University of Pavia

Department of Electrical, Computer and Biomedical Engineering

Ph.D. course in Microelectronics
XXXVI Cycle

A test platform for CMOS SPADs and digital SiPMs in a 110 nm technology

Ph.D. Candidate:
Gianmarco Torilla

Supervisors:
Prof. Lodovico Ratti
Prof. Carla Vacchi
Ph.D. Coordinator:
Prof. Piero Malcovati

July 2024

Contents

1 Single photon avalanche diodes: operating principle and applications **1**

1.1 Fundamental physics of semiconductor detectors 4
1.2 Avalanche photodetectors 8
1.2.1 The multiplication process 9
1.2.2 Linear mode operation 13
1.2.3 Geiger mode operation 13
1.3 Single Photon Avalanche Diodes 14
1.3.1 Key parameters and Figures of Merit (FoMs) 17
1.3.2 Quenching circuits 27
1.3.2.1 SPAD orientation 27
1.3.2.2 Passive quenching circuits 29
1.3.2.3 Active quenching circuits 33
1.3.3 SPADs in CMOS technology 36
1.3.4 Silicon Photomultipliers (SiPMs) 37
1.3.4.1 Digital Silicon Photomultipliers (dSiPMs) . . . 39
1.3.5 Main fields of application 46
1.4 The APIX2/ASAP project 51
1.4.1 Description of the coincidence-based dual layer SPAD structure 52
1.4.2 Dark noise mitigation in a coincidence-based structure . 54

2 The ASAP chip **57**

2.1 Chip overview 57
2.2 Single SPADs and linear APDs 59
2.3 SPAD arrays and core circuits 61
2.3.1 Row and column selection registers 63
2.3.2 Array 1 (A1) 69
2.3.2.1 A1 in-pixel electronics 70

2.3.3 Array 2 (A2) 77
2.3.3.1 A2 in-pixel electronics 77
2.3.4 Array 3 (A3) 80
2.3.4.1 A3 in-pixel electronics 81
2.3.5 Array ACT-Q (AAQ) 81
2.3.5.1 AAQ in-pixel electronics 82
2.3.6 Array SiPM (ASIPM) 85
2.3.6.1 SiPM electronics 86
2.3.6.2 SiPM electronics - alternative design 102
2.3.7 Array SEPARATED-STI (ASSTI) 104
2.3.8 Array PIXEL-TEST (APXT) 104
2.3.9 Time to digital converter (TDC) 106

3 Measurement setup and characterization results 115
3.1 Measurement setup 115
3.2 Characterization results 121
3.2.1 Measurement methods 123
3.2.2 Breakdown voltage 123
3.2.3 Photon detection 128
3.2.4 Dark count rate 132
3.2.4.1 DCR measurement procedures 132
3.2.4.2 DCR measurements 135
3.2.4.3 ASAP110LF vs APIX2LF 144
3.2.4.4 RTS measurements 148
3.2.4.5 Crosstalk measurements 153
3.2.5 Digital SiPM measurements 155

Introduction

Single Photon Avalanche Diodes (SPADs) are semiconductor devices enabling the detection of individual photons. The p-n junction, representing the core of the sensing structure, is operated with a reverse voltage exceeding the breakdown voltage of the sensor. Under proper bias conditions, the avalanche which is triggered upon the absorption of a photon is self-sustaining, thus generating an output signal which is highly non-linear with respect to the incident radiation amplitude. In principle, since the information provided by the SPAD is of the binary type, a fully digital circuit can be used to read out the sensor output signal.
The fabrication of SPADs in well-established CMOS technologies has enabled the development of complete detection systems integrating on the same substrate both the sensing structures and the readout electronics. For applications requiring high granularity and spatial resolution, CMOS SPADs can be arranged in highly dense arrays, where the signals generated by the individual pixels can be processed by readout circuits attaining the functionality of silicon photomultipliers (SiPMs) or position-sensitive detectors. Noise, in the form of dark pulse generation, is the most limiting factor for the production of SPADs in commercial CMOS technologies. Fabrication processes, with not sufficiently high degree of purity, may affect the noise performance of the sensors, thus strongly limiting their detection capabilities.
Due to their high level of miniaturization, design flexibility, convenience of use and elevated signal processing capabilities, SPADs have become the detectors of choice for Time-Correlated Single Photon Counting (TCSPC). The properties of SPADs, mostly employed in the photon detection domain for applications such as fluorescence lifetime imaging, optical ranging, Raman spectroscopy and positron emission tomography, may be beneficially exploited also for charged particle tracking. Since the sensitive volume of a SPAD is limited to the thin depleted region around the p-n junction, the amount of detector material can in principle be substantially reduced, by thinning down the sensor substrate to a few tens of microns, without undermining the signal to noise ratio.
This thesis work presents the development of a test platform for the characterization of CMOS SPADs fabricated in a 110 nm CMOS Image Sensor (CIS) technology. As compared to standard CMOS technologies, a CIS technology enables the production of CMOS sensors featuring improved noise and optical performance, by using specifically designed fabrication steps. The

characterization of SPADs is of paramount importance to evaluate the sensor performance in different conditions, thus identifying the main limits of the employed technology and provide a classification of different sensor structures. The results from the characterization may be used as a starting point for the improvement of the sensor design and set the state-of-the-art. A novel digital SiPM architecture, based on parallel counters, is described in this thesis work. The new structure, providing the number of simultaneously firing SPADs with very low latency time, makes it possible to perform real time photon counting with the use of a fully digital readout network.
The research work presented in this thesis was conducted as part of the ASAP project, funded by the Italian Institute for Nuclear Physics (INFN). In the framework of this project a new type of position-sensitive sensor for charged particle detection, based on the vertical integration of SPAD arrays, was conceived. The aim of the project is to reduce the impact of dark noise on the detection system by exploiting the coincidence readout of overlapped SPADs.
The first chapter of this manuscript begins with a general discussion about the working principle of avalanche photodetectors (APDs). The different working regions of SPADs, as well as the typical figures of merit, will be presented in a dedicated section. After a broad discussion about the main architectures of quenching circuits, SiPMs will be introduced, together with a brief description of the most notable digital SiPM architectures that have set the benchmark in the last decade. As a final part of the chapter, more details about the ASAP project will be provided, after a brief overview of the most common applications exploiting the SPAD detection capabilities.
In the second chapter, the design of the chip developed during this thesis work, the ASAP110LF chip, fabricated in a 110 nm CIS technology, will be discussed. The different structures included in the chip, involving both SPAD arrays and single sensors, will be described, together with the digital signals needed for the correct chip operation. A general overview explaining the basic principles of parallel counters will be provided, before the description of the above mentioned novel digital SiPM architecture, which is implemented in one of the arrays of the ASAP110LF chip.
In the third and last chapter, the results from the characterization of the structures integrated in the new chip will be presented. Before discussing the measurement results, the automatic FPGA-based measurement system, which was developed to accomplish the SPAD characterization, will be described. The results from a preliminary characterization of a SiPM sample will be provided in the last part of the chapter.

Chapter 1

Single photon avalanche diodes: operating principle and applications

Single photon sensing, which represents the ultimate detection limit for the electromagnetic radiation [1], has emerged as a key focus of research and development across various fields. From quantum computing to medical diagnostics, from 3D imaging to fundamental scientific exploration, the importance of single photon detection in the modern world cannot be overstated. Through the years several acquisition techniques, based on Time-Correlated Single Photon Counting (TCSPC), have taken advantage of the increasing capabilities offered by specific radiation sensitive devices to collect and study the significant amount of information carried by single photon emission. Raman spectroscopy, fluorescence lifetime imaging, time of flight ranging, quantum imaging are just a few of the scientific fields heavily relying on the detection of individual light particles. The ability to manipulate and detect single photons has unlocked new avenues for scientific research and technological breakthroughs.
In the second half of the previous century, three types of Single Photon Detectors (SPDs) emerged, each exploiting different physical principles. Vacuum-based detectors like photomultiplier tubes (PMT) and microchannel plates (MCP) were the initial choices for photon counting. PMTs, in particular, gained prominence in the 1960s due to their high gain (up to $10^5 - 10^6$) and detection efficiency, up to 35% for visible wavelengths, making them the most reliable single-photon detectors for several decades [2]. In contrast, solid-state SPDs, in the particular flavour of avalanche detectors, faced several challenges during their development, due to limitations in available junctions and quench-

