-Monopix2.

Threshold trimming of individual pixels is performed with the help of a tuning DAC (TDAC), shown in Fig. 4.99 (b), which controls the discriminator active load (comparison) current I_{DISC}. In order to minimize its area, the TDAC does not generate I_{DISC} but works as an analog multiplexer, consisting of simple PMOS transistor switches, which select one of seven I_{DISC} lines that are provided by the main 8-bit biasing DAC. Each I_{DISC} (1-7) is generated by summing two currents, one that sets the coarse threshold ($I_{\mathrm{DISC}_{\mathrm{coarse}}}$) and one that is sets the fine tuning step ($I_{\mathrm{DISC}_{\mathrm{fine}}}$). Therefore, the available I_{DISC} current values that are distributed across the pixel matrix are equal to:

$$I_{\mathrm{DISC}_n} = I_{\mathrm{DISC}_{\mathrm{coarse}}} + (n-1) \cdot I_{\mathrm{DISC}_{\mathrm{fine}}}, \; 1 \leq n \leq 7 \tag{4.56}$$

I_{DISC_7} corresponds to the highest threshold, while I_{DISC_1} corresponds to the lowest threshold. The current that is selected by the TDAC (and is applied to the front-end) depends on the 3-bit tuning code that is stored in the pixel and is equal to:

$$I_{\mathrm{DISC}} = I_{\mathrm{DISC}_{\mathrm{coarse}}} + (\mathrm{TCODE} - 1) \cdot I_{\mathrm{DISC}_{\mathrm{fine}}}, \; 1 \leq \mathrm{TCODE} \leq 7 \tag{4.57}$$

where TCODE is the decimal representation of the TDAC code. The remaining value, TCODE = 0, is used to select the pixel for masking in order to save area (spare a separate masking register) and optimize routing of the matrix configuration lines. Each bit of the TDAC code is connected to the front-end masking switches M_8, M_9, M_{10} (Fig. 4.85). Therefore, if all bits are 0, the discriminator is disabled and the pixel is masked. The drawback of this approach is that the available tuning levels are reduced by one (from $2^{n_{\mathrm{TDAC}}} = 8$ to 7).

The TDAC code is stored in the in-pixel configuration memory, which is illustrated in Fig. 4.99 (b) and consists of three SR-latches with enable (which essentially function similar to a D-latch). The ENABLE signal is used to select the row to be configured, while the TCODE data bits (SET, RESET) are distributed to every pixel of the same column. Therefore, configuration is written to the matrix row by row, similar to a rolling shutter scheme.

4.5.1.3 In-pixel readout logic

The in-pixel digital readout (R/O) logic that is implemented in TJ-Monopix2 is shown in Fig 4.100. While the basic functionality is similar to TJ-Monopix1 (section 4.3.1.2), it has been redesigned from scratch in order minimize its area, improve its robustness and address the TJ-Monopix2 requirements and characteristics such as the small pixel size and large column height.

Two main new features have been added in order to deal with issues related to the large propagation delay of signals through the column-bus (long column effects). The first is the inclusion of the HIT delay circuit that is used to compensate for the delay of the BCID time-stamp across the column (see the relevant paragraph below). The second feature is the implementation of the fast token-pass logic that reduces the token propagation delay.

The drawback of a simple token-pass logic, similar to TJ-Monopix1, is that the token signal has to pass through each lower priority pixel of the double column. Therefore, since one gate delay is added at each step, the maximum token propagation time (top to bottom) can significantly increase as the

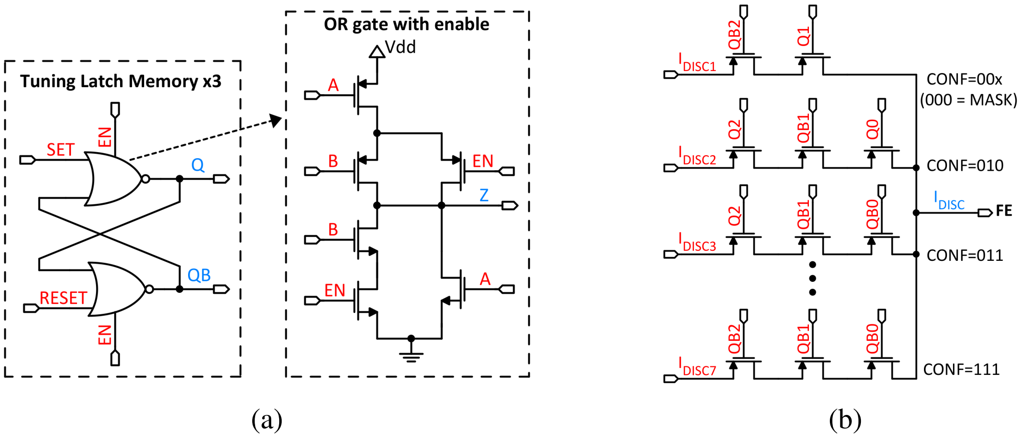


Figure 4.99: In pixel threshold tuning circuit used in TJ-Monopix2: a) schematic of the tuning memory latch, b) schematic of the 3-bit tuning DAC.

column size becomes larger. In the case of TJ-Monopix2, a simple token-pass logic would result in a propagation delay (top to bottom) higher than 100 ns. As discussed in section 4.2, a large token delay can result extra empty read cycles at the end of each readout sequence and therefore reduce the readout speed (at least in the case of a triggered readout). In order to ensure that the TJ-Monopix2 pixel matrix is compatible with future developments that feature a fast triggered readout, the token propagation delay has to be reduced below 50 ns for a maximum column bandwidth of 10 MHz. The fast token logic works by taking advantage of pixel grouping in 2×2 cores. Pixels are arbitrated in two levels using a local and a fast (group) token. The local token defines the priority in the lower level, within the 2×2 core, while the fast token defines the higher (group) level priority and propagates across the double column. The fast token is asserted if any pixel within the core is hit and passes through only one gate (NAND-NOR logic) of each lower priority core. Therefore the total number of gate delays is divided by 4 and the total propagation time is reduced below 35 ns across all process corners.

SRAM memory cell The LE/TE SRAM memory cell used in TJ-Monopix2 is shown in Fig. 4.101. Similar to TJ-Monopix1, the memory element (bi-stable latch) is formed by transistors $M_1 - M_4$ and can be written using transistors M_7 and M_8. However, there are two main differences: the memory readout circuit ($M_8 - M_9$) and the introduction of transistors M_5 and M_6 that are used to improve the BCID time-stamp signal integrity during multiple simultaneous write operations.

As discussed, TJ-Monopix2 uses single-ended pixel data transmission in order to reduce the column-bus width and achieve a small pixel size of $33.04 \times 33.04\,\mu\text{m}^2$. The data-bus readout is based on a current-mode sensing scheme, which is analyzed in section 4.5.2.1. SRAM LE/TE hit data bits are buffered and transmitted to the EoC using a controllable current source, implemented by transistor M_8, and a switch (M_9) that controls whether the bit-line is connected to M_8 depending of the memory state (MEM).

The current source (M_8) is activated only during the read phase by the internal read signal (READINT) that controls pixel access to the data-bus. While READINT is high (active), the gate of M_8 is connected to the bias voltage I_{READ} through the transmission gate formed by M_{10} and M_{11}, while M_{12} is switched off. In contrast, while READINT is low (inactive), the transmission gate is switched

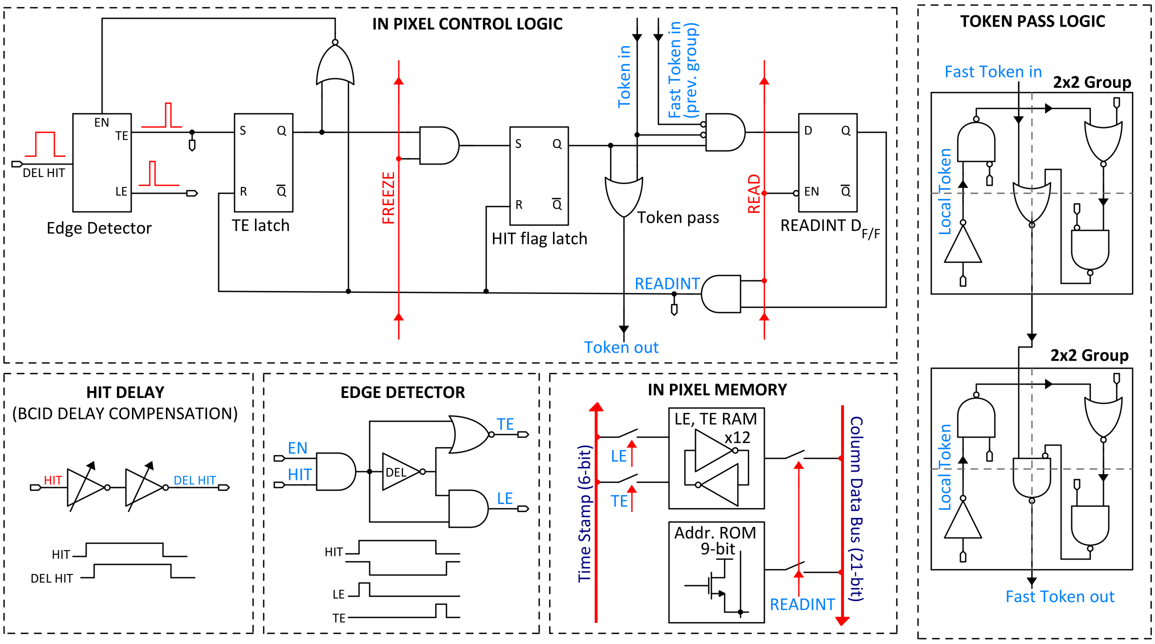


Figure 4.100: Schematic of the TJ-Monopix2 in-pixel R/O logic.

off and the gate of M_8 is pulled down by M_{12}. I_{READ} determines the transient current provided by the pixel during the readout phase and can be adjusted by the 8-bit biasing DAC. An additional I_{READ} mirror stage is added at the column bottom, within the active matrix area, in order to compensate for the body-effect due to p-well backbias. Therefore, the readout current is directly controlled within the pixel and can be set more accurately compared to the TJ-Monopix1 source-follower scheme. The value of I_{READ} should be high enough in order to achieve an adequate data-bus bandwidth ($\geq$ 10 MHz), while keeping crosstalk low.

During the readout phase, if the memory state is high (MEM=1), the bit-line is connected to M_8 through M_9 and is pulled-down by the current I_{READ} which is sensed by the EoC column-bus readout circuitry. In contrast, if the memory state is low (MEM=0), the bit-line is disconnected and no current is sensed at the EoC.

In the case of a conventional SRAM cell (such as the one used in TJ-Monopix1), the BCID lines directly "force" the state of the bi-stable latch (memory element) during the write phase. If the memory bit has to be flipped, the pull-up and pull-down transistors that are active depending on the current memory state, initially resist being toggled and draw current for a short time through the BCID lines until the metastable point is crossed after which the latch quickly settles to the new state through positive feedback. This loading effect on the BCID lines can, under certain conditions, cause a significant voltage drop that results in higher timing uncertainty or even wrong data to be recorded by flipping the sign of the differential BCID line voltage. In the case of TJ-Monopix2, the BCID line signal integrity during the write phase has to be considered due to the high line resistance as a result of the aggressive routing (minimum line width) and large column height. In order to demonstrate the BCID line loading effect, a simulation has been performed for an extreme case of 25 pixels at the column top, that use conventional SRAM cells, being written at the same time. Simulation results are

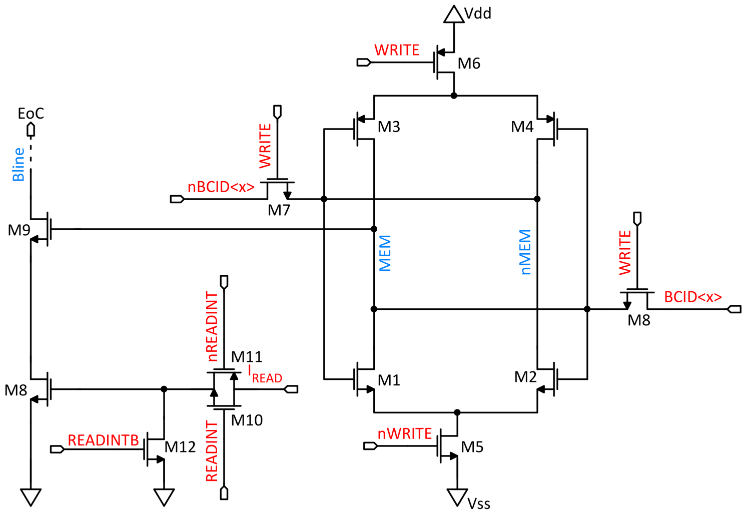


Figure 4.101: Schematic of the LE/TE SRAM cell used in TJ-Monopix2.

shown in Fig. 4.102. As can be observed, if the write operation occurs close to the BCID crossing point, the wrong bit is being recorded due to the voltage drop that results from loading the BCID lines and exceeds the small differential BCID voltage amplitude.

Although this effect is only significant is edge cases of multiple pixels being written simultaneously, the SRAM cell used in TJ-Monopix2 has been modified in order to eliminate loading of the BCID lines and improve timing accuracy around the crossing point. The idea is to disconnect the SRAM cell power and ground rail while the WRITE pulse is active using transistors M_5 and M_6 such that no current can be drawn from the BCID lines. Since power is disconnected, the complementary memory nodes of the SRAM latch (MEM, nMEM) simply follow the differential BCID line voltage. At the falling edge of the WRITE pulse, the BCID lines are disconnected and at the same time power is reapplied. Due to positive feedback, the SRAM latch will quickly settle to the new state according to the voltage difference developed on the MEM and nMEM nodes. Therefore, the TJ-Monopix2 SRAM cell works in a "sensing" mode during write, similar to a sense-amplifier. As can be observed from Fig. 4.102, when the new SRAM cell is used, the correct data bit is always recorded even if the write operation happens close to the crossing point.

Pixel address ROM The TJ-Monopix2 pixel address ROM implementation is shown in Fig. 4.103. Instead of using a separate ROM for each individual pixel, as in TJ-Monopix1, a common set of address ROM transistors is included in each 2 × 2 pixel core in order to save area. The pixel position inside the 2 × 2 core is encoded using a simple logic such that the address MSB defines the physical column (left of right) and the LSB defines if the pixel is on the top or bottom row of the core. The remaining 8 bits represent the group address, which is unique for each core. The group address is defined by hardwiring (either connecting or not) the ROM transistor switches to the corresponding bit-lines using binary encoding. This approach is significantly more area efficient because instead of 80 ROM transistors in total (2 transistors per bit, 10-bits per pixel), only 20 transistors and 3 digital gates are required.

Transmission of the ROM address data is similar to the case of LE/TE data (SRAM) and is based

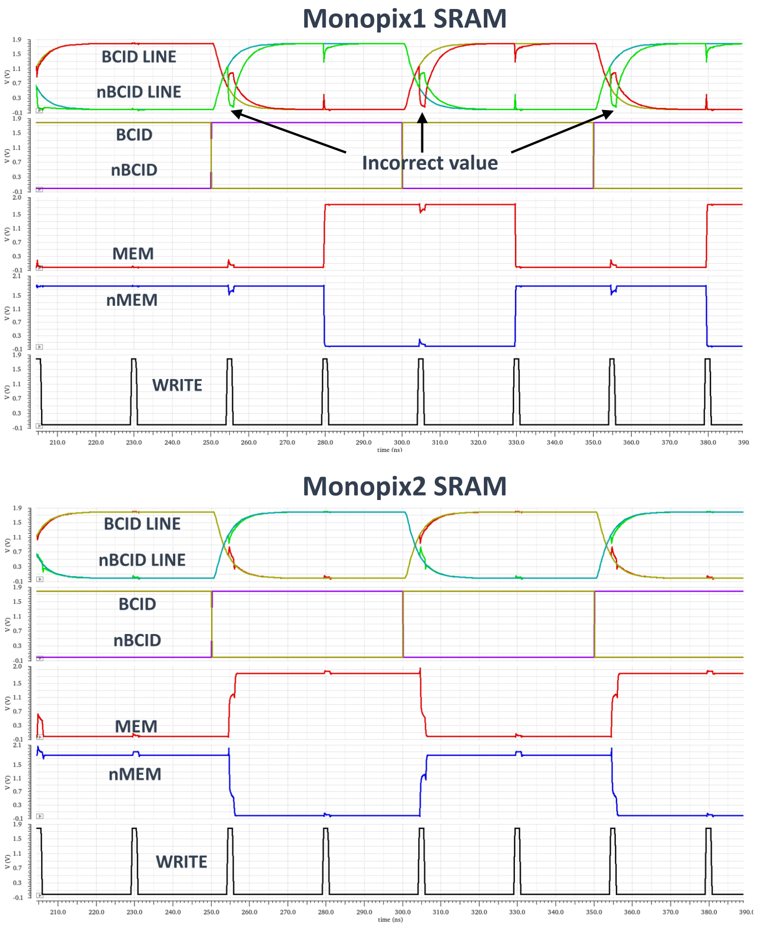


Figure 4.102: Simulation of simultaneous LE/TE memory write operations at the 25 topmost pixels of a column. The TJ-Monopix2 SRAM cell design prevents bit flipping due to the loading effect on the BCID line by disconnecting power for the duration of the write pulse ("sensing" operation).

on a current source with configurable current (I_{READ}) that pulls-down the bit line combined with a switch that enables access to the data-bus according to the pixel READINT signal.

4.5.1.4 BCID delay compensation

As mentioned, the RC effect of the long column combined with aggressive column-bus routing result in a high signal propagation delay. A simulation of the BCID LSB propagation time across the column is shown in Fig. 4.104. The delay between the top and bottom of the column is equal to 16.5 ns, which is approximately 4 times higher than the specification of the maximum BCID timing variation ≊ 4 ns. In order to reduce the BCID timing dispersion across the pixel matrix, a compensation scheme has been implemented that adds a delay to the HIT pulse which is approximately equal to the local BCID delay, and therefore ensures that the hit timing information remains accurate.

The HIT pulse delay is realized by a circuit that contains a chain of 4 inverter stages, shown in Fig. 4.105. In order to achieve a high propagation delay time up to 16.5 ns while keeping the area small, the first two stages are comprised of current-starved inverters that are biased by a configurable

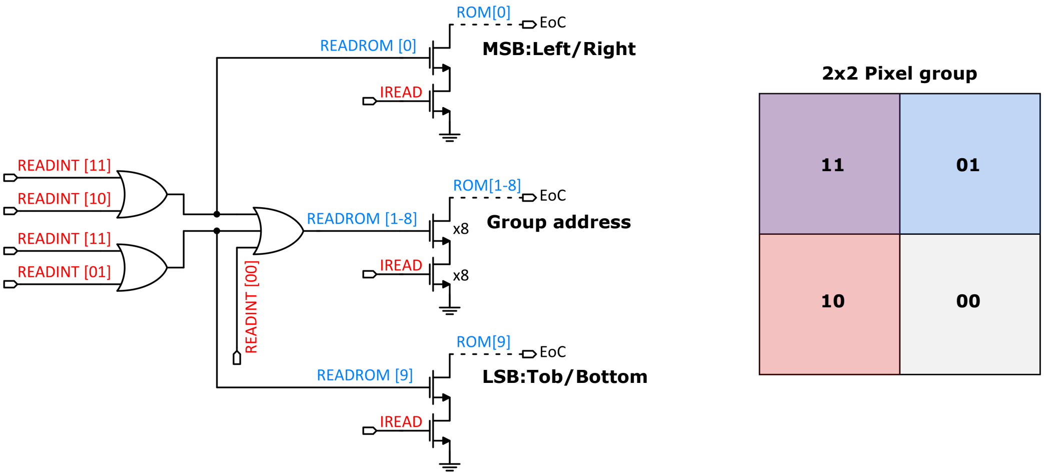


Figure 4.103: Schematic of the pixel address ROM implemented in TJ-Monopix2. The pixel position withing the 2 × 2 core is encoded in order to save area.

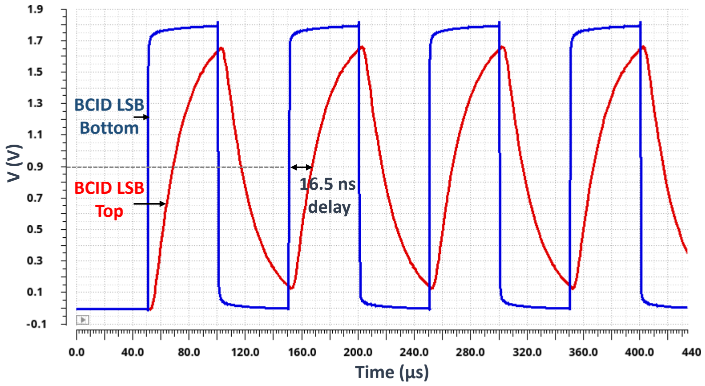


Figure 4.104: Simulation of BCID time-stamp propagation across a pixel column of TJ-Monopix2. The waveforms correspond to one of the two differential components of the BCID LSB bit line at the bottom and top of the column. The propagation delay is equal to approx. 16.5 ns.

current I_{DEL}. The delay of each current-starved inverter stage is proportional to the current I_{DEL} and the total capacitance at its output, which depends on the transistor area:

$$t_{\text{del}} \approx I_{\text{DEL}} \cdot \left(C_{\text{gd}_1} + C_{\text{gd}_2} + C_{\text{gg}_3} + C_{\text{gg}_4}\right) \approx I_{\text{DEL}} \cdot WL_{M_1-M_6} \tag{4.58}$$

where the first stage (transistors M_1, M_2, M_7, M_8) has been used as an example. $C_{\text{gd}_1}, C_{\text{gd}_2}$ is the gate-to-drain capacitance of the first stage transistor switches M_1 and M_2, while $C_{\text{gg}_3}, C_{\text{gg}_4}$ is the total input (gate) capacitance of the following stage.

Most of the total propagation delay is essentially contributed by the first two current-starved inverter stages. The third inverter is included in order to provide a capacitive load for the second stage, equal to the capacitance that is added by the second stage to the output of the first stage. Therefore, transistors $M_1 - M_6$ are designed with the same dimensions. The final inverter is included in order to restore the HIT pulse polarity and buffer the output which is then connected to the R/O logic edge detector. In order to adjust the HIT pulse delay across the column, the gate area of $M_1 - M_6$ ($WL_{M_1-M_6}$) is scaled accordingly. The column is split to 32 groups, consisting of 16 pixels each, where the same HIT pulse delay is applied. The delay group size is a compromise between the maximum BCID timing dispersion within the same group and the amount of different delay circuit variations that have to be created.