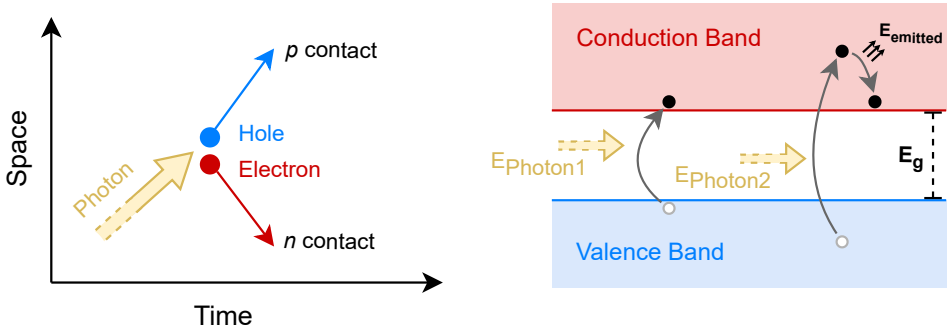
ing circuits. However, their potential became evident from the outset as they offered solutions to the drawbacks of vacuum-based SPDs, specifically in terms of size and cost. While PMTs and MCPs were efficient photodetectors with high gain, low noise, and good timing characteristics, their bulky nature, the high manufacturing precision requirements, and the demand for high power consumption made them unsuitable for compact systems.
Similarly, solid-state detectors were preferred over cryogenic-temperature-based SPDs, such as superconducting nanowire single-photon detectors and transition-edge sensors, mainly due to the impractical dimensions and power consumption of the cooling system [3]. Although cryogenic detectors demonstrated high quantum efficiency in the visible and near-infrared wavelength range, remarkable noise performance, and picosecond pulse-to-pulse timing jitter, their integration was hindered by the cooling system's limitations. Therefore, in the last decades, to meet the requirements of modern applications, including cost-effectiveness, miniaturization, reliability, design flexibility, integration density, and signal processing capabilities, there was a strong inclination towards a fully solid-state solution. Such an approach was aimed at overcoming the limitations of vacuum-based and cryogenic-temperature-based SPDs, thus offering flexible, compact and cost-effective single-photon detectors that can serve proficiently to diverse applications [4].
In the wide scenario of solid state detectors, avalanche photodiodes (APDs) turned out to be the most suitable monolithic solution in applications where a substantial internal gain is highly desirable, such as single photon counting. Based on the self-sustained avalanche mechanism, Signal Photon Avalanche Diodes, namely SPADs, represent the ultimate frontier in terms of TCSPC. The distinctive characteristic of SPADs is the huge internal gain. These sensors are mainly used to capture faint optical signals, since even a single photon is very likely to trigger a potentially never-ending avalanche in the active region of the sensor. Recently, the remarkable gain of SPADs has been successfully exploited also in the field of charged particle tracking [5][6][7]. In this case, charged particles serve the same purpose, from the standpoint of the electron-hole pair generation, as single photons transferring energy to the lattice of the detector. Since the output current generated by SPADs is highly non-linear with respect to the impinging radiation, no pre-amplification circuit is needed, thus resulting in strong savings in terms of power consumption and in a significant reduction of the design complexity featured by the processing electronics. Indeed, the intrinsic digital nature of SPADs opens up the potential for a fully digital design of the entire readout channel, including both the front-end circuits and the electronics used to process the data. In addition,

since the sensitive volume of the SPAD is limited to the depletion region of the p-n junction making up the device, the overall thickness of the detector can be potentially reduced to few micrometers, with a substantial reduction in the amount of detector material. Furthermore, remarkable performance, as far as the spatial and time resolution are concerned, have been widely recorded in the literature [8][9][10][11][12][13], where a high number of complex systems relying on SPADs are extensively described.
Eventually, the fabrication of SPADs in commercial CMOS technologies, that have nowadays a recognized level of maturity and reliability, has enabled the production of monolithic detection systems, which integrate successfully both the readout electronics and the sensing element in a common substrate [14]. Despite all the compromises brought about with the CMOS integration, mainly related to the performance of the sensor in terms of Dark Count Rate (DCR) and Photon Detection Efficiency (PDE), the possibility of designing high density arrays with a remarkable fill factor, the low power consumption allowed by the relatively low voltages needed for the device operation and the capability of having a complete sensing system inside a single silicon die have elected CMOS SPADs as the best candidates for modern and future applications of photon counting.
Thanks to their exceptional capabilities and elevated adaptability in a wide range of fields, SPADs have revolutionized industry, research, and everyday life, paving the way for groundbreaking innovations and unprecedented possibilities.

In this chapter, the basic principle of SPAD operation will be described. Initially, particular focus will be given to the fundamental physics of avalanche photodiodes and the commonly utilized front-end circuits that fully exploit SPADs capabilities. After a section devoted to the most important figures of merit used to compare sensor performance, the impact which CMOS technologies had on the SPAD development will be explained. Before a general description of various practical applications involving SPADs, a thorough discussion about silicon photomultipliers will be held. The chapter will conclude with a brief explanation of the leading ideas behind the ASAP project, the broader framework to which this work contributes.

1.1 Fundamental physics of semiconductor detectors

Semiconductor detectors rely on the internal photoelectric effect. If the impinging electromagnetic radiation has an energy larger than the bandgap energy of the semiconductor ($E_g = 1.12\ eV$ for silicon), it may be absorbed in the active area of the device, hence triggering the generation of an electron-hole pair (Fig. 1.1a). The photogenerated carriers, i.e. carriers generated after the absorption of a photon, are free to move in the lattice. Under the action of an electric field, they can produce a current, flowing through the device. Indeed, from the energy point of view, the generation of the electron-hole pair corresponds to promoting an electron from the valence band to the conduction band, while the opposite happens to the respective hole [15]. In order to lift an electron to the conduction band, the energy of the absorbed electromagnetic radiation must be, in principle, equal to or larger than the bandgap energy of the material. If enough energy is provided, the electron can jump into one of the empty states of the conduction band, before moving towards its edge, thus emitting energy in the form of lattice vibrations (Fig. 1.1b). The generated current, also referred to as photocurrent, can be sensed by specifically designed electronics and get further processed.
Assuming that the incident light comes from an infinitely thick medium, the optical power through the section of the semiconductor material decays expo-



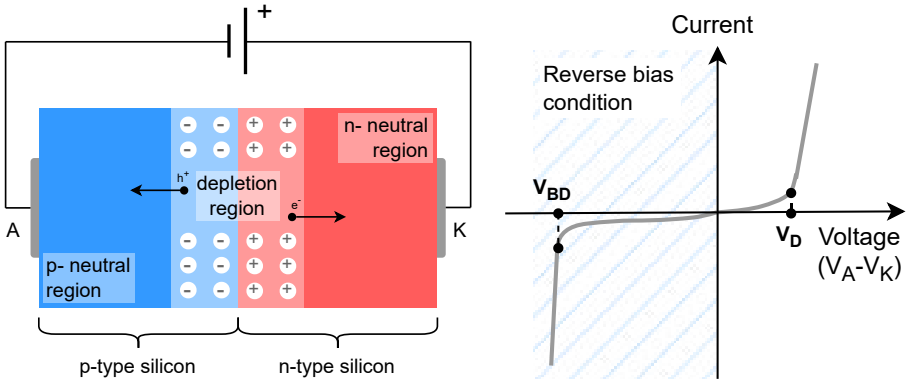
(a) (b)

Figure 1.1: Absorption of a photon in a semiconductor material: a) space-time representation, b) simplified energy band diagram showing the photogeneration of carriers in the case $E_{Photon2} > E_{Photon1} \geq E_g$.

nentially, following the Lambert-Beer law [16]:

$$P(\lambda, y) = P_0 e^{-\alpha(\lambda)y}, \tag{1.1}$$

where P_0 is the optical power at the semiconductor surface (taken as $y = 0$), $\alpha(\lambda)$ represents the optical absorption coefficient, which is a function of the wavelength, and y is the depth inside the material (under the assumption that the direction of the incident light is orthogonal to the semiconductor surface). Starting from (1.1), important parameters determining the detector behavior can be extracted, such as the penetration depth of light inside the semiconductor and the photogeneration rate per volume.
In a semiconductor photodetector, the p-n junction operated with a reverse bias is the most suitable structure that effectively harnesses the photogeneration phenomenon. A simplified diagram of a p-n junction biased with a reverse voltage, together with its $I - V$ characteristic, is shown in Fig. 1.2. Due to the reverse bias condition of the junction, the current flowing through it, in the absence of a photogeneration event, is minimal and referred to as dark current. The latter is a key parameter of the detector, and it depends on the junction saturation current, on the applied bias voltage and on the absolute temperature. As already mentioned, in case photons are absorbed in the active area of the sensor, some free carriers are produced. Under the effect of the applied reverse bias, the resulting current sums up to the dark current, which however represents the highest contribution. Therefore, the exposition of the photodiode to electromagnetic radiation may shift the $I-V$ characteristics by



(a) **(b)**

Figure 1.2: a) simplified diagram of a p-n junction under reverse bias condition, b) I-V characteristics.

an amount determined by the photogenerated current, which is proportional to the incident radiant power.
The active region of the photodiode, which is the area where photogeneration is more likely to occur, corresponds to the depletion region of the p-n junction. This layer is usually referred to as space charge region, since in a steady state condition no carriers can be found here. Given its crucial role in the photogeneration phenomenon, it may be important to compute the expression for the width of the depletion layer of the junction, as a function of the applied voltage. This calculation provides insights into how the applied voltage affects the detection performance of the photodiode by influencing its active region. The electric field, developed inside the junction, is not uniformly distributed along all the structure, according to the following expressions [17]:

$$\mathcal{E}_n(x) = -\frac{qN_d}{\epsilon_s}(x_n - x), \qquad 0 < x < x_n, \tag{1.2}$$

$$\mathcal{E}_p(x) = -\frac{qN_a}{\epsilon_s}(x + x_p), \qquad -x_p < x < 0, \tag{1.3}$$

where q is the charge of the electron ($1.60217663 \cdot 10^{-19}$ C), N_d and N_a are respectively the donor and the acceptor concentrations, ϵ_s is the silicon permittivity, x_n and x_p represent the edges of the n-doped and p-doped sides of the depletion region (with reference to Fig. 1.3), and x is the axis perpendicular to the junction surface, which, conventionally, is taken as the zero for this system of reference. As it can be noticed from (1.2) and (1.3), the electric field exhibits a linear increase in both the n-doped and p-doped sides, extending towards the junction surface where it reaches its maximum value. As shown in Fig. 1.3, the electric field falls down to zero at the boundaries of the depletion region. As a consequence, a carrier generated within the depletion region experiences a significant electric field, resulting in strong acceleration. Conversely, a charge generated in the neutral region, beyond the space charge region, moves at a slower pace due to the absence of a substantial electric field. This charge, far away from the depletion region, diffuses gradually through the diode, often recombining before reaching the high field layer. As a result, the photogenerated charge originating from the depletion region plays a significant role in the overall photocurrent, while the charge generated in the neutral region contributes negligibly. Photons incident on the device within the neutral region mostly remain undetected.
Considering the continuity of the electric field on the junction surface (leading to $N_a x_p = N_d x_n$), and knowing the expression of the built-in potential of a p-n junction [18], the total width of the depletion region, with an applied reverse

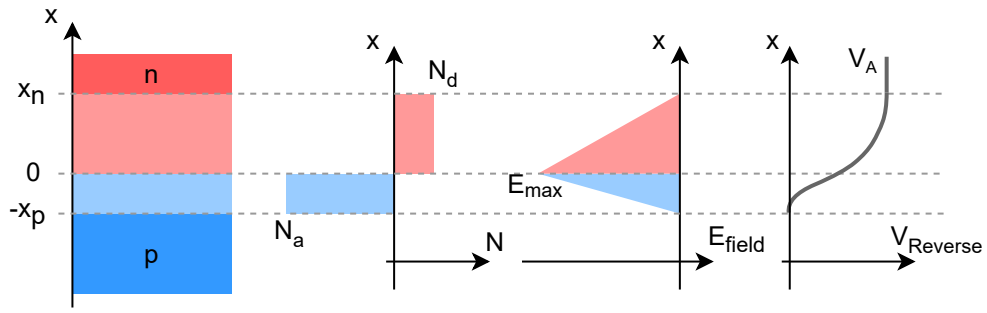


Figure 1.3: Doping concentration, electric field, and potential distribution within the depletion region of a reverse biased p-n junction.

bias, can be calculated as:

$$x_d = x_n + x_p = \sqrt{\frac{2\varepsilon_s}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)(\phi_i - V_A)}, \tag{1.4}$$

where x_d is the total width of the depletion region, ϕ_i is the built-in potential and V_A is the applied bias voltage. Tipically, in silicon sensors, the junction is formed by a shallow p+ diffusion with high doping in a bulk material that has lower doping levels [19]. As a result, when the junction is reverse biased, the depletion region expands deeper into the n-side of the junction, since, as depicted in Fig. 1.3, it has a lower concentration of dopants. Therefore, the term $1/N_a$ in (1.4) can be neglected. Additionally, considering that the external voltage applied to the junction, V_A, is typically much greater, in absolute value, than the built-in voltage, ϕ_i, the total width of the depletion region can be approximated as:

$$x_d \approx x_n \approx \sqrt{\frac{2\varepsilon_s}{qN_d}|V_A|}. \tag{1.5}$$

From this formula, it can be deduced that the total width of the depletion region depends on the square root of $|V_A|$. Specifically, as the reverse voltage applied to the junction increases, the space charge region expands, thus leading to an increase of the active volume.
A direct outcome of altering the width of the active region, utilizing voltage as a means, is a change in the photodetector responsivity. This parameter serves as a gauge of the effectiveness of the conversion process from light power to

electrical current. The responsivity R of a detector can be defined as

$$R = \frac{I_{PH}}{P_{in}} \quad \left[\frac{A}{W}\right], \tag{1.6}$$

where P_{in} is the power of the incident radiation and I_{PH} is the current produced by the photogeneration phenomenon. Beside changing with the active area, R is also a function of the temperature and of the wavelength featured by the incident light [20].

1.2 Avalanche photodetectors

Avalanche photodiodes (APDs) are particular semiconductor detectors relying on a junction structure which is specifically designed to fully exploit the avalanche mechanism. Sensor systems based on APDs are well-suited for applications demanding a significant internal gain, like those involving single particle detection.
In a reverse-biased photodiode, an electric field increasing with the applied voltage is established within the depletion region. If particles (electrons or holes) are generated within this region, after light absorption phoenomena or by thermal generation, they can acquire enough kinetic energy to break atomic bonds, leading to the formation of new electron-hole pairs. This phenomenon,

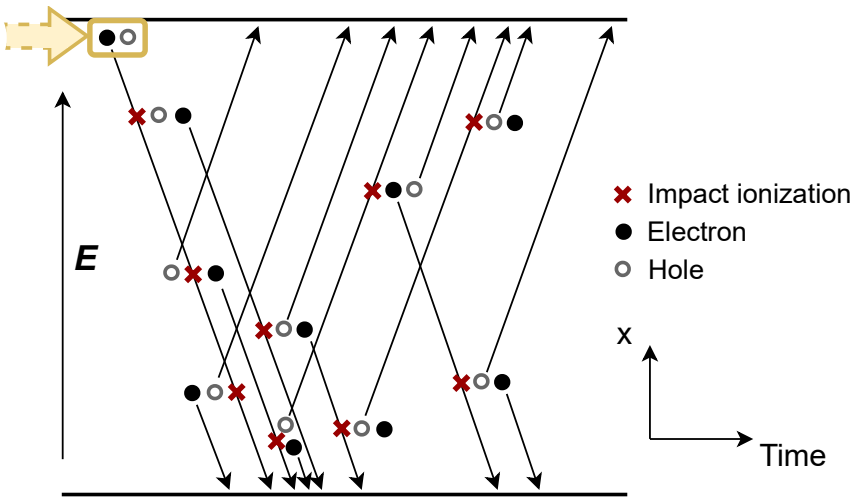


Figure 1.4: A series of impact ionization events giving rise to the avanlanche mechanism.

known as impact ionization, occurs when high energy carriers collide with the lattice, triggering the creation of additional carriers. The newly generated carriers, combined with the initial ones, undergo further acceleration, triggering subsequent impact ionization events. This cascading effect creates a positive feedback loop, progressively amplifying the number of carriers within the structure, hence the term avalanche. Fig. 1.4 schematically shows the generation of an avalanche from a primary electron-hole pair. It can be noticed that the number of carriers leaving the high field region is significantly increased while the avalanche phenomenon takes place. Although the multiplication process may produce bad effects in non-suitably designed semiconductor devices, leading to electrical breakdown of the structure, it can be successfully exploited for signal amplification.

1.2.1 The multiplication process

The multiplication process, which is responsible for the avalanche breakdown within a semiconductor detector, can be explained through the band diagrams depicted in Fig 1.5 [15]. Here, the energy diagrams are represented as a function of position. The slope of the energy bands is proportional to the electric field strength, according to $\mathcal{E}_x = \frac{1}{q}\frac{\partial E_i}{\partial x}$, where $\mathcal{E}_x$ indicates the electric field in the x direction and E_i is the intrinsic energy. Fig 1.5a and Fig 1.5b show the carrier generation phenomenon in an avalanche photodiode under the assumption of, respectively, low electric field and high electric field. In the energy diagrams, each electron is represented as a solid dot while holes are shown as empty circles. Solid horizontal arrows are used to represent the acceleration of a particle, and, consequently, the increasing vertical distance between the arrow and the edge of the conduction band gives the kinetic energy of the particle at the respective position x. Dashed vertical lines describe energy release occurrences after the collision of a particle with the lattice. Solid vertical lines indicate points where a multiplication event takes place. The primary carriers are supposed to be generated at position x_0, due to the absorption of electromagnetic radiation.
With reference to Fig. 1.5a, in the case of relatively low electric field, free carriers are not allowed to acquire significant kinetic energy due to the shallow inclination of the energy bands. As a consequence, the generation of secondary electron-hole pairs cannot take place. In other words, after each collision, no additional carrier can join the conduction band and the multiplication mechanism cannot be triggered. In the condition of higher field strength, depicted in Fig. 1.5b, the slope of the energy bands allows the charge carriers to gain high values of kinetic energy between collisions. After the first interaction with the

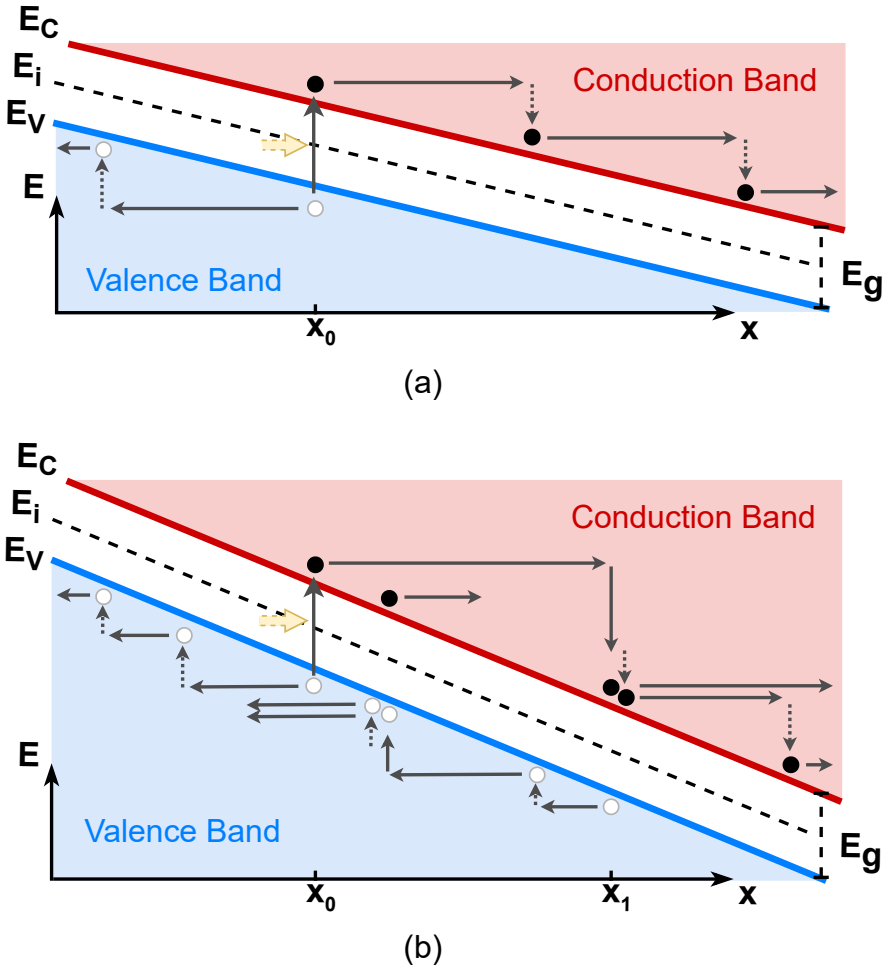


Figure 1.5: Energy diagrams, as a function of position, describing the multiplication mechanism in two cases: (a) low electric field, (b) high electric field.