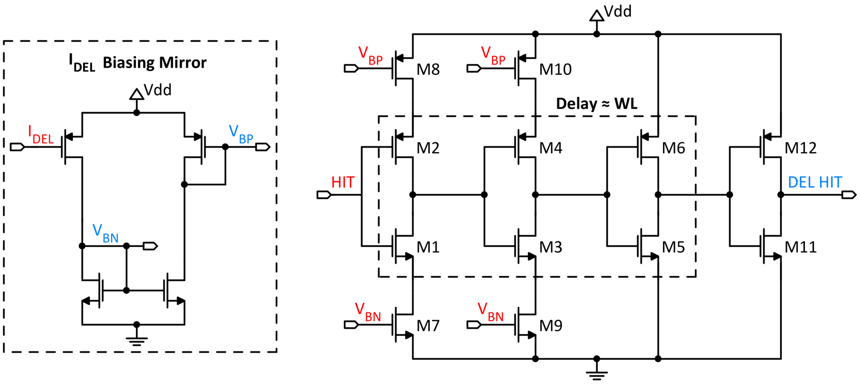


Figure 4.105: Schematic of the HIT pulse delay circuit used to compensate for the BCID propagation delay across the column. The highlighted inverter transistor pairs are scaled to adjust the delay of each 16-pixel group.

Simulation results of the BCID timing at each 2 × 2 pixel core across a double column before and after compensation are shown in Fig. 4.105. Although the maximum total BCID timing variation is appox. equal to 1 ns for the selected delay group size of 16 pixels, the effect of transistor mismatch has to be considered and is included in the simulation. The minimum and maximum traces correspond to the BCID timing after compensation, where the value at $\pm 3\sigma$ (99.7% probability) of the delay distribution due to mismatch is added. Because the variance of the current provided by transistors (M_7, M_8, M_9, M_{10}) is the same for all delay groups, while the capacitive load ($\approx WL_{M_1-M_6}$ is increased from bottom to top of the column, the delay variance scales with the mean delay value (i.e. the same current variation will result in a larger delay variation at the column top where the delay is larger). The maximum total BCID timing variation after the inclusion of mismatch effects is approximately equal to 4 ns, which is equal to the ATLAS specification.

While the relative delay between each group is hard-coded by sizing transistors $M_1 - M_6$, the absolute delay is can be adjusted by I_{DEL}, which is generated by the biasing DAC, in order to correct for process, temperature and voltage (PVT) variations. Similar to the readout current (I_{READ}), an additional I_{DEL} mirror stage in included at the column bottom in order to compensate for the body effect due to backbias.

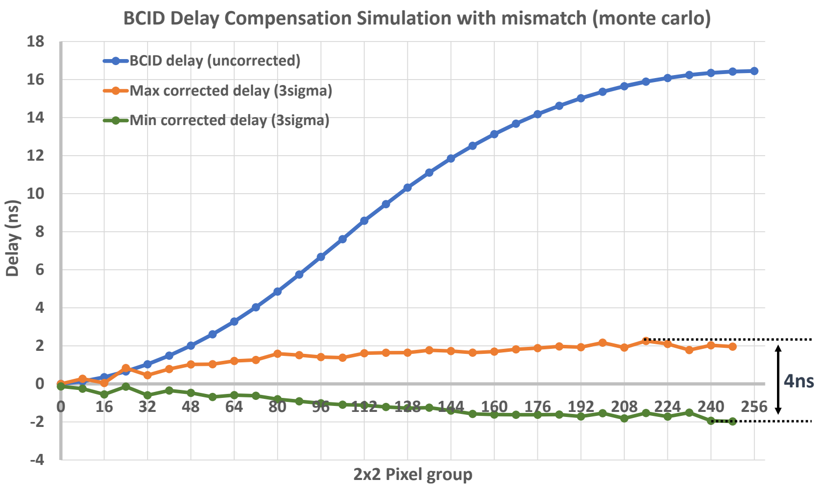


Figure 4.106: Simulation of the BCID timing delay at each 2×2 pixel core the across a double column of TJ-Monopix2 before and after compensation is applied. The compensated delay includes transistor mismatch effects ($\pm 3\sigma$ of the distribution). The maximum BCID timing dispersion after correction is approx. equal to 4 ns.

4.5.1.5 Test features

As in the case of TJ-Monopix1 (see section 4.3.1.3), the additional test features include artificial hit injection, a HIT-OR logic that provides access to the pixel discriminator output and the ability to monitor the sensor and front-end analog output transient voltage using additional monitoring pixels that include an analog buffer.

The hit injection circuit is similar to the one shown in Fig. 4.42. In order to improve the linearity of the injected charge as a function of the injection pulse amplitude ($V_H - V_L$), the PMOS switches M_1 and M_2 have been replaced by transmission gates that are composed of both transistor types (NMOS, PMOS). The injection capacitance (C_{inj}) is equal to 230 aF. In contrast to TJ-Monopix1, C_{inj} is the same for both DC-coupled and AC-coupled HV front-end variants due to the improved MOM capacitor structure layout that allows C_{inj} to be placed close to the collection electrode node while being shielded from the front-end input node. The injection step of TJ-Monopix2 is finer compared to TJ-Monopix1 due to the higher voltage DAC resolution (LSB=7.03 mV). The injected charge Q_{inj} can be calculated using equation 4.44 as follows:

$$Q_{\text{inj}} = 1.4375\ {}^{e^-}\!/_{\text{mV}} \cdot 7.03\,\text{mV} \approxeq 10.1\ {}^{e^-}\!/_{DU} \tag{4.59}$$

In contrast to the wired-OR logic used in TJ-Monopix1, HIT-OR is implemented using full CMOS digital gates in a NAND-NOR logic configuration that minimizes the gate area and signal propagation delay. The gates used by the HIT-OR circuit are designed to be balanced in order to avoid distortion of the HIT-OR pulse width (ToT). Pixels can be selected for HIT-OR using a two vector projection scheme with column and row enable.

Table 4.10 provides a list of the analog monitoring pixels that are included in TJ-Monopix2. Four cells have been placed at each side of the matrix. The front-end output is monitored using the analog buffer shown in Fig. 4.45. Apart from all four variations of the TJ-Monopix2 front-end, a copy of the TJ-Monopix1 front end[8] has been included for comparison. Additionally, a sensor monitoring cell that allows probing of the collection electrode transient voltage has been added in order to provide the ability to characterize the sensor raw performance. The sensor monitoring circuit consists of a series of four source-follower buffers and is shown in Fig. 4.107. Finally a cell that monitors the V_{CASN} bias voltage is included for debugging purposes.

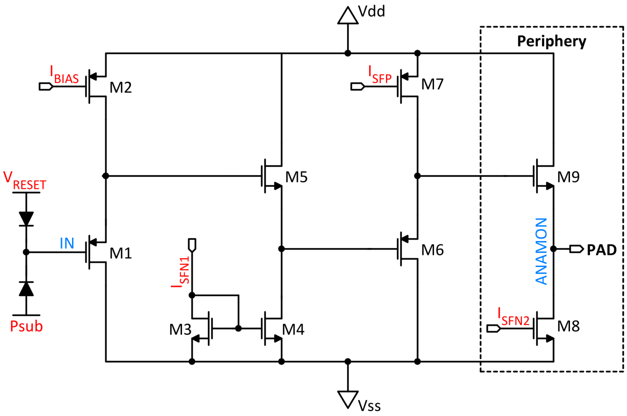


Figure 4.107: Schematic of the analog buffer used to monitor the sensor voltage at the collection electrode.

Table 4.10: Analog monitoring front-end cells included TJ-Monopix2.

Left side	Right side
TJ-Monopix1 FE	V_{CASN} monitoring
TJ-Monopix2 FE	HV TJ-Monopix2 FE
TJ-Monopix2 cascode FE	HV TJ-Monopix2 cascode FE
Sensor monitoring	HV sensor monitoring

4.5.2 Chip design and architecture

The chip architecture of TJ-Monopix2 is shown in Fig. 4.46. The pixel matrix is organized in 256 double columns, each consisting of 1024 pixels that share the same column-bus. Pixel hit data (10-bit address, 7-bit LE time-stamp, 7-bit TE time-stamp) is sensed by a current mode data-bus readout circuit at the EoC. The EoC block additionally contains BCID and control signal buffers and the EoC readout logic which is part of the synthesized digital chip bottom. Since no trigger memory is

[8] Because an I_{RESET} bias required by the PMOS input reset circuit is not available in TJ-Monopix2, the TJ-Monopix1 front-end monitoring cell uses a diode input reset.

implemented, the EoC logic is used to arbitrate the readout at the double column level and append the 8-bit column address ($2^8 = 256$ double columns) to the pixel data as described in section 4.3.2.

Two selectable readout modes have been implemented. The default mode is continuous (direct) readout i.e. hit data arriving at the EoC is immediately transmitted to the DAQ without being stored in the periphery. Additionally, a shutter mode is available that is useful to test the TJ-Monopix2 matrix suitability for certain imaging applications such as proton computed tomography. In general, the shutter mode can be used to capture a complete "snapshot" (within a certain time duration) when the hit rate significantly exceeds the continuous readout capability and works as follows: Initially, the in-pixel R/O logic is held in a reset state and no hits can be recorded. The reset is then released and the detected hits are stored in the pixel memory, while no READ operation is allowed. After a configurable time t_{shutter}, the FREEZE signal is asserted throughout the whole matrix, and therefore any new hits are discarded. Subsequently, the readout operation is initiated and hit data that has been recorded during t_{shutter} is transferred to a FIFO memory in the chip periphery. The FIFO memory is used as an intermediate buffer that allows to "empty" the matrix quickly without being limited by the output data link speed.

The digital chip bottom has been designed for easier system integration and is significantly more complex (> 7500 lines of HDL code) compared to TJ-Monopix1. Configuration and control is performed through an LVDS 160 MBps command data stream that is interpreted by a command decoder which is a modified version of the decoder used in the RD53A [92, 93] LHC phase-II upgrade prototype ROIC. Since no Clock Data Recovery (CDR) is included, the 160 MHz command synchronization clock has to be additionally provided by the DAQ system. The 40 MHz base clock that is used to drive the 7-bit BCID gray counter is generated by dividing the command clock by 4. Output data is 8b/10b encoded using an encoder derived from the FE-I4 chip [28] in order to achieve DC balancing and easier synchronization with the DAQ FPGA receiver (frame based data transmission). Apart from the hit data, there is the possibility to read back the configuration register values for debugging. After encoding, the output data is serialized and transmitted off-chip by a fast LVDS transmitter rated for speeds up to 5 GBps. However, the output data path including the serializer has been synthesized for a maximum speed of 320 MBps.

A R/O controller with configurable readout sequence timing has been also included on-chip. Therefore, TJ-Monopix2 can be essentially operated using only the uplink (command clock and data), the serializer clock and the downlink (output data stream). As a backup, the internal R/O controller can be bypassed and the control signals can be provided externally from the DAQ FPGA through CMOS I/O pads. Furthermore, every LVDS I/O interface is supplemented by a redundant CMOS I/O pad. LVDS or CMOS communication can be selected by an external static signal.