lattice, occurring at position x_1, part of the energy involved in the impact is transferred into lattice vibration. The remaining energy causes the creation of an additional electron-hole pair. All the secondary carriers, generated within the high field region after an impact ionization event, are accelerated by the electric field and have a certain probability of creating other pairs of charge carriers, thus potentially starting an avalanche. Electrons and holes have a different probability for creating secondary charges. Depending on the magnitude of the electric field, a bias condition may be found where only a type

of carrier can cause multiplication events. In silicon-based detectors, electrons are more likely to produce impact ionization if compared to holes, thus moderate levels of electric field can be sufficient to let electrons be the only type of carrier capable to sustain the avalanche. In such a condition, the charge generated as a consequence of the multiplication process is proportional to the primary generated carriers. Besides the electric field, also the spatial extent of the high field region plays a fundamental role in determining the probability associated to the multiplication phenomenon, which strongly depends on the mean free path length.
At significantly high fields, when both types of carriers gain enough kinetic energy while travelling through the depleted region, the avalanche breakdown occurs. In order to obtain the onset condition for the avalanche breakdown, an asymmetrically doped reverse-biased p-n junction, of the type already presented in Fig. 1.3, can be considered. If the multiplication mechanism is triggered, the electron and hole concentrations increase by the same amount within the depletion region. The variation featured by the electron and hole concentrations, in a thin region dx of the space charge region, can be expressed as [15]

$$dn = dp = \alpha_n n(x)dx + \alpha_p p(x)dx, \tag{1.7}$$

where α_n and α_p are the field-dependent multiplication coefficients for electrons and holes respectively, while $n(x)$ and $p(x)$ are the carrier concentrations as a function of the position. In principle, due to their dependence on the electric field, also α_n and α_p can be written as a function of x. Under the assumption of doping asymmetry, the rate of the electrons moving by diffusion from the neutral p region into the space charge region is much larger than the rate of holes diffusing from the undepleted n region. Therefore, the hole concentration at the n-side boundary of the depleted region can be considered as zero. Since, in this approximation, holes are originated only as a consequence of multiplication processes, the hole concentration at position x can be expressed as

$$p(x) = n(L) - n(x), \tag{1.8}$$

where $x = x_n = L$ denotes the boundary of the n-side space charge region. From (1.7), it can be found that

$$\frac{dn}{dx} = (\alpha_n - \alpha_p)n(x) + \alpha_p n(L). \tag{1.9}$$

In the particular case of $\alpha_n = \alpha_p = \alpha(x)$, an analytic solution to the (1.9) is represented by the following

$$n(L) = n(0) + n(L) \int_0^L \alpha(x)dx. \tag{1.10}$$

The ratio between electrons leaving and electrons entering the depleted region, namely the multiplication factor or gain M, can be obtained as

$$M = \frac{n(L)}{n(0)} = \frac{1}{1 - \int_0^L \alpha(x)dx}. \tag{1.11}$$

By letting $M \longrightarrow \infty$, one can reach the condition for the avalanche breakdown:

$$\int_0^L \alpha(x)dx = 1. \tag{1.12}$$

In general, during the avalanche transient, an interplay occurs within the photodiode, involving the generation of carriers through successive impact ionization events and the departure of carriers from the structure upon reaching the device electrodes. From these two phenomena, the relevant rates can be extracted, named respectively carrier ionization rate (also referred to as the birth rate) and extraction rate [22]. Depending on the applied bias voltage, one of these rates prevails in the competition, thereby determining the

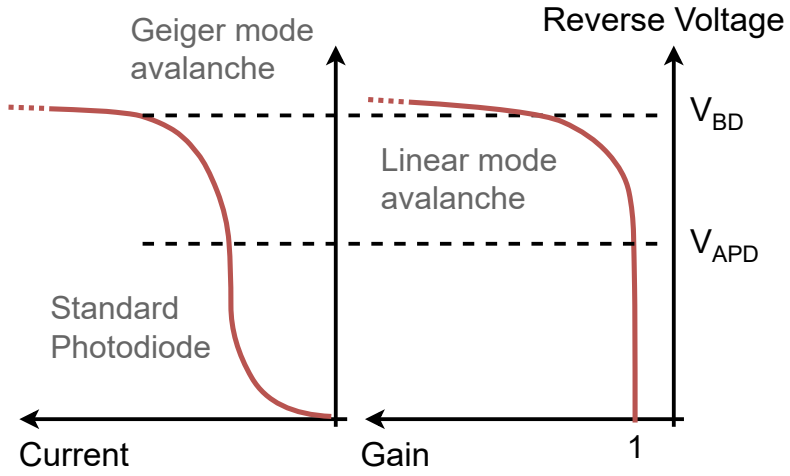


Figure 1.6: I-V and Gain-V characteristics of an avalanche photodiode in quasi static conditions [21].

operational mode of the detector. The device can operate in either the linear mode or the Geiger mode, distinguished by a threshold value known as the breakdown voltage (V_{BD}), as depicted in Fig. 1.6.

1.2.2 Linear mode operation

When the applied bias voltage is below the breakdown voltage, the extraction rate prevails over the carrier ionization rate, resulting in the damping of the avalanche. Therefore, the number of electrons and holes within the junction decreases. Upon the arrival of a photon, the immediate outcome is the generation of a finite number of subsequent electron-hole pairs. The internal gain, M, typically ranges from tens to hundreds, depending upon the bias voltage as illustrated in Fig. 1.6. Impact ionization is a statistical process, thus the exact number of collisions encountered by an electron within the junction remains non-deterministic. Consequently, the value of M serves as an average representation, as the actual internal gain may differ between two avalanche events occurring in the same junction under identical operating conditions. The fluctuations in the internal gain, inherent to the linear mode, generate excess noise, commonly referred to as multiplication noise. This noise becomes more pronounced as M increases due to elevated bias voltages. The extent of multiplication noise depends not only on the bias voltage but also on the specific material employed in constructing the junction. In particular, M features a low variance in materials where the ionization events are triggered by electrons. In contrast, higher values for the multiplication noise are obtained in materials where impact ionization events are produced also by holes.
The operating mode discussed so far is known as linear mode, since the photogenerated current exhibits a linear dependence on the incident light intensity.

1.2.3 Geiger mode operation

As shown in Fig. 1.6, when the reverse voltage exceeds the breakdown threshold, the current generated within the junction swiftly rises, following an exponential law. In this bias condition, the carrier ionization rate surpasses the extraction rate, thus electrons and holes are generated at a higher rate than they can be extracted, resulting in a potentially never-ending avalanche. This operational state is commonly referred to as Geiger mode and the devices relying on this principle are indicated as Geiger-mode avalanche photodiodes (GAPDs or Gm-APDs). In Geiger mode, the notion of gain becomes less meaningful as the device has the capability to produce an exceptionally large number of electron-hole pairs in response to the primary photogenerated car-

riers. Therefore, the gain is considered to be infinite. Due to their highly non-linear response with respect to the impinging radiation, Geiger avalanche detectors are widely recognized for their remarkable performance in single photon detection.

1.3 Single Photon Avalanche Diodes

Geiger mode APDs specifically employed to perform single photon detection are commonly known as Single Photon Avalanche Diodes. Due to their outstanding gain, which is attained by biasing the p-n junction with a reverse voltage higher than the breakdown voltage V_{BD}, SPADs are intrinsically digital devices. The output current reaches values different from zero only if an avalanche is triggered as a consequence of photon absorption or thermal generation of carriers. In this case, the output signal provided to the subsequent electronics can be interpreted as logic 1. If no electromagnetic radiation is absorbed within the active region or no dark pulse event is triggered, the output current remains at the low logic value, represented by the leakage current of the structure. By exploiting the intrinsic SPAD digital nature, it is possible to design complete readout channels consisting of only digital blocks. The main advantages brought about by a fully digital approach are represented by the very low power consumption, which is a typical feature of CMOS digital electronics, and the powerful processing capabilities offered by the digital domain. The fundamental operating principle of SPADs is illustrated in Fig. 1.7. Four distinct phases can be identified: *idle*, *avalanche* (*build-up*), *quenching*, and *recharge* (or *reset*).

- *Idle*. During this phase, the SPAD is subjected to a bias voltage exceeding the breakdown threshold. The discrepancy between the reverse bias voltage and the breakdown voltage is referred to as the excess voltage (V_{EX}). Consequently, the device is maintained in an idle state with an applied excess voltage greater than zero. In this stage, the electric field within the depletion region is significantly high. However, since no particles are detected, no carriers are generated within this region (assuming negligible noise contributions), resulting in no current flow. Essentially, the SPAD remains in an OFF (idle) state, waiting for the arrival of incident particles.