4.5.2.1 Current mode column-bus readout

The column data-bus speed of TJ-Monopix2 is compromised by the large column height and the design and layout modifications that are necessary to achieve a small pixel size. Single-ended data transmission results in a reduction of the signal amplitude by approximately 2 times. Furthermore, due to the minimum metal line width and spacing the bit-line resistance and capacitance per unit length are relatively large. The high time constant of the bit-line elementary component ($\tau = RC$), combined with its length of approx. 17 mm result in a high data-bus propagation delay (or equivalently low bandwidth). Additionally, crosstalk between neighboring bit-lines is increased due to the small spacing (large coupling capacitance) and single-ended transmission. As a result, the overall signal to

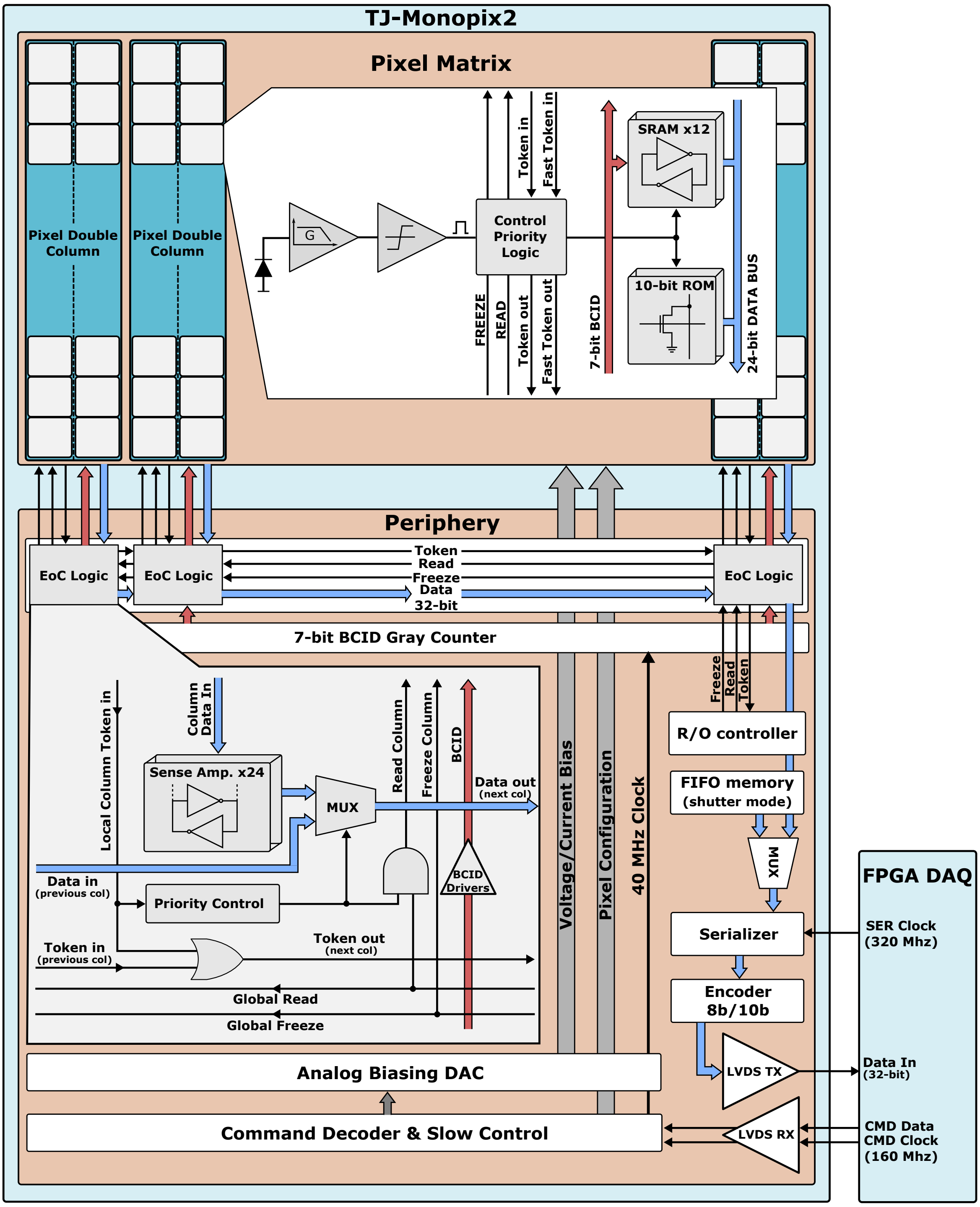


Figure 4.108: Chip Architecture of TJ-Monopix2

noise ratio (ability to distinguish the bit value) is reduced because the differential signal amplitude between two lines corresponding to opposite bit values becomes smaller.

In order to reduce the data-bus delay, or equivalently increase the signal amplitude for a given READ pulse duration, a current-mode sensing scheme has been implemented at the EoC instead of a standard voltage-mode sense amplifier. The benefit of low impedance current-mode sensing becomes evident by examining the equivalent sensing circuit of Fig. 4.109. Data is transmitted by the in-pixel current source I_{READ} (see section 4.5.1.3) which has an equivalent parallel resistance R_B. The bit-line is modeled by a series of distributed RC elements and R_L is the sense amplifier input resistance at the receiving end (EoC) which terminates the line. The transmission delay transfer function is given by [94]:

$$\delta t = \frac{R_T C_T}{2}\left(\frac{R_B + \frac{R_T}{3} + R_L}{R_B + R_T + R_L}\right) + R_B C_T\left(\frac{R_L}{R_B + R_T + R_L}\right) \tag{4.60}$$

where R_T and C_T are the total bit-line resistance and capacitance. Since an ideal voltage sense amplifier has close to infinite input resistance while an ideal current sense amplifier has zero input resistance, the delay equation for each case is simplified to:

$$\begin{gathered}\delta t_{voltage} = \frac{R_T C_T}{2}\left(1 + \frac{2R_B}{R_T}\right) \\ \delta t_{\mathrm{current}} = \frac{R_T C_T}{2}\left(\frac{R_B + \frac{R_T}{3}}{R_B + R_T}\right)\end{gathered} \tag{4.61}$$

Because typically $R_B \gg R_T$ (due to the high MOSFET output resistance r_{ds}), it is obvious that the current-mode delay is significantly smaller than the voltage-mode delay[9]. The lower limit of current sensing delay is equal to $R_T C_T/2$ which corresponds to the inherent propagation speed of the line (Elmore delay [95]) [10]. Intuitively, the reason of the small delay associated with current-mode sensing is that as a result of the low termination resistance R_L, the voltage across the line is small and therefore the distributed capacitances are charged to the final value ($I_{\mathrm{READ}} \cdot R_L$) in a relatively short time.

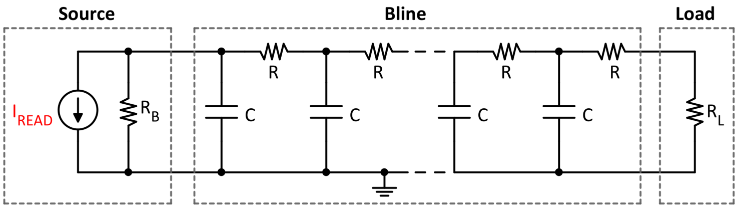


Figure 4.109: Equivalent bit-line sensing circuit used to derive the data transmission delay function.

A comparison of current and voltage sensing is performed by a simulation of the readout of a pixel at the column top using two termination resistance values, $R_L = 1\,\mathrm{k\Omega}$ and $R_L = 1\,\mathrm{M\Omega}$ respectively.

[9] Reducing R_B in order to improve the voltage-mode delay is impractical because it would result in a reduction of the signal amplitude.

[10] The Elmore delay is also equal to $R_T C_T/2$ in the case of voltage sensing ($R_L \rightarrow \infty$) if a voltage source is used ($R_B \rightarrow 0$). However, a voltage source, e.g. strong digital buffer, would result in a very high transient current in the pixel.

Two neighboring bit-lines have been used in the simulation in order to include the effect of coupling (crosstalk). The first line is pulled down (transmits a bit value of 1) while the second line is not connected to I_{READ} (transmits a bit value of 0). Simulation results are shown in Fig. 4.110. The readout speed (READ pulse width) is essentially determined by the time that is required to develop an adequate differential signal amplitude between the two lines in order to distinguish the correct bit value. As can be observed, in the case of current-mode sensing the delay is smaller and a higher signal amplitude is developed which results in faster data transmission (higher column bandwidth).

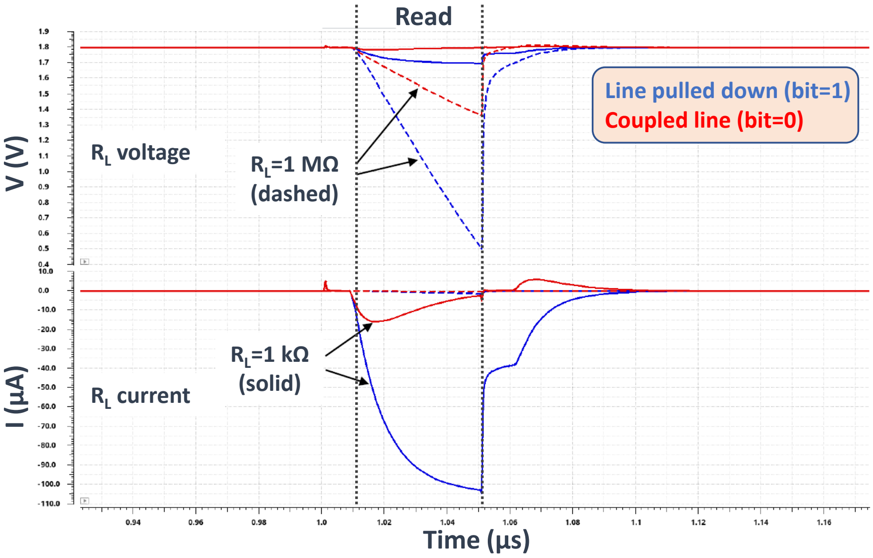


Figure 4.110: Readout simulation of a pixel at the column top using current (solid lines) and voltage (dashed lines) sensing. The voltage and current at the receiving end (R_L) of two coupled neighboring bit-lines are measured. A higher signal amplitude is achieved by current sensing.

The schematic of the complete data-bus readout circuit is shown in Fig. 4.111. The bit-line current is sensed by the diode-connected PMOS transistor M_5 and is subsequently mirrored to M_8 with a mirror ratio of 4 that provides current amplification. Transistors M_8 and M_9 form a common source stage that works as a current comparator (similar to the front-end discriminator) and compares the bit-line current, amplified by 4 times, to the configurable current I_{COMP}. Due to the high output resistance of the current comparator ($r_{\mathrm{ds}_8} /\!/ r_{\mathrm{ds}_9}$), a large voltage amplitude is developed, which is compatible with digital CMOS levels. From a different perspective, the sensed bit-line current (see Fig. 4.110) is amplified and driven to a large resistance (drain of M_8) in order to generate a high amplitude voltage signal. The current comparator output is digitally buffered and is subsequently latched by a flip-flop which is triggered by the falling edge of the READ signal.

Between each read cycle, the current sensing transistor M_5 is disconnected by switching off M_4 and the bit-line is precharged by transistor M_3 to an adjustable voltage V_{PC} which is supplied externally. Additionally, the gate of M_5 and M_8 is pulled up to V_{dd} by M_7 in order to ensure that no current is flowing to M_9. Therefore, the static power consumption is zero. During the read phase, M_3 and M_7 are switched off and the bit-line is connected to M_5. If a bit value of 1 is transmitted, the line is pulled down by the in-pixel transistor M_1, which sinks a current equal to I_{READ}. As the bit-line is discharged, the sensed current increases and becomes equal to I_{READ} when the bit-line voltage settles to its final value ($V_{\mathrm{BL_{fl}}}$) that is given by:

$$V_{\mathrm{BL_{fl}}} \approxeq V_{\mathrm{dd}} - V_{\mathrm{gs}}^{M_5} = V_{\mathrm{dd}} - \sqrt{\left(\frac{L}{W}\right)_{M_5} \frac{1}{\mu_n C_{\mathrm{ox}}} I_{\mathrm{READ}}} - V_{\mathrm{TH}}^{M_5} \tag{4.62}$$

where V_{gs} is the gate-to-source voltage, L, W are the transistor dimensions, μ_n is the electron mobility, C_{ox} is the gate capacitance per unit area and V_{TH} is the transistor threshold. If a bit value of 0 is transmitted, M_1 is disconnected from the bit-line by M_2 and therefore no current is sinked. If the bit-line voltage (at the receiving end) becomes lower than $V_{\mathrm{TH}}^{M_5}$, e.g. due to coupling, a small current will be drawn by M_5 until the line is pulled back up to approx. $V_{\mathrm{dd}} - V_{\mathrm{TH}}^{M_5}$.