- *Avalanche*. Once an absorption event transfers sufficient energy to the lattice, an initial electron-hole pair is generated within the depletion region. The free carriers are rapidly accelerated by the electric field,

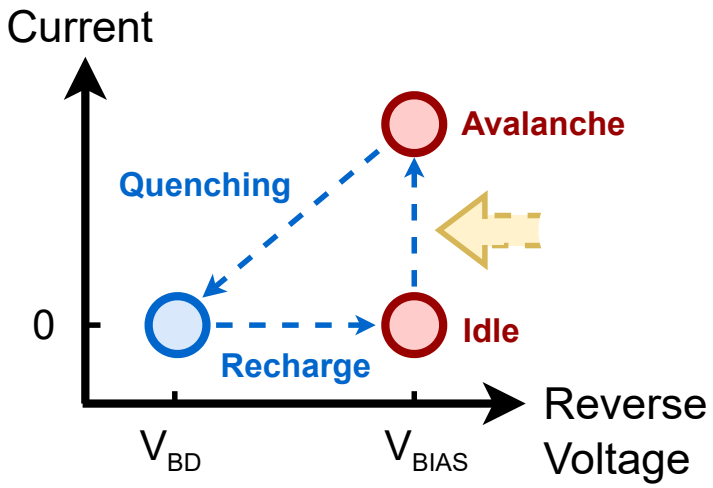


Figure 1.7: SPAD output current, as a function of the applied reverse voltage, in the four operating phases. Red circles represent the states where the SPAD can stay indefinitely if no external activity interferes (internal noise is neglected).

thus causing a cascade of impact ionization events which initiate a self-sustaining avalanche. In this phase, also referred to as build-up phase, the current exhibits exponential growth (as shown in Fig. 1.6). This current value is determined by the excess voltage and the intrinsic resistance of the device, R_S. The equivalent circuit of a SPAD is depicted in Fig. 1.8 [23][24][25]. The SPAD model includes the diode junction capacitance C_D and the series resistance R_S. The negative feedback provided by R_S, associated with the space-charge region, ensures the stability of the current against fluctuations. If the current increases, the voltage drop across R_S gets higher, resulting in a reduction of the voltage across the depletion region and subsequently causing a decrease in the current. The rising transient of the current, evolving from the OFF state to its steady-state value, typically lasts tens of picoseconds. During the build-up phase, the SPAD becomes insensitive to additional impinging particles since the avalanche has already been triggered and the generated current is independent of the number of particles detected, unlike in linear mode. Indeed, if another particle arrives, it may lead to the generation of another electron-hole pair. However, the new additional carriers are overshadowed by the large number of pairs already generated

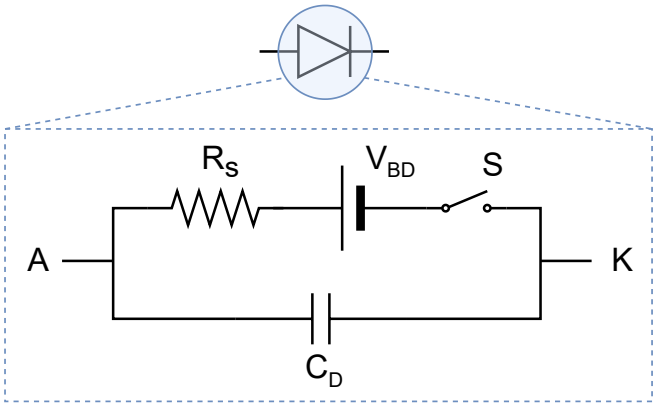


Figure 1.8: Equivalent circuit of a SPAD.

during the avalanche process, resulting in no discernible change in the current.

- *Quenching.* During the quenching phase, the junction bias voltage gets reduced by an external circuit, referred to as quenching circuit, in order to damp the avalanche. The reverse voltage applied across the junction is decreased until it approaches the breakdown voltage. As the quenching operation is complete, $V_{EX} = 0$ and the Geiger mode conditions are no longer met. As a result, the current decreases, going back to the OFF value, which is represented by the leakage level (intended as zero-current condition). More information about quenching circuits, involving a comparison between different implementations, will be provided in the following of this chapter.

- *Recharge.* Once the avalanche is quenched, the excess voltage V_{EX} needs to be restored to its original value. The time needed to bring the SPAD back to the initial condition, referred to as reset or recharge time, is the main parameter determining the maximum rate of detectable particles. The overall parasitic capacitance attached to the output node of the SPAD (which can be the anode or the cathode) affects the reset time, which in general should be kept as low as possible. Once V_{EX} is restored, the SPAD is ready for the detection of another event.

In summary, a SPAD can be likened to a bistable circuit. The two stable states, where the sensor can stay indefinitely, are represented as red circles in Fig. 1.7. In the first state the SPAD is idled, waiting for the detection

of electromagnetic radiation or a dark noise event. As soon as the avalanche is triggered the SPAD enters the second state, which is characterized by a potentially never-ending flow of current. An external circuit, suppressing the multiplication process, is needed to let the SPAD switch from the avalanche state to the idle one, passing through an intermediate state. Therefore, in principle, the readout operation can be accomplished using a comparator, as the output signal of a SPAD is simply a fixed-amplitude current pulse.

1.3.1 Key parameters and Figures of Merit (FoMs)

SPAD features can be discussed by defining a set of figures of merit that capture the key properties of these devices. These parameters, some of which have already been introduced in previous sections, will now be defined and examined in further details [2][4][26][27][28][29][30][31][32][33][34][35][36][37][38][39][40] [41][42][43][44][45][46][47].
Some of the properties which are going to be discussed in this section can be extended to the more general case of APDs. Hence, if applicable, a broader discussion, beyond the restricted field of SPADs, will be provided.
Gain. Gain represents the ratio between the output current measured at the detector terminals and the internal photocurrent prior to the avalanche multiplication process. This parameter is particularly relevant for linear APDs where the photocurrent is directly proportional to the incident photons. However, in the case of Geiger avalanche detectors, the concept of gain becomes meaningless, since the multiplication continues until the avalanche is externally quenched. Under this assumption, the gain is assumed to be infinite.
Photon Detection Probability (PDP). The photon detection probability (PDP), which is a function of the incident light bandwidth, represents a measure of the percentage of incoming photons that triggers the generation of an output pulse in the photodiode. The PDP depends on the product between the quantum efficiency (η_Q), expressing the probability of a photon being absorbed in the detector, hence generating an electron-hole pair, and the avalanche probability (P_{AV}), which indicates the likelihood of the primary carriers triggering an avalanche that does not terminate prematurely. The quantum efficiency is mainly influenced by the detector material, as different materials exhibit different responses across a range of wavelengths. For silicon detectors, typical values of quantum efficiency can be found around 80% in a set of wavelengths enclosing the visible spectrum ($400 - 1000\ nm$). Beyond this range, undesired effects, such as reflections and absorption taking place within the device optical layers, can limit the detection of the incident radiation [26]. The avalanche probability can be maximized by maintaining high

electric field values within the active region of the detector. In the specific case of SPADs, a high excess voltage is desirable to optimize the probability of triggering an avalanche. Rather than the absolute value of V_{EX}, the ratio V_{EX}/V_{BD} is the real valuable parameter to consider. Indeed, an excess voltage of 3 V can be remarkable for a breakdown voltage around 18 V, but it may not be significant for a breakdown voltage ranging around 100 V. However, it is worth specifying that the avalanche probability increases linearly with the excess voltage until it eventually saturates. Beyond this limit, a further increase in the excess voltage will lead to degradation in dark count rate (DCR), with very small advantage in terms of PDP [27]. The PDE (photon detection efficiency) can be obtained starting from the PDP, by taking into account also the fill factor of the sensor. Indeed, dead areas between cells, where photons may go undetected, may affect the overall detection efficiency of the system [28]. The relationship between PDE and PDP can be expressed as follows [48]:

$$PDE = FF \times PDP, \tag{1.13}$$

where FF represents the fill factor of the sensing structure. In array detectors, as well as in CMOS sensors fabricated in the same substrate of the processing electronics, the PDE can be sensibly lower than the PDP, due to the presence of non negligible dead areas strongly reducing the fill factor. Examples of PDE values found in the literature include: 52% for a 200 μm-diameter SPAD with $V_{EX} = 5\ V$ at $\lambda = 500\ nm$ [27] and 44% for a 6.4 μm-diameter SPAD in a 90 nm CMOS technology at $\lambda = 690\ nm$ [29].

Dynamic Range (DR). Typically, the dynamic range provides insight into the sensor ability to effectively capture both bright highlights and dark shadows within a scene. In the case of APDs, it is determined by the ratio between the sensor maximum counting rate and the level of noise caused by dark counts. The dynamic range is mainly influenced by the bias voltage and the technology employed.

Timing resolution (jitter). In APDs, the timing resolution, also known as jitter, usually refers to the statistical variations in the time interval between the arrival of photons at the detector surface and the leading edge of the output pulse [30]. It quantifies the precision of the timing measurement performed by the detector. The latency time, which is the time delay between the arrival of a photon and the onset of the avalanche, is not a constant number, due to physical factors associated with the photogeneration and avalanche processes. A factor contributing to latency fluctuations is the uncertainty in the location where the initial electron-hole pair is generated within the detector structure. If the photon is absorbed in the depletion layer, a short latency time is observed because the carrier experiences a strong electric field and is rapidly

accelerated towards the multiplication region. On the other hand, if the carrier is generated in the neutral region, it undergoes diffusion and takes some time to reach the depletion region (this happens only in a limited number of cases) where it can be accelerated [2]. In addition, the statistical behavior of the avalanche process itself contributes to the time variations in the latency time. Jitter can be characterized by using a laser source generating an incident photon beam. The corresponding output signal, provided by the sensor, is then measured by observing the Full Width Half Maximum (FWHM) of the time variation in the resulting avalanche pulses. In [31], a 10 μm-diameter square-shaped SPAD, produced with a 150 nm CMOS technology and operating at 468 nm, achieved a jitter of 52 ps (FWHM), while authors in [49] have reached a time resolution of 12 ps (FWHM) by using a wavelength of 515 nm and circular SPADs with a diameter of 25 μm, fabricated in a 180 nm CMOS technology.