The small-signal input resistance of M_5 is approximately equal to:

$$r_i^{M_5} \approxeq \frac{1}{g_{m_5}} = \frac{1}{\sqrt{2\mu_n C_{\mathrm{ox}} \left(\frac{L}{W}\right)_{M_5} I_D^{M_5}}} \tag{4.63}$$

and therefore depends on the drain current of M_5 ($I_D^{M_5}$), which at the start of the read phase depends on the value of V_{PC}. In order to reduce the sense amplifier input resistance, M_5 should be designed with a high W/L ratio.

The data-bus readout is tuned by adjusting I_{READ}, V_{PC} and I_{COMP}. The readout speed can be improved by increasing the pull-down current (I_{READ}). However I_{READ} should not be set too high in order to avoid crosstalk between the pixel analog front-end and digital readout. The pre-charge voltage (V_{PC}) should be: a) lower than V_{dd} in order to decrease $r_i^{M_5}$ and additionally reduce the time needed to reach $V_{\mathrm{BL_{fl}}}$ (for a pulled-down line) and b) high enough to avoid a large current being drawn by a disconnected bit-line (bit value=0) while accounting for coupling. Typically V_{PC} is set to approximately 1.6 V. It can also be shorted to V_{dd}, if required, with the cost of reduced speed. The comparator current I_{COMP} should be set high enough to avoid errors due to coupling.

A simulation of the TJ-Monopix2 data-bus readout is shown in Fig 4.113. Data is transmitted by a pixel at the column top, which represents the worst case. During the first read cycle a bit value of 1 is transferred, while in the second read cycle the bit value is 0. For the selected I_{READ} current of 100 µA the data transmission delay is approximately equal to 15 ns. However, the propagation delay of the READ pulse to the column top, which is approximately equal to 10 ns should be added to the minimum READ pulse duration. In order to account for noise and mismatch, the nominal READ pulse width is set to 50 ns (2 clock cycles) which corresponds to a maximum column bandwidth of 10 MHz.

4.5.2.2 Double column simulation

In order to verify the correct double column operation and investigate for possible issues due to crosstalk, a simulation has been performed that includes a detailed RC model of the power grid. The double column is placed in the middle of the matrix, which represents the worst case in terms of voltage drop. Simulation results are shown in Fig. 4.113. Four hits in total are injected at various pixel locations, and the correct data is received at the EoC. The BCID propagation delay is compensated by the HIT delay circuit. Crosstalk during the read phase can be observed at the front-end pre-amplifier output which is mainly attributed to coupling through the common substrate (p-well). However, the resulting output voltage fluctuation is small (approx. 15 mV) and does not have a significant impact

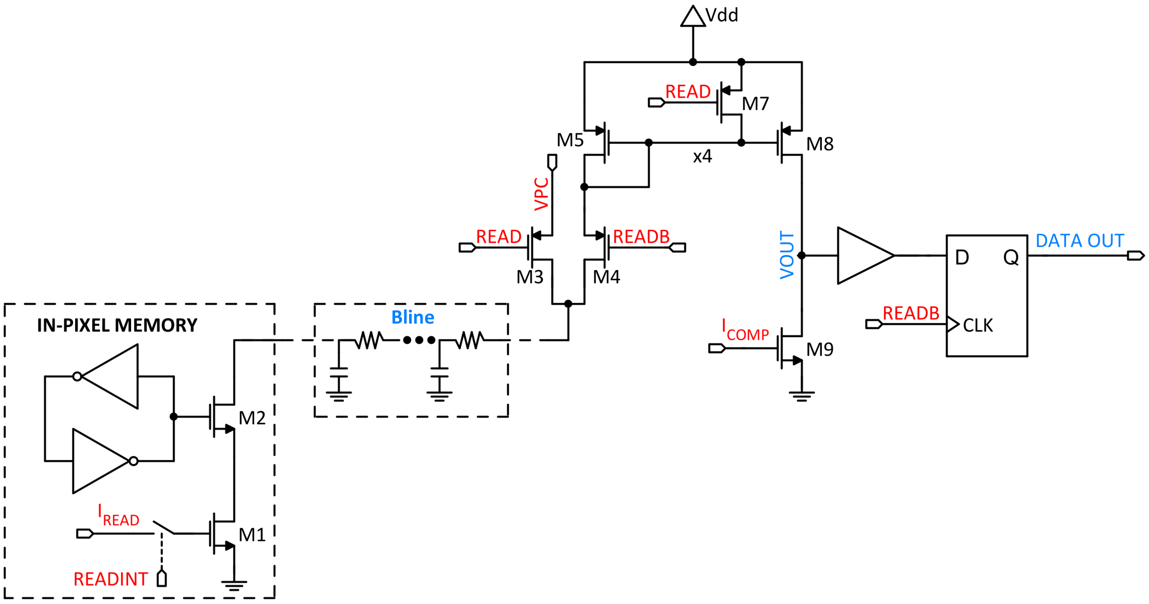


Figure 4.111: The current-mode data-bus readout scheme used in TJ-Monopix2.

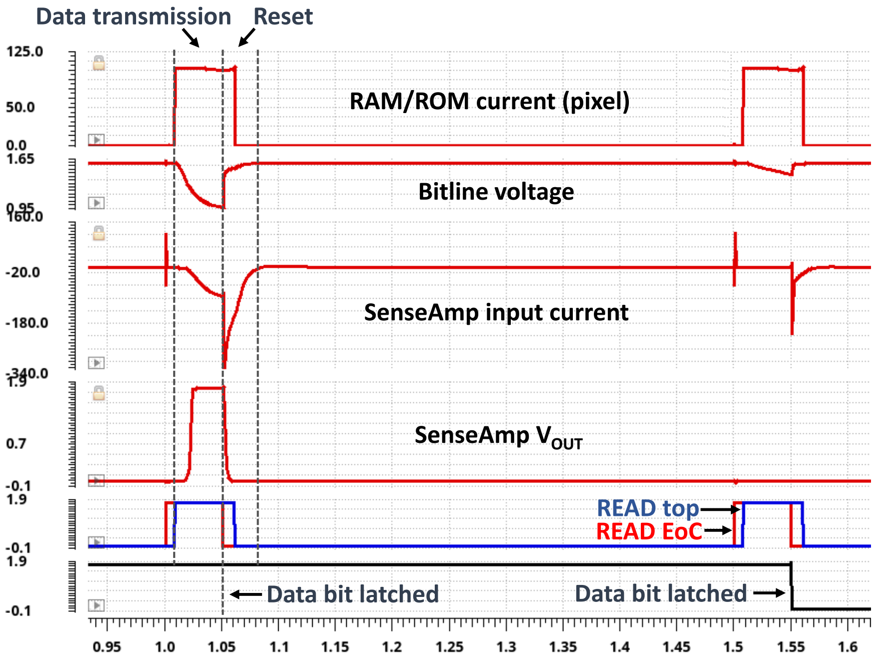


Figure 4.112: Post-layout simulation of a data bit transmission for the topmost pixel to the EoC using current mode sensing.

on the pixel performance.

4.5.2.3 Analog bias generation

An 8-bit DAC, derived from the mini-Malta chip [67], is used to generate the analog current and voltage biases. It is designed to be modular, in order to be easily adapted to different matrix sizes, and consists of a main current DAC (IDAC) and voltage DAC (VDAC) unit and the final mirrors and buffers which are distributed below the pixel matrix. The generated voltage values have good linearity with respect to the DAC code setting. The maximum (absolute) differential non-linearity (DNL) is equal to 4.5% while the maximum integrated non-linearity (INL) is equal to 0.2%. Furthermore, its performance after irradiation has been tested up to 80 Mrad TID and no significant degradation was observed.

The architecture of the IDAC is shown in Fig. 4.114. The reference current, which is equal to 20 nA, is mirrored to 256 current sources that are controlled by the IDAC code (SET). The IDAC code ranges from 1 to 256 and is thermometer encoded by a dedicated encoder that is included in the DAC circuit. The total current, $I_{\mathrm{REF}} \cdot \mathrm{SET}$, is summed at the input of a cascode current mirror in order to generate the bias voltages that drive the mirror at the EoC (below the pixel matrix). A PMOS transistor is used to provide isolation from fluctuations of the ground rail voltage between the main unit and the distributed mirror stages. The resolution (LSB) and range of each bias current is determined by the EoC mirror ratio.

Similar to TJ-Monopix1 (see section 4.3.2.3), the voltage drop across the matrix analog power grid (V_{DDA}) is compensated by using the local analog matrix supply to power the final EoC DAC mirror stage. Additionally, pixel columns are grouped in order to reduce the bias current dispersion due to mismatch of the mirror transistors. Due to the systematic threshold variations that were observed in TJ-Monopix1 as a result of biasing mismatch effects, the group size has been increased to 32 columns.

The VDAC architecture is similar to Fig 4.53. The resistor divider is included in the main VDAC unit and the source-follower buffers that are required for bias voltages with high current requirements ($V_{\mathrm{RESET}}, V_H, V_L$), are distributed at the EoC. The VDAC code (SET) ranges from 0 to 255 and is internally one-hot encoded. Therefore, the voltage DAC LSB is equal to $1.8\,\mathrm{V}/256 \cong 7.03\,\mathrm{mV}$. Each bias voltage/current can be monitored or externally supplied using a set of dedicated pads. The monitor/override functionality is enabled by setting the corresponding configuration bits.

4.5.2.4 Data transmission

A set of CMOS and LVDS I/O are used to communicate with the DAQ FPGA. CMOS I/O pads are used for static signals (e.g. reset), debugging, and as a backup to each LVDS transceiver. They are normally operated at frequencies up to 40 MHz. LVDS is used for high frequency signals such as the command input and clock, the data output and the serializer clock. The LVDS transceiver circuit has been imported from TJ-MALTA1 and is referred to as pseudo-LVDS for the ATLAS Pixel Apparatus (LAPA) [96]. The output driver has been rated for speeds up to 5 GBps and is based on the TIA/EIA 644 LVDS transmission standard. It is optimized to operate with 400 mV differential amplitude, which corresponds to a steering current of 4 mA over an 100 Ω termination resistor and consumes approx. 30 mW. The common mode voltage (V_{CM}) is set to 0.8 V, instead of 1.2 V specified by the standard, and is regulated by a common-mode feedback circuit.

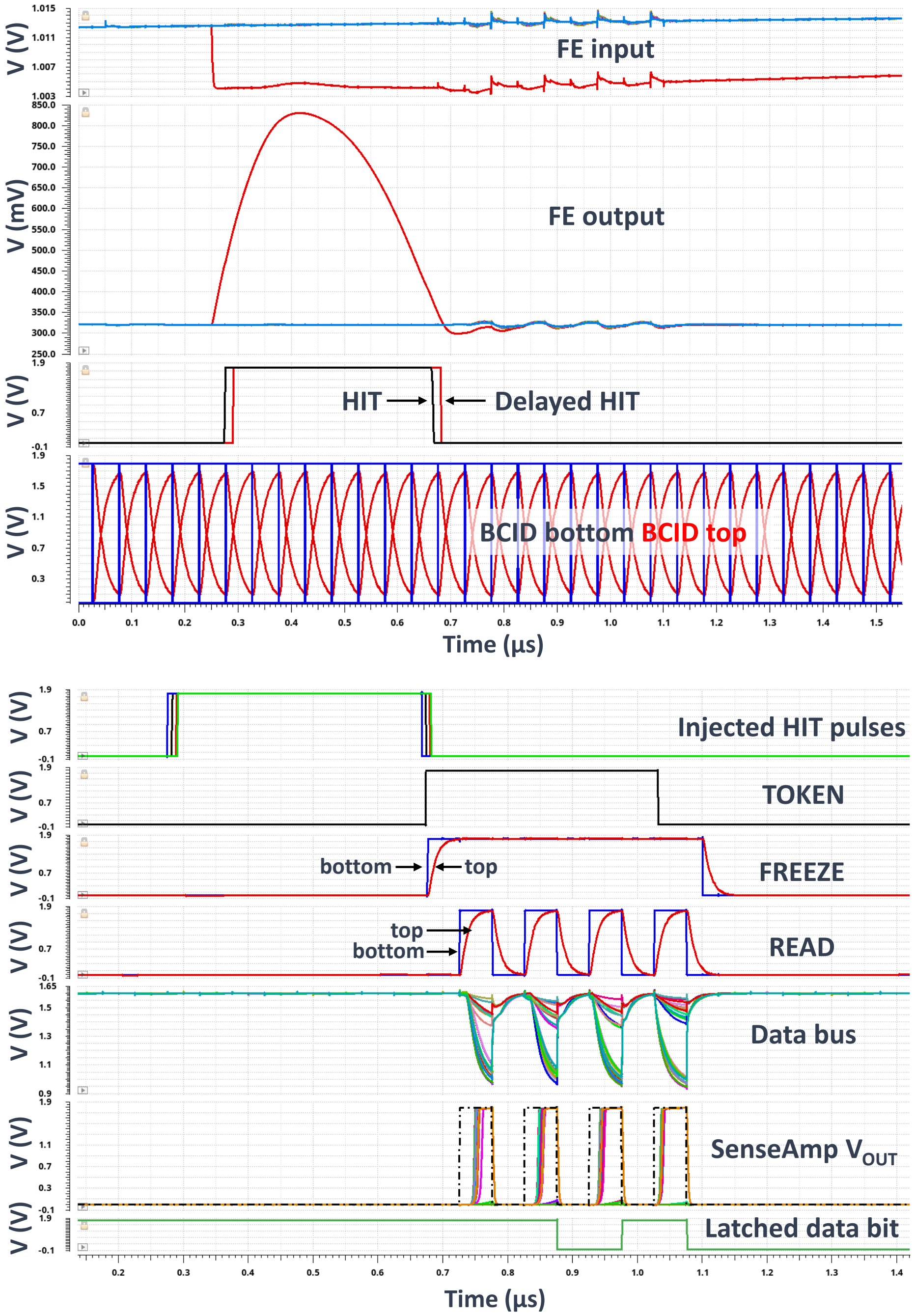


Figure 4.113: Post-layout simulation of a full double column in the middle of the matrix including a power grid RC model. Four hits are injected and the corresponding hit data is correctly read out.

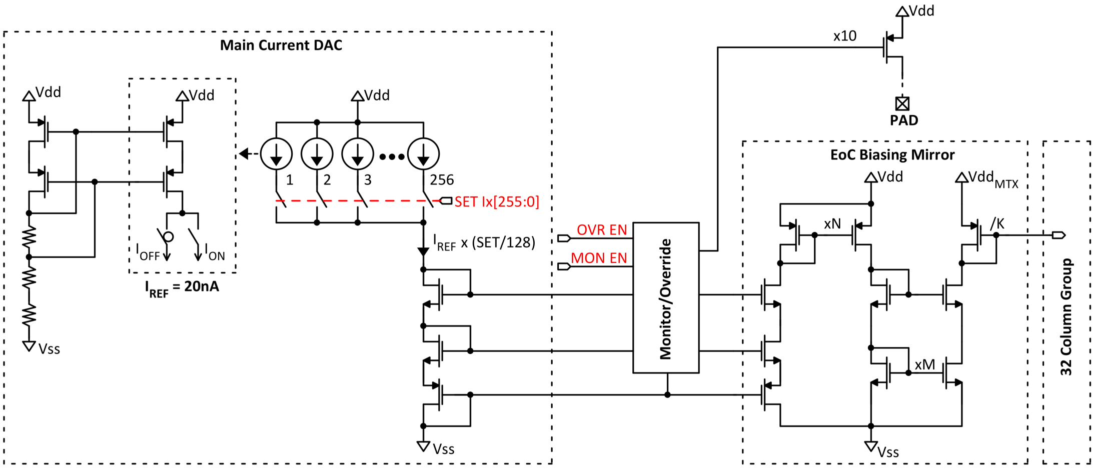


Figure 4.114: Schematic of the bias current DAC used in TJ-Monopix2.

The main driver is based on an H-Bridge scheme shown in Fig. 4.115. The output (steering) current I_L is equal to the difference between the ON (I_{on}) and OFF (I_{off}) current values and is switched depending on the input data. A non-zero OFF current is used to improve the switching speed and set the operating point in the OFF state of each branch. In order to reduce the performance loss when a low bandwidth channel is used (e.g. long or low mass cables), pre-emphasis can be applied with the help of a capacitively coupled charge injection circuit which is composed of a digital inverter and a coupling capacitor. The pre-emphasis strength is configurable by selecting a number of the available 16 independent charge injection stages.

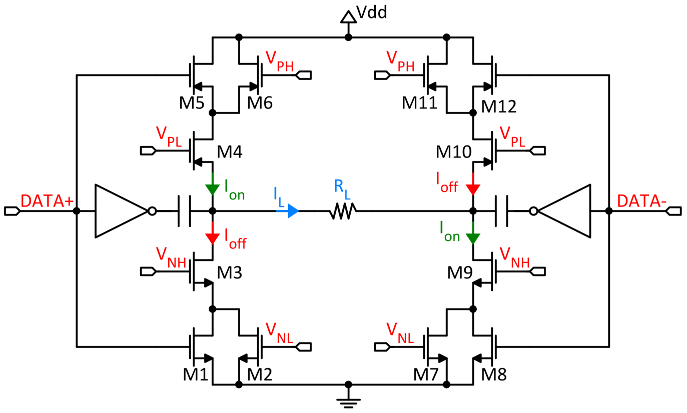


Figure 4.115: Schematic of the LVDS driver. Pre-emphasis can be applied by a capacitively coupled digital inverter [96].

CHAPTER 5

Conclusions and outlook

Detector development is currently driven by the upcoming High Luminosity upgrade of the Large Hadron Collider (HL-LHC) in order to address the unprecedented requirements in terms of hit rate and radiation intensity. The new generation of pixel detectors should combine a small pixel size, high radiation tolerance, low mass, low power consumption and be able to be produced in large volumes with low cost. While hybrid pixels are the state-of-the-art for tracking in high rate and high radiation environments, Depleted Monolithic Active Pixel Detectors (DMAPS), that combine the sensor and readout electronics in the same silicon crystal have emerged as a promising alternative due to their low cost, low material and reduced production complexity.

Two large scale, small collection electrode DMAPS prototype chips, called TJ-Monopix1 and TJ-Monopix2, have beed developed to target the requirements of the HL-LHC ATLAS Inner Tracker (ITk) outer layer: 25 ns timing resolution, NIEL fluence of $10^{15}\,n_{eq}/\mathrm{cm}^2$, TID dose up to 80 Mrad and particle rate of about $100\,\mathrm{MHz/cm}^2$. TJ-Monopix chips are fabricated on a TowerJazz 180 nm CMOS imaging process which has been modified in order to achieve full depletion of the sensitive layer. They feature a standalone, fast "column-drain" readout architecture and aim to combine low noise and low power consumption due to the very small sensor capacitance of approx. 3 fF with fast charge collection by drift, hence high radiation tolerance.

TJ-Monopix1 is a half-scale chip ($1 \times 2\,\mathrm{cm}^2$) and is composed of 224×448 pixels with $36 \times 40\,\mu\mathrm{m}^2$ size and the supporting periphery blocks. The small detector capacitance is exploited by a compact, power-efficient analog front-end that is based on a voltage amplifier and yields a low ENC of approx. $10\,e^-$ and analog power consumption equal to $70\,\mathrm{mW/cm}^2$. Extensive measurement and characterization has demonstrated the full functionality and high analog performance of TJ-Monopix1. However, after irradiation to $10^{15}\,n_{eq}/\mathrm{cm}^2$, the detection efficiency is reduced to approx. 70% due to the low lateral field at the pixel edges and the relatively high threshold. After applying improvements to the modified process layout in order to enhance the lateral field, the efficiency increased to approx. 85%. An even higher efficiency of about 97% has been achieved by using a thicker Czochralski substrate material instead of the standard 25 µm epitaxial layer due to the resulting higher signal.

In order to further increase performance and address the shortcomings of TJ-Monopix1, a full scale ($2 \times 2\,\mathrm{cm}^2$) next generation prototype, called TJ-Monopix2 has been designed. The active area consists of 512×512 pixels with $33.04 \times 33.04\,\mu\mathrm{m}^2$ size. Apart from the smaller pixel, which leads to faster charge collection, the threshold has been reduced by a factor of 3 as a result of the improved analog front-end circuit. The digital periphery packs more functionality and allows easy system integration and high speed LVDS communication. TJ-Monopix2 has been recently fabricated (Q1 2021) and is fully operational. Initial testing to characterize its performance has currently started.

A summary of the TJ-Monopix development line specifications is given in Table 5.1. This work has shown very promising results towards small collection electrode DMAPS for LHC type high rate and high radiation environments, which has been previously considered as a real challenge. Furthermore, TJ-Monopix is ideally suited for other experiments, such as lepton colliders (e.g. future upgrades if the Belle II detector) where the small pixel size, low material and low power consumption it offers is of utmost importance.

Table 5.1: TJ-Monopix DMAPS prototype chip specifications and performance summary

	TJ-Monopix1	TJ-Monopix2
Chip size	$1 \times 2\,\mathrm{cm}^2$	$2 \times 2\,\mathrm{cm}^2$
Matrix arrangement	224×448 pixels	512×512 pixels
Pixel size	$36 \times 40\,\mu\mathrm{m}^2$	$33.04 \times 33.04\,\mu\mathrm{m}^2$
Total matrix power (analog & BCID)	$130\,\mathrm{mW/cm}^2$	$170\,\mathrm{mW/cm}^2$
ToA/ToT resolution	6-bit	7-bit
Noise (ENC)	$\approxeq 9\,e^-$	$\approxeq 5\,e^{-*}$
Threshold dispersion	$\approxeq 30 - 35\,e^-$	$\approxeq 5 - 10\,e^{-*}$
Operational threshold	$\approxeq 300\,e^-$	$\approxeq 100\,e^{-*}$
In-time threshold	$< 350\,e^-$	$< 170\,e^{-*}$
Efficiency at $10^{15}\,n_{eq}/\mathrm{cm}^2$ NIEL (25 µm epi, n$^-$gap/extra-dpw)	$\approxeq 85\%$	$> 97\%^*$
Efficiency at $10^{15}\,n_{eq}/\mathrm{cm}^2$ NIEL Cz substrate, n$^-$gap/extra-dpw)	$\approxeq 97\%$	$> 97\%^*$

* Expected values

Bibliography

[1] C. Wilson, *On the Cloud Method of Making Visible Ions and the Tracks of Ionizing Particles*, Nobel Lecture, URL: `https://www.nobelprize.org/prizes/physics/1927/wilson/lecture/` (cit. on p. 1).

[2] D. Glaser and D. Rahm, *Characteristics of Bubble Chambers*, Phys. Rev. **97** (1955) 474 (cit. on p. 1).

[3] G. Charpak and F. Sauli, *High resolution electronic particle detectors*, Ann. Rev. Nucl. Part. Sci. **34** (1984) 285 (cit. on p. 1).

[4] J. England et al., *A silicon strip detector with 12-micrometer resolution*, Nucl. Instrum. Meth. **196** (1982) 149 (cit. on p. 1).