Dead time. Once the avalanche is triggered, SPADs become inert to subsequent incoming photons, since other electron-hole pairs, that can be photogenerated within the depleted region, do not change the current significantly. The time interval between the triggering of an avalanche and the following reset phase, including the latter, is known as the dead time. It should be noted that during the reset phase, the SPAD is not completely insensitive to the incoming light, since the excess voltage gradually increases and the probability of triggering another avalanche increases as the bias voltage approaches the steady-state level. The duration of the dead time is primarily influenced by the specific quenching circuit employed. Further discussion about the impact of quenching circuits on this parameter can be found in the following of this chapter. Typically, values for the dead time can range from few nanoseconds [50] to some microseconds [31].

Fill Factor (FF). Fill factor is calculated as the ratio between the active area of the SPAD and the total area of the pixel. The active area can be constrained by design choices, such as the inclusion of guard rings, which can be implemented to ensure a uniform electric field distribution across the active region of the SPAD, or by the presence of in-pixel electronics like quenching and processing circuits, which are part of the pixel but not part the photodiode itself. Maximizing the fill factor can be a hard task, especially in the design of monolithic sensors, where the implementation of processing circuits is carried out at the pixel level. In [32], the authors attained a fill factor of 19.5% for a sensor chip, fabricated in a 150 nm CMOS technology, consisting of pixels equipped with p+/nwell square-shaped SPADs and dedicated time to digital converters (TDCs).

Noise. Noise in SPADs appears as random occurrences of dark current pulses. The rate of avalanche events triggered by non-photon-generated carriers is referred to as the Dark Count Rate (DCR), and it is measured in counts per second (cps) or Hertz (Hz). The main challenge lies in distinguishing between an avalanche caused by an incident photon and one resulting from a dark count event. In order to ensure proper detector operation, minimizing DCR is of paramount importance. Indeed, in some cases, DCR represents the main limiting factor in using SPAD-based detectors.
The noise sources include primary events, i.e. noise events not correlated to previous avalanche events taking place in the junction, and secondary (or correlated) events [39]. Regarding primary dark events, various noise sources can be identified, depending on the carrier generation process and the location of non-equilibrium carriers [40]. Fig. 1.9 shows a band diagram representation describing the different carrier generation mechanisms which can take place in a p+/nwell SPAD. Firstly, thermally excited minority carriers situated in the bulk region of the pn junction might migrate, by diffusion, towards the depletion layer, thereby initiating an avalanche event. In second place, thermal generation of non-equilibrium carriers can occur within the depletion layer due to direct band to band generation, trap-assisted tunneling, or trap-assisted (or indirect) generation. Eventually, under conditions of extremely high elec-

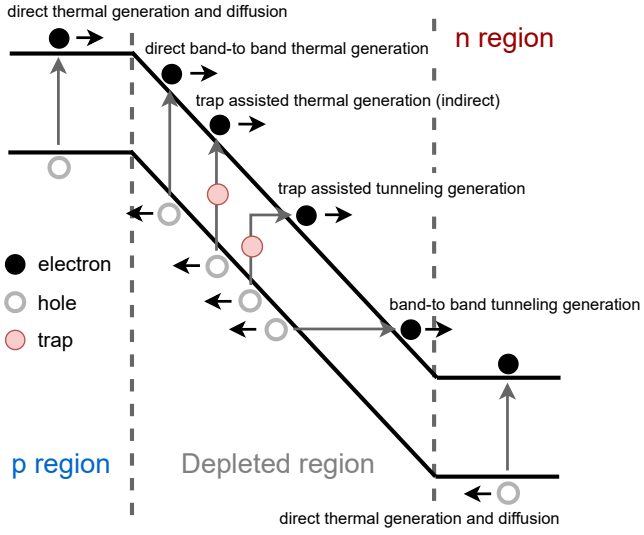


Figure 1.9: Energy band diagram showing the different carrier generation mechanisms underlying the DCR, in a p+/nwell SPAD.

tric fields, non-equilibrium carriers may be produced through band to band tunneling within the depletion region. Actually, the contribution to DCR represented by the direct thermal generation of carriers in the two neutral regions is negligible, due to the high recombination rate featured by minority carriers travelling through the bulk sections of the device.
Under the assumption that primary generation inside the depletion region is the sole origin of dark count pulses, the probability density function of the dark counts is expected to adhere to Poisson statistics. Hence, the probability of having n avalanche events within the specified time frame T can be expressed as follows:

$$P(n,T) = \frac{T^n DCR_{av}^n}{n!} e^{-DCR_{av}T}, \tag{1.14}$$

where DCR_{av} represents the average DCR, calculated as the product between the avalanche triggering probability and the average number of carriers in the depletion region per unit time. From (1.14), the probability of observing at least one avalanche event in a time window of 1 s ($n \geq 1, T = 1\ s$) can be found as:

$$P(n \geq 1, T = 1\ s) = 1 - P(0) = 1 - e^{-DCR_{av} \times 1s}. \tag{1.15}$$

In the noise distribution, any departure from the Poisson statistics may indicate the contribution of, for instance, afterpulsing phenomena. Therefore, studying the distribution of the noise events is a valuable way to extract information about the nature of the DCR pulses.
In the case of silicon material, the indirect thermal generation phenomenon overwhelmingly prevails over the direct band to band one, thus the latter can be neglected with very small error. The trap-assisted thermal generation rate can be expressed, according to the Shockley-Read-Hall (SRH) recombination theory, as

$$G_{SRH} = \frac{pn - n_i^2}{\tau_n \left[p + n_i e^{\frac{-(E_t - E_i)}{kT}}\right] + \tau_p \left[n + n_i e^{\frac{(E_t - E_i)}{kT}}\right]}, \tag{1.16}$$

where τ_n and τ_p are respectively the electron and hole lifetimes, which inversely depend on the density of recombination centers and on the capture cross area of the carrier, n_i is the intrinsic carrier concentration, and E_t is the energy level associated to the recombination center. If a relatively strong electric field is considered (not higher than $9 \times 10^5\ V/cm$), a field effect enhancement factor (Γ) can be introduced in (1.16), in order to model the thermal generation of

carriers by trap assisted tunneling effects:

$$G_{SRH,TAT}(x) = \frac{pn - n_i^2}{\frac{\tau_n}{1+\Gamma}\left[p + n_i e^{\frac{-(E_t - E_i)}{kT}}\right] + \frac{\tau_p}{1+\Gamma}\left[n + n_i e^{\frac{(E_t - E_i)}{kT}}\right]}, \tag{1.17}$$

with

$$\Gamma = 2\sqrt{3\pi}\frac{|\mathcal{E}(x)|}{F_\Gamma} e^{\left(\frac{|\mathcal{E}(x)|}{F_\Gamma}\right)^2}, \tag{1.18}$$

where F_Γ is a constant. For the sake of completeness, it is worth specifying that (1.17) is not relevant to the Poole-Frenkel effect, as this contribution, in presence of relatively strong electric fields, is negligible if compared to tunneling effects. The resulting DCR, taking into account both traps-related thermal generation (DCR_{SRH}) and trap assisted tunneling generation (DCR_{TAT}), can be calculated as:

$$DCR_{SRH} + DCR_{TAT} = S_{depl} \cdot \int_{W_1}^{W_2} P_{AV}(x) \cdot G_{SRH,TAT}(x)\, dx, \tag{1.19}$$

where S_{depl}, W_1 and W_2 represent respectively the area and the boundary positions of the depletion region, while P_{AV} is the avalanche triggering probability, describing the likelihood of a carrier generated within the high field region to effectively initiate an avalanche.
If the electric field applied to the depletion region exceeds the conventional value of 9×10^5 V/cm, direct band to band tunneling becomes a non negligible noise source. In silicon devices, the generation rate of band to band tunneling can be modeled as:

$$G_{BTBT}(x) = \alpha \cdot |\mathcal{E}(x)|^\beta \cdot e^{\frac{-\gamma}{\mathcal{E}(x)}}, \tag{1.20}$$

where, at room temperature, $\alpha = 4 \cdot 10^{14}$ $cm^{-1/2} \cdot V^{-5/2} \cdot s^{-1}$, $\beta = 2.5$ and $\gamma = 1.9 \cdot 10^7$ V/cm [51].
As a result, when a high electric field is applied, the band to band tunneling process should be considered, along with the Shockley-Read-Hall thermal generation and the trap assisted tunneling. The overall dark count rate (DCR_{TOT}) is then obtained by adding the various contributions coming from all the noise mechanisms discussed so far:

$$\begin{gathered} DCR_{TOT} = DCR_{SRH} + DCR_{TAT} + DCR_{BTBT} = \\ S_{depl} \cdot \int_{W_1}^{W_2} P_{AV}(x) \cdot (G_{SRH,TAT}(x) + G_{BTBT}(x))\, dx. \end{gathered} \tag{1.21}$$

Therefore, the electric field across the junction plays a crucial role in determining the noise performance of the sensor, affecting both the thermal generation of carriers and the band to band tunneling phenomenon. Generally, at room temperature, trap-assisted tunneling is the most significant noise contribution for several SPAD architectures implemented in deep sub-micron CMOS technologies. However, as the technology node is scaled down, the doping level of the active area increases, with a consequential enhancement of the electric field within the depletion region, for a given V_{EX}. This entails a significant rise in dark counts originated from band to band tunneling, which may become the primary noise source in CMOS technologies with very large scale of integration [41][40].
Among the noise events labelled as secondary, or correlated, afterpulsing and crosstalk provide the main contributions to DCR. Optical crosstalk is a dark pulse source that turns to be effective especially when SPADs are arranged in array configuration [43]. An interaction between neighboring photodiodes may take place, so as the operating condition of one SPAD, called screamer, can affect the performance of the adjacent ones. High-energy carriers, generated within the high field region, might induce the emission of electroluminescent photons, which, in turn, could be absorbed by a neighbouring pixel, hence initiating an avalanche [42]. Even though, in the adjacent SPADs, the multiplication process is triggered by photogenerated carriers, the corresponding output pulse has to be treated as an actual noise event, since the latter is a direct consequence of a prior screamer pulse. Crosstalk mainly depends on the array density, on the SPAD structure and on the chip thickness. As a matter of a fact, the silicon substrate can act as a planar waveguide, greatly enhancing the long-distance transmission of secondary photons. In [44], a crosstalk coefficient of $0.4 \cdot 10^{-3}$ has been measured for a SPAD located at $0.4\ mm$ far from the screamer one, in a chip with 50 μm substrate fabricated in a 150 nm CMOS technology.
Afterpulsing phenomena have origin from deep energy levels located between the band edge and the mid-gap. When an avalanche occurs, some carriers may get trapped in deep levels and released after a random time, that depends on the time constant of the trapping level. The new carrier generated in this way can gain enough energy, due to the effect of the electric field, to cause an impact ionization and, thus, trigger another avalanche. This kind of dark pulse is referred to as "secondary", since it is a follow-up event of a preceding avalanche [35].
During an avalanche, the probability of capturing a carrier by a deep level of

energy can be written as [52]:

$$P_{cap}(t_g) = 1 - e^{-\frac{t_g}{\tau_{cap}}}, \tag{1.22}$$

where t_g is the time duration of the avalanche event, referred to as gate length, and τ_{cap} represents the capture time constant calculated as $\tau_{cap} = q/J \cdot \sigma$, with J symbolizing the current density, and σ the defect cross-sectional area. In order to minimize the capture probability, thus acting on the afterpulsing phenomenon incidence, a possible way consists of reducing the overall avalanche charge flowing through the space charge region, by means of suitable current control circuits. The probability of carrier releasing after a time t, from the moment when it is captured, can be expressed as:

$$P_{rel}(t) = \frac{1}{\tau_{rel}} e^{-\frac{t}{\tau_{rel}}}, \tag{1.23}$$

where τ_{rel}, namely the trap lifetime constant, indicates the average time during which a carrier stays in the trap level and can be expressed as

$$\tau_{rel} = \tau_0 \, e^{\frac{E_r}{kT}}, \tag{1.24}$$

with τ_0 a time coefficient, E_r the defect activation energy, k the Boltzmann's constant and T the absolute temperature. Since the likelihood of an afterpulsing phenomenon occurring after the quenching operation is higher at time intervals close to the primary avalanche, a circuit able to provide the sensor with a hold-off time can potentially mitigate the triggering of secondary pulses. Although such a quenching system is employed, afterpulsing may occur in any case, when the off time interval is over. As the temperature increases, the afterpulsing effect may reduce due to the trap lifetime reduction, making deep levels more likely to release trapped carriers during the hold-off time. Since afterpulsing highly depends on the number of trapping centers located within the depletion region, a cleaner manufacturing process, with very low concentration of impurities and crystal defects, significantly contributes to the minimization of secondary avalanches [31].
As shown in Fig. 1.10, beside representing additional noise sources, secondary dark events are also responsible for the stretch out of the reset time, undermining the SPAD detection capabilities. Each photon falling within this time interval may be lost by the detector, since the bias condition of the SPAD may not ensure the same PDE featured in the idle phase. Depending on the threshold level of the comparator used to discriminate the SPAD output signal and on the architecture of the readout circuits, some secondary events can go

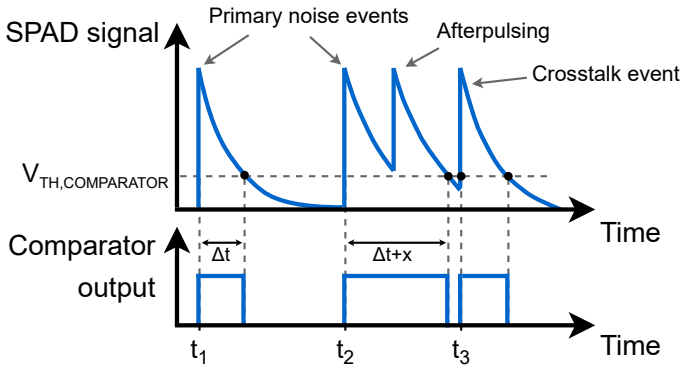


Figure 1.10: Effect of secondary pulses on the reset time.

unrevealed by the subsequent electronics, even though contributing to reset time elongation.
In summary, the SPAD noise, in the form of DCR, is a non-desired, complex phenomenon originated from both primary and correlated sources. It represents the main limiting factor for SPADs produced in CMOS technologies, since the latter are based on fabrication processes with a relatively high concentration of lattice defects. The main parameters to play with, in order to tune the SPAD noise performance, are: the active area, the excess voltage, affecting also the PDE with a reverse proportionality factor if compared to DCR [53], the temperature, the technology and the fabrication process.
DCR values around 7.7 $Hz/\mu m^2$ can be found in [45] for SPADs fabricated in a 180 nm CMOS technology with high voltage option, exploiting a p+ diffusion over a deep n-well junction ($V_{EX} = 1.4\ V$, $T = 25°C$). In [46], where a DPW/BNW (deep p-well over a buried n-well) junction is used, SPADs fabricated in a standard 55 nm BCD (Bipolar-CMOS-DMOS) technology feature a DCR median value of 2.6 $Hz/\mu m^2$, measured on 10 samples located in different chips ($V_{EX} = 7\ V$, $T = 20°C$).
Breakdown voltage (V_{BD}). In an avalanche photodiode, the breakdown voltage represents a threshold value determining the behaviour of the detector. If the bias voltage is maintained below the breakdown voltage, the photodiode operates in linear region, with the output current exhibiting a proportional response to the incoming light. However, in this mode, the photodiode cannot be operated as single photon detector, due to its moderate gain. On the other hand, if the bias voltage exceeds the breakdown voltage, the photodiode

enters the non-linear region, thus becoming able to generate a self-sustaining avalanche even with a single photon being absorbed in the active region. As mentioned before, the difference between the bias voltage and the breakdown voltage (both considered as absolute values) gives the excess voltage, V_{EX}, which expresses how deep inside the non-linear mode the device is being operated.

The breakdown voltage is a key parameter for SPADs, deserving particular attention especially during the sensor design phase, in order to optimize performance and allow a proper working condition. In particular, when designing a pixelated detection system, the uniformity of the breakdown voltage is one of the most important features to take into account. Since the same bias voltage is applied to all the SPADs of the array, any variation in the breakdown voltage can result in non-homogeneous PDE and DCR across different regions of the chip. Within the single SPAD, premature edge breakdown represents a common uniformity-related problem affecting the breakdown voltage. The distribution of the electric field across the junction of a SPAD is non-uniform, with corners and edges experiencing higher electric fields compared to the center of the multiplication region. This uneven electric field distribution increases the avalanche probability in the peripheral regions, leading to a higher level of dark count rate (DCR) and potentially compromising the detector operation. To mitigate this effect, a circular shape of the sensor is commonly used [54][55][56][57][58], as it reduces the number of critical points, making the overall junction geometry smoother. However, some layout rules, typical of CMOS processes, may restrict the use of circular shapes [37]. In such cases, alternative techniques can be adopted. The specific choice of PEBP (premature edge breakdown prevention) techniques depends on the process features available for SPAD fabrication.

The breakdown voltage of a detector is affected by several factors, including the fabrication process, the doping profile of the junction and the temperature. As the temperature increases, the lattice atoms experience more thermal vibration, thus reducing the mean free path length of carriers. As a result, at higher temperature, a free carrier accelerated by the electric field will feature a smaller kinetic energy between collisions, hence decreasing the probability of causing impact ionization. Since the breakdown voltage is strictly related to the impact ionization probability, it can be expressed as a function of temperature, according to

$$V_{BD} = V_{BD0}[1 + \beta(\Delta T)], \tag{1.25}$$

where V_{BD0} represents the breakdown voltage at room temperature, β is a coefficient depending on the fabrication process, typically around 10^{-3} V/K

for silicon sensors, and ΔT is the temperature variation [38]. A breakdown voltage of about 20 V ($T = 20°C$) and a β ranging between 15.9 mV/K and 16.2 mV/K have been obtained in [47], where the authors show the results from the characterization of SPADs fabricated in a 110 nm CIS (CMOS Image Sensor) technology, relying on a p+/n-implant junction.

1.3.2 Quenching circuits

Once a certain number of electron-hole pairs is generated within the high-field region, a positive feedback loop is initiated, triggering a rapid surge in current. Since the internal gain of a SPAD is assumed to be infinite, the number of generated carriers inside the junction does not scale proportionally with the detected radiation. As a result, once the avalanche is triggered, the current persists, rendering the device incapable of detecting subsequent particles. In order to suppress the avalanche and restore the SPAD to its original quiescent state, an external circuit, known as the quenching circuit (QC), can be used. The time required to complete the quenching operation is referred to as the quenching time. While the current escalation occurs rapidly within some picoseconds, the suppression of the avalanche and the subsequent transition to the idle phase can be relatively slower operations, which may take time intervals around tens or hundreds of nanoseconds. These two phases significantly impact the timing performance of the device and can influence fundamental SPAD characteristics such as the power dissipation and the DCR. Therefore, the choice of the most appropriate quenching circuit represents an important step of the front-end circuit design, since it can affect the overall performance of the device.
Two groups of quenching circuits can be identified: passive quenching circuits (PQCs) and active quenching circuits (AQCs). Depending on the requirements of the specific application, a trade-off between performance, size, and cost has to be evaluated before choosing the most appropriate quenching architecture.