[5] L. Rossi, P. Fischer, T. Rohe, and N. Wermes, *Pixel Detectors: From Fundamentals to Applications*, Particle Acceleration and Detection, Springer-Verlag, 2006, ISBN: 978-3-540-28332-4, 978-3-540-28333-1 (cit. on pp. 1, 2, 5, 32–35, 38, 39, 41).

[6] *LHC Machine*, JINST **3** (2008) S08001, ed. by L. Evans and P. Bryant (cit. on pp. 2, 13).

[7] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716** (2012) 1, arXiv: `1207.7214 [hep-ex]` (cit. on p. 2).

[8] CMS Collaboration, *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC*, Phys. Lett. B **716** (2012) 30, arXiv: `1207.7235 [hep-ex]` (cit. on p. 2).

[9] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003 (cit. on pp. 2, 3, 13).

[10] CMS Collaboration, *The CMS Experiment at the CERN LHC*, JINST **3** (2008) S08004 (cit. on pp. 2, 13).

[11] ATLAS Collaboration, *Event display of a H → 4e candidate event*, ATLAS Photos, URL: `https://cds.cern.ch/record/1459495` (cit. on p. 2).

[12] N. Wermes, “Pixel vertex detectors,” *34th SLAC Summer Institute on Particle Physics: The Next Fronterier: Exploring with the LHC*, 2006, arXiv: `physics/0611075` (cit. on pp. 2, 8, 25).

[13] C. J. Kenney et al., *A Prototype monolithic pixel detector*, Nucl. Instrum. Meth. A **342** (1994) 59 (cit. on pp. 2, 45).

[14] R. Turchetta et al., *A monolithic active pixel sensor for charged particle tracking and imaging using standard VLSI CMOS technology*, Nucl. Instrum. Meth. A **458** (2001) 677 (cit. on pp. 2, 45).

[15] N. Wermes, *Pixel detectors ... where do we stand?* Nucl. Instrum. Meth. A **924** (2019) 44, arXiv: `1804.10640 [physics.ins-det]` (cit. on p. 2).

[16] W. Snoeys, *Monolithic pixel detectors for high energy physics*, Nucl. Instrum. Meth. A **731** (2013) 125, ed. by Y. Unno, H. Toyokawa, Y. Arai, and T. Hatsui (cit. on pp. 2, 40, 41, 48).

[17] I. Perić et al., *A high-voltage pixel sensor for the ATLAS upgrade*, Nucl. Instrum. Meth. A **924** (2019) 99 (cit. on p. 2).

[18] M. Havránek et al., *DMAPS: a fully depleted monolithic active pixel sensor—analog performance characterization*, JINST **10** (2015) P02013, arXiv: `1407.0641 [physics.ins-det]` (cit. on p. 2).

[19] N. Wermes, *Depleted CMOS pixels for LHC proton–proton experiments*, Nucl. Instrum. Meth. A **824** (2016) 483, ed. by M. G. Bisogni et al. (cit. on p. 2).

[20] M. Garcia-Sciveres and N. Wermes, *A review of advances in pixel detectors for experiments with high rate and radiation*, Rept. Prog. Phys. **81** (2018) 066101, arXiv: `1705.10150 [physics.ins-det]` (cit. on pp. 2, 7, 9–11, 13, 20, 29, 30, 43, 47).

[21] W. Snoeys, *Development of monolithic sensors for high energy physics in commercial CMOS technologies*, Nucl. Instrum. Meth. A **938** (2019) 41 (cit. on p. 2).

[22] *The High Luminosity LHC (HL-LHC) project*, URL: `https://hilumilhc.web.cern.ch/` (cit. on pp. 3, 14).

[23] *High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1*, tech. rep., 2017 (cit. on pp. 3, 10, 14, 16).

[24] ATLAS Collaboration, *Technical Design Report for the ATLAS Inner Tracker Pixel Detector*, tech. rep. CERN-LHCC-2017-021. ATLAS-TDR-030, CERN, 2017, URL: `https://cds.cern.ch/record/2285585` (cit. on pp. 3, 14, 15).

[25] CMS Collaboration, *The Phase-2 Upgrade of the CMS Tracker*, tech. rep. CERN-LHCC-2017-009. CMS-TDR-014, CERN, 2017, URL: `https://cds.cern.ch/record/2272264` (cit. on p. 3).

[26] J. (Chistiansen and M. (Garcia-Sciveres, *RD Collaboration Proposal: Development of pixel readout integrated circuits for extreme rate and radiation*, tech. rep. CERN-LHCC-2013-008. LHCC-P-006, The authors are editors on behalf of the participating institutes. the participating institutes are listed in the proposal: CERN, 2013, URL: `https://cds.cern.ch/record/1553467` (cit. on pp. 3, 16).

[27] H. Pernegger, *The Pixel Detector of the ATLAS experiment for LHC Run-2*, JINST **10** (2015) C06012 (cit. on pp. 3, 14, 15).

[28] M. Garcia-Sciveres et al., *The FE-I4 pixel readout integrated circuit*, Nucl. Instrum. Meth. A **636** (2011) S155, ed. by T. Ohsugi, H. Sadrozinski, and Y. Unno (cit. on pp. 3, 42, 133, 164).

[29] W. Snoeys et al., *A process modification for CMOS monolithic active pixel sensors for enhanced depletion, timing performance and radiation tolerance*, Nucl. Instrum. Meth. A **871** (2017) 90 (cit. on pp. 3, 50, 53).

[30] K. Moustakas et al., "Development in a Novel CMOS Process for Depleted Monolithic Active Pixel Sensors," *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference*, 2017 8533114 (cit. on pp. 3, 50).

[31] K. Moustakas et al., *CMOS Monolithic Pixel Sensors based on the Column-Drain Architecture for the HL-LHC Upgrade*, Nucl. Instrum. Meth. A **936** (2019) 604, ed. by G. Batignani et al., arXiv: `1809.03434 [physics.ins-det]` (cit. on pp. 3, 50).

[32] T. Wang et al., *Depleted fully monolithic CMOS pixel detectors using a column based readout architecture for the ATLAS Inner Tracker upgrade*, JINST **13** (2018) C03039, arXiv: `1710.00074 [physics.ins-det]` (cit. on pp. 3, 50).

[33] I. Caicedo et al., *The Monopix chips: Depleted monolithic active pixel sensors with a column-drain read-out architecture for the ATLAS Inner Tracker upgrade*, JINST **14** (2019) C06006, arXiv: `1902.03679 [physics.ins-det]` (cit. on pp. 3, 57, 134–136).

[34] C. Bespin et al., *DMAPS Monopix developments in large and small electrode designs*, Nucl. Instrum. Meth. A **978** (2020) 164460, arXiv: `2006.02297 [physics.ins-det]` (cit. on pp. 3, 57, 131, 133–136).

[35] H. Kolanoski and N. Wermes, *Particle Detectors*, Oxford University Press, 2020, isbn: 978-0-19-885836-2 (cit. on pp. 6–9, 20–24, 26–29, 33, 35–37, 39, 40, 43, 45, 46, 53).

[36] F. Zimmermann et al., *High-Energy LHC Design*, J. Phys. Conf. Ser. **1067** (2018) 022009 (cit. on p. 10).

[37] FCC Collaboration, *FCC-hh: The Hadron Collider: Future Circular Collider Conceptual Design Report Volume 3*, Eur. Phys. J. ST **228** (2019) 755 (cit. on p. 10).

[38] Belle II Collaboration, *Belle II Technical Design Report*, tech. rep., 2010, arXiv: `1011.0352 [physics.ins-det]` (cit. on p. 10).

[39] CLIC and CLICdp Collaborations, *The Compact Linear Collider (CLIC) - 2018 Summary Report*, **2/2018** (2018), ed. by P. Burrows et al., arXiv: `1812.06018 [physics.acc-ph]` (cit. on p. 10).

[40] H. Abramowicz et al., *The International Linear Collider Technical Design Report - Volume 4: Detectors*, tech. rep., 2013, arXiv: `1306.6329 [physics.ins-det]` (cit. on p. 10).

[41] RD50 Collaboration, *Development of radiation hard sensors for very high luminosity colliders: CERN-RD50 project*, Nucl. Instrum. Meth. A **511** (2003) 97, ed. by S. Olsen and D. Bortoletto (cit. on p. 12).

[42] ALICE Collaboration, *The ALICE experiment at the CERN LHC*, JINST **3** (2008) S08002 (cit. on p. 13).

[43] LHCb Collaboration, *The LHCb Detector at the LHC*, JINST **3** (2008) S08005 (cit. on p. 13).

[44] *The Inner Detector of the ATLAS experiment at the LHC*, URL: `https://atlas.cern/discover/detector/inner-detector` (cit. on pp. 14, 15).

[45] K. Olive, *Review of Particle Physics*, Chinese Physics C **38** (2014) 090001, URL: `https://doi.org/10.1088%2F1674-1137%2F38%2F9%2F090001` (cit. on pp. 21, 22).

[46] Wikipedia, *p-n juction*, URL: `https://en.wikipedia.org/wiki/P%E2%80%93n_junction` (cit. on p. 25).

[47] E. Simoen, B. Dierickx, C. L. Claeys, and G. J. Declerck, *Explaining the amplitude of RTS noise in submicrometer MOSFETs*, IEEE Transactions on Electron Devices **39** (1992) 422 (cit. on p. 37).

[48] C. Leyris et al., "Impact of Random Telegraph Signal in CMOS Image Sensors for Low-Light Levels," *2006 Proceedings of the 32nd European Solid-State Circuits Conference*, 2006 376 (cit. on p. 37).

[49] T. Carusone, D. Johns, and K. Martin, *Analog Integrated Circuit Design*, Analog Integrated Circuit Design, Wiley, 2011, ISBN: 9780470770108, URL: `https://books.google.de/books?id=1OIJZzLvVhcC` (cit. on pp. 38, 97).

[50] L. Blanquart et al., *Pixel readout electronics for LHC and biomedical applications*, Nucl. Instrum. Meth. A **439** (2000) 403, ed. by P. Holl et al. (cit. on p. 41).

[51] M. Barbero et al., *Design and test of the CMS pixel readout chip*, Nucl. Instrum. Meth. A **517** (2004) 349 (cit. on p. 42).

[52] I. Peric et al., *The FEI3 readout chip for the ATLAS pixel detector*, Nucl. Instrum. Meth. A **565** (2006) 178, ed. by J. Grosse-Knetter, H. Krueger, and N. Wermes (cit. on pp. 42, 50, 60).

[53] M. Barbero et al., *Radiation hard DMAPS pixel sensors in 150 nm CMOS technology for operation at LHC*, JINST **15** (2020) 05, arXiv: `1911.01119 [physics.ins-det]` (cit. on pp. 42, 46, 48, 60, 68, 112).

[54] M. Prathapan et al., *Towards the large area HVCMOS demonstrator for ATLAS ITk*, Nucl. Instrum. Meth. A **936** (2019) 389, ed. by G. Batignani et al. (cit. on p. 42).

[55] I. Berdalovic et al., *Monolithic pixel development in TowerJazz 180 nm CMOS for the outer pixel layers in the ATLAS experiment*, JINST **13** (2018) C01023 (cit. on pp. 42, 50).

[56] W. Snoeys et al., *Layout techniques to enhance the radiation tolerance of standard CMOS technologies demonstrated on a pixel detector readout chip*, Nucl. Instrum. Meth. A **439** (2000) 349, ed. by P. Holl et al. (cit. on p. 44).

[57] S. Parker, *A Proposed VLSI Pixel Device For Particle Detection*, Nucl. Instrum. Meth. A **275** (1989) 494, ed. by E. Heijne, I. Debusschere, and H. Kraner (cit. on p. 45).