1.3.2.1 SPAD orientation

Regardless of the quenching approach used, a quenching circuit can be connected either to the anode or the cathode terminal of the SPAD. Fig. 1.11 shows two examples of quenching circuits with different SPAD orientation. In (a), the quenching operation is attained through a PMOS transistor that is connected to the SPAD by means of the the cathode contact. The sensor is biased by applying the negative pole of the bias voltage to the anode and the positive one to the source terminal of the transistor. In (b), the reverse

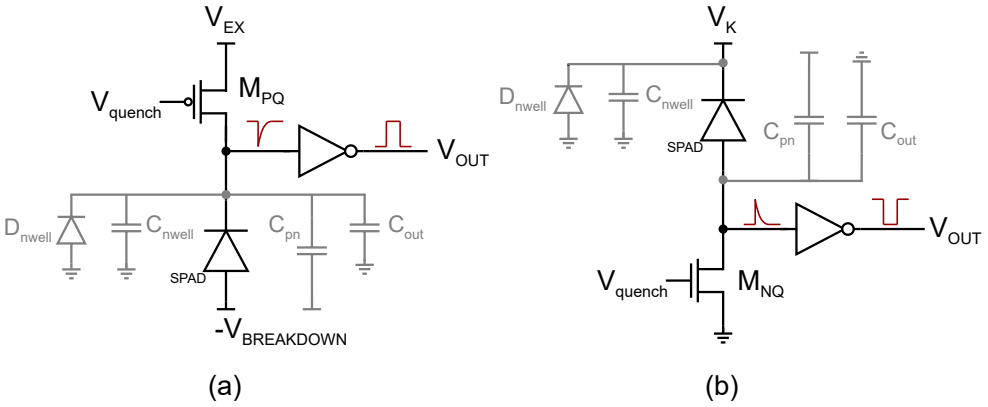


Figure 1.11: Quenching circuits with different SPAD orientation: in (a) the sensor is connected through the cathode to a PMOS transistor, in (b) the sensor is connected through the anode to an NMOS transistor.

bias is achieved by applying a positive voltage at the cathode of the SPAD, while the ground voltage is provided at the anode of the SPAD through an NMOS transistor, assuming a similar function to the PMOS employed in (a). In general, after the quenching circuit, which consists of a simple transistor in both the circuit diagrams in Fig. 1.11, the signal produced by the SPAD can be read in various ways, involving a comparator, an inverter or a couple of inverters working as a digital buffer. Both the quenching approaches shown in Fig. 1.11, have been extensively used in the literature (in [59] and [60] different applications of the circuit in (a) are shown, while configuration (b) is used, among others, in [61], [62] and [63]), although they feature different characteristics. In configuration (a), the cathode serves as SPAD output node, since it is physically connected to the input of the inverter. During the dynamic operation of the SPAD, a low pass filtering effect is provided to the sensor output signal by various parasitic capacitances, including the parasitic capacitance between the n-well and p-substrate C_{nwell}, the capacitance of the p-n junction C_{pn}, and the equivalent capacitance resulting from the inverter physical gates and interconnections (indicated as $Cout$ in Fig. 1.11). These capacitances, providing the signal generated by the SPAD with a relatively large time constant, may cause a significant slowdown in device operation. As a result, the exponential growth of the SPAD output signal is stretched out in time. Another consequence of a large equivalent capacitance is represented by the higher charge flowing through the detector, leading to increased power consumption and a higher afterpulsing probability.

In configuration (b), the output node is connected to the anode of the SPAD, allowing for optimization of the reset time and of the charge amount involved in the avalanche event. The overall parasitic capacitance connected to the SPAD output node is reduced if compared to the (a) configuration, as C_{nwell} gives no more contribution to the overall parasitic capacitance. Additionally, NMOS transistors exhibit higher electron mobility (μ_n) than PMOS ones, resulting in a lower equivalent quenching resistance. The advantages brought about by the approach in (b) enables the creation of small circuits, based on the NMOS-anode approach, having the same current capability, with lower time constants, as circuits relying on the PMOS-cathode configuration [4]. Furthermore, limitations on the lowest level of $V_{breakdown}$ must be considered in circuit (a), in order to prevent the n-well to p-substrate parasitic diode from being set in forward bias. This kind of drawback makes little sense in the (b) case, since V_K is kept at a voltage level always larger than the ground voltage. Regardless of the SPAD orientation, it is important not to exceed the maximum gate oxide potential featured by the subsequent core electronics to prevent permanent damage to the readout channel. Hence, it is common practice to use thick gate oxide transistors, if available in the technology in order to ensure a safe operating region for the following transistor gates.

1.3.2.2 Passive quenching circuits

A straightforward method to suppress the SPAD avalanche can be obtained by connecting a high-value ballast resistor, R_B, in series with the photodiode [64], as illustrated in Fig. 1.12. During the idle phase, R_B biases the SPAD to the steady-state voltage without any current passing through the resistor. Once an avalanche is initiated, the current rapidly reaches the maximum level, which is determined by the excess voltage V_{EX} and the internal resistance of the photodiode:

$$I_{PEAK} = \frac{V_E}{R_S}. \tag{1.26}$$

If a high impedance is provided by the following electronics, generally implemented by means of digital gates, I_{PEAK} flows entirely through the quenching resistor R_B, causing the discharge of the parasitic capacitance, C_P, connected to the output node. The time constant of the discharge phase depends on the equivalent resistance with respect to ground, which is featured at the anode,

$$\tau = C_P(R_S//R_B) \simeq C_P R_S, \tag{1.27}$$

which is a reasonable approximation, since, generally, $R_B \gg R_S$. When the transient is over, the total voltage across the series connection SPAD-R_B has

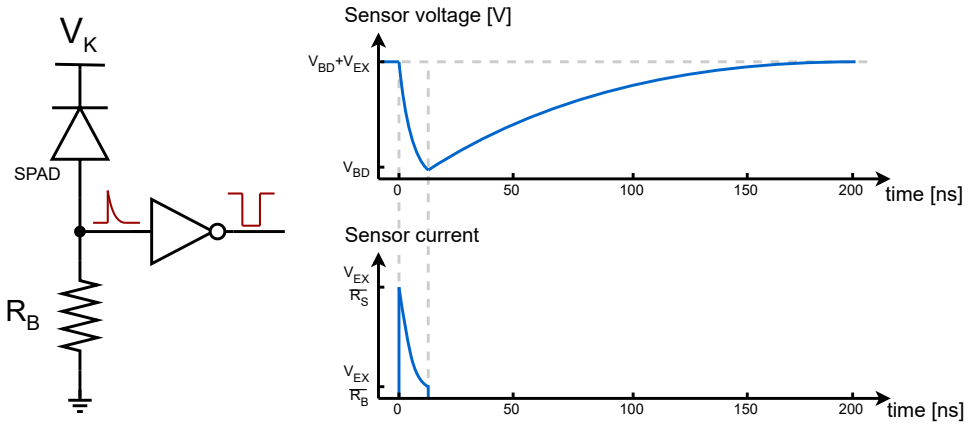


Figure 1.12: Example of a passive quenching circuit consisting of a simple resistor R_B, with the relevant waveforms.

not changed. At the end of the discharge phase a voltage, equal to V_{EX}, is transferred from the SPAD to the quenching resistor. The final current flowing out from the anode node is

$$I_F \simeq \frac{V_{EX}}{R_B}. \tag{1.28}$$

It can be noticed that the reverse voltage applied to the SPAD never drops below the breakdown voltage of the junction. Therefore, in principle, the avalanche cannot be suppressed and the current I_F may keep flowing through the sensor indefinitely. If the final magnitude of the current is sufficiently large, the avalanche can be self-sustaining, as a high number of free carriers are simultaneously present within the junction. In such a scenario, if the power dissipation exceeds the device thermal limits, excessive heating can result in permanent damage to the detector [35]. Conversely, if the current I_F is reasonably small, the amount of charge is insufficient to sustain the avalanche, leading to its quenching after a given time window. For SPADs manufactured in recent technologies, a threshold value for the final current, conventionally capable of discriminating between a not conclusively avalanche and a successful quenching operation, can be set to 100 μA [64]. As it can be inferred from the (1.28), given a certain range of excess voltage within the SPAD is expected to work, I_F can be set by choosing a suitable value for R_B. Even with the provision of appropriate conditions for the avalanche damping, the exact quenching time cannot be precisely determined due to the statistical nature of the avalanche process. However, it is possible to provide a first order

estimation of the time interval taken to perform the quenching operation. The quenching time can be approximated as the duration between the moment when the current reaches its maximum value and the moment when it drops below the threshold value I_{TH}, after the exponential decay:

$$t_Q = \tau ln(\frac{I_{PEAK} - I_F}{I_{TH} - I_F}). \tag{1.29}$$

Depending on the manufacturing technology and on the dimension of the photodiode, typical values for the quenching time range between few nanoseconds and tens of nanoseconds. The most limiting factor is represented by the intrinsic junction capacitance, which dramatically affects the discharge time constant and poorly scales down with the technology scaling.
Once the avalanche is successfully quenched, the SPAD current drops to zero, marking the beginning of the reset period. During this phase, the voltage at the anode decreases, thereby restoring the initial operating conditions of the SPAD. PQCs like the one discussed so far suffer from disadvantages related to the reset time. In these circuits, the time interval required to discharge the anode node back to its steady-state voltage is considerably long, often several orders of magnitude greater than the quenching time. During the reset phase, no current flows through the detector, and the time constant of the exponential transient is determined by $R_B C_P$. A large reset time impacts on the maximum counting rate achievable by SPADs. Furthermore, during the reset phase, the SPAD remains biased above the breakdown voltage, thus potentially allowing for the avalanche triggering, even before the excess voltage reaches its steady-state value. Since the avalanche triggering probability strongly depends on the excess voltage applied to the sensor, the SPAD behaviour, during the reset phase, cannot be determined a priori. Depending on the time of arrival of the impinging photons, the resulting pulses of current may go unnoticed by the subsequent inverter. As a result, a high number of count losses can occur, particularly at higher counting rates where the majority of avalanches are triggered during the early phase of reset. If