[58] J. Schambach et al., *A MAPS Based Micro-Vertex Detector for the STAR Experiment*, Phys. Procedia **66** (2015) 514, ed. by B. L. Doyle et al. (cit. on p. 45).

[59] M. Mager, *ALPIDE, the Monolithic Active Pixel Sensor for the ALICE ITS upgrade*, Nucl. Instrum. Meth. A **824** (2016) 434, ed. by M. G. Bisogni et al. (cit. on pp. 45, 49).

[60] I. Mandić et al., *Neutron irradiation test of depleted CMOS pixel detector prototypes*, JINST **12** (2017) P02021, arXiv: `1701.05033 [physics.ins-det]` (cit. on p. 46).

[61] A. Schöning et al., *MuPix & ATLASpix: Architectures and Results*, PoS **Vertex2019** (2020) 024, arXiv: `2002.07253 [physics.ins-det]` (cit. on p. 48).

[62] G. Aglieri Rinella, *The ALPIDE pixel sensor chip for the upgrade of the ALICE Inner Tracking System*, Nucl. Instrum. Meth. A **845** (2017) 583, ed. by G. Badurek et al. (cit. on p. 49).

[63] C. Gao et al., *A novel source–drain follower for monolithic active pixel sensors*, Nucl. Instrum. Meth. A **831** (2016) 147, ed. by Y. Unno et al. (cit. on p. 50).

[64] H. Pernegger et al., *First tests of a novel radiation hard CMOS sensor process for Depleted Monolithic Active Pixel Sensors*, JINST **12** (2017) P06008 (cit. on pp. 50, 55–57).

[65] R. Cardella et al., *MALTA: an asynchronous readout CMOS monolithic pixel detector for the ATLAS High-Luminosity upgrade*, JINST **14** (2019) C06019 (cit. on pp. 50, 57).

[66] M. Munker et al., *Simulations of CMOS pixel sensors with a small collection electrode, improved for a faster charge collection and increased radiation tolerance*, JINST **14** (2019) C05013, arXiv: `1903.10190 [physics.ins-det]` (cit. on pp. 50, 57, 58, 136).

[67] M. Dyndal et al., *Mini-MALTA: Radiation hard pixel designs for small-electrode monolithic CMOS sensors for the High Luminosity LHC*, JINST **15** (2020) P02005, arXiv: `1909.11987 [physics.ins-det]` (cit. on pp. 50, 126, 147, 170).

[68] G. Aglieri Rinella et al., *Charge collection properties of TowerJazz 180 nm CMOS Pixel Sensors in dependence of pixel geometries and bias parameters, studied using a dedicated test-vehicle: the Investigator chip*, (2020), arXiv: `2009.10517 [physics.ins-det]` (cit. on p. 53).

[69] L. Snoj, G. Žerovnik, and A. Trkov, *Computational analysis of irradiation facilities at the JSI TRIGA reactor*, Appl. Radiat. Isot. **70** (2012) 483 (cit. on pp. 55, 130).

[70] E. Schioppa et al., *Measurement results of the MALTA monolithic pixel detector*, Nucl. Instrum. Meth. A **958** (2020) 162404, ed. by M. Krammer et al. (cit. on p. 57).

[71] S. Zhang, *Charge collection efficiency simulation of irradiated monolithic silicon pixel detectors with TCAD*, BONN-IB-2019-05, Master Thesis: University of Bonn, 2019, URL: `https://www.hep1.physik.uni-bonn.de/results/data/internal/Zhang_Master.pdf` (cit. on p. 57).

[72] H. Pernegger et al.,
Radiation hard monolithic CMOS sensors with small electrodes for High Luminosity LHC,
Nucl. Instrum. Meth. A **986** (2021) 164381 (cit. on pp. 59, 132, 135, 153, 154).

[73] T. Wang and T. Hemperek, “Column-drain readout architecture efficiency simulation for the TJ-Monopix and LF-Monopix DMAPS development,” Internal communication, 2017 (cit. on pp. 63, 65, 66).

[74] D. Arutinov et al., *Digital architecture and interface of the new ATLAS pixel front-end IC for upgraded LHC luminosity*, IEEE Trans. Nucl. Sci. **56** (2009) 388 (cit. on p. 65).

[75] C. Ay et al., *Monte Carlo generators in ATLAS software*,
J. Phys. Conf. Ser. **219** (2010) 032001, ed. by J. Gruntorad and M. Lokajicek (cit. on p. 66).

[76] D. Kim et al., *Front end optimization for the monolithic active pixel sensor of the ALICE Inner Tracking System upgrade*, JINST **11** (2016) C02042 (cit. on p. 70).

[77] Y. Tsividis and C. McAndrew, *Operation and modeling of the MOS transistor; 3rd ed.*
Oxford series in electrical and computer engineering, Oxford Univ. Press, 2011,
URL: `https://cds.cern.ch/record/1546736` (cit. on p. 74).

[78] R. G. Carvajal and J. Ramirez-Angulo and others,
The flipped voltage follower: a useful cell for low-voltage low-power circuit design,
IEEE Transactions on Circuits and Systems I: Regular Papers **52** (2005) 1276 (cit. on p. 74).

[79] H. Hillemanns et al.,
“Radiation hardness and detector performance of new 180nm CMOS MAPS prototype test structures developed for the upgrade of the ALICE Inner Tracking System,”
2013 IEEE Nuclear Science Symposium and Medical Imaging Conference and Workshop on Room-Temperature Semiconductor Detectors, 2013 (cit. on p. 100).

[80] V. Filimonov, *Development of a serial powering scheme and a versatile characterization system for the ATLAS pixel detector upgrade*, BONN-IR-2017-06,
PhD Thesis: University of Bonn, 2017 (cit. on p. 117).

[81] Silizium Laboratory Bonn, *Basil, a data acquisition and system testing framework*,
URL: `https://github.com/SiLab-Bonn/basil` (cit. on p. 117).

[82] W. J. Snoeys, *A New integrated pixel detector for high-energy physics*, PhD Thesis, 1992 (cit. on p. 121).

[83] W. Hillert, *The Bonn electron stretcher accelerator ELSA: Past and future*,
Eur. Phys. J. A **28S1** (2006) 139, ed. by H. Arenhoevel et al. (cit. on p. 133).

[84] N. Heurich, F. Frommberger, P. Hänisch, and W. Hillert,
“The New External Beamline for Detector Tests at ELSA,”
7th International Particle Accelerator Conference, 2016 THPOY002 (cit. on p. 133).

[85] H. Jansen et al., *Performance of the EUDET-type beam telescopes*,
EPJ Tech. Instrum. **3** (2016) 7, arXiv: `1603.09669 [physics.ins-det]` (cit. on p. 133).

[86] J. Baudot et al., “First test results Of MIMOSA-26, a fast CMOS sensor with integrated zero suppression and digitized output,”
2009 IEEE Nuclear Science Symposium and Medical Imaging Conference, 2009 1169 (cit. on p. 133).

[87] Silizium Laboratory Bonn,
A powerful and adaptable analysis software in Python for beam test data,
URL: `https://github.com/SiLab-Bonn/testbeam_analysis` (cit. on p. 133).

[88] T. Wang et al., *Depleted Monolithic Active Pixel Sensors in the LFoundry 150 nm and TowerJazz 180 nm CMOS Technologies*, PoS **Vertex2019** (2020) 026 (cit. on pp. 134–136).

[89] R. Diener et al., *The DESY II Test Beam Facility*, Nucl. Instrum. Meth. A **922** (2019) 265, arXiv: `1807.09328 [physics.ins-det]` (cit. on p. 135).

[90] C. Kelley, *Iterative Methods for Optimization*, Frontiers in Applied Mathematics, SIAM, 1999, ISBN: 9781611970920, URL: `https://books.google.de/books?id=7oqAh1MAZG0C` (cit. on p. 143).

[91] *Cadence® Virtuoso® analog design environment*, URL: `https://www.cadence.com` (cit. on p. 143).

[92] E. Monteil,
RD53A: a large scale prototype for HL-LHC silicon pixel detector phase 2 upgrades,
PoS **TWEPP2018** (2019) 157 (cit. on p. 164).

[93] L. Gaioni et al., *Test results and prospects for RD53A, a large scale 65 nm CMOS chip for pixel readout at the HL-LHC*, Nucl. Instrum. Meth. A **936** (2019) 282, ed. by G. Batignani et al. (cit. on p. 164).

[94] Arsovski, Igor, *High-speed low-power sense amplifier design*, URL: `https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.387.4771&rep=rep1&type=pdf` (cit. on p. 166).

[95] J. Rabaey, *Digital Integrated Circuits: A Design Perspective*,
Prentice Hall electronics and VLSI series, Prentice Hall, 1996, ISBN: 9780133942712,
URL: `https://books.google.de/books?id=MJt4QgAACAAJ` (cit. on p. 166).

[96] R. Cardella et al., *LAPA, a 5 Gb/s modular pseudo-LVDS driver in 180 nm CMOS with capacitively coupled pre-emphasis*, PoS **TWEPP-17** (2017) 038 (cit. on pp. 170, 172).

List of Figures

List of Tables

Acknowledgements

This dissertation marks the completion of a fascinating yet challenging journey during which i had the opportunity to work on cutting edge detector technologies, meet great people and collaborate with leading experts in international projects. At this point i would like to thank everyone that contributed and helped to make it happen.

I would like to express my deepest gratitude and appreciation to my supervisor Prof. Dr. Norbert Wermes for giving my the opportunity to be part of his group. His guidance and support helped me overcome any difficulty in the process of my PhD studies. He has been a bright example both as an academic and as a person and what i have learned from him will be invaluable in my future career. It has been an honor to work with you.

Many thanks to my PhD committee members, Prof. Dr. Klaus Desch, Prof. Dr. Bernard Metsch and Prof. Dr. Joachim Anlauf.

This work would not have been possible without the help and support of Dr. Tianyang Wang and Dr. Tomasz Hemperek to whom i am extremely grateful. They introduced me to detector development, were happy to share their expertise and have always been there to answer my questions. Their active contribution was vital for the developments presented in this work. Thank you for the interesting, sometimes philosophical, discussions we had.

I would like to extend my sincere thanks to Dr. Hans Krüger for his valuable advice, and insightful suggestions. His extensive knowledge and analytic and problem solving mindset has inspired me to become a better scientist.

I am also grateful to all my fellow designers and office mates for the excellent collaboration and working environment. Special thanks to Piotr Rymaszewski who has always been willing to help with everything, from IC design to IT and laboratory equipment.

I very much appreciate the contribution of Christian Bespin, Ivan Caicedo and Dr. Toko Hirono that was instrumental for the measurement and characterization of the developed pixel detector chips. Thank you for your effort especially during demanding test beam and irradiation campaigns.

I would also like to extend my deepest gratitude to Dr. Walter Snoeys, Dr. Heinz Pernegger and their group members for their excellent collaboration. They have introduced me to the specifics of the TowerJazz modified process sensor concept and front-end design, shared critical blocks and always provided valuable advice. My secondment at CERN was a unique and unforgettable experience. It has been a great pleasure to work with you.

Special thanks to Prof. Dr. Thomas Noulis and Prof. Dr. Stylianos Siskos, my former supervisors and mentors, without whom i would not be where i am now.

I would like to offer my heartfelt thankfulness and gratitude to my parents Grigorios and Eleftheria and my brother Vasileios, for their unconditional support, understanding and sacrifices throughout all these years. Special thanks to Aidonia, my love, for her understanding, patience and support that kept me going even at difficult times.

This project has been supported by the Smart Sensor Technologies and Training for Radiation Enhanced Applications and Measurements (STREAM) Innovative Training Network (ITN) as part of the European Union’s Horizon 2020 research and innovation programme under grant agreement No 675587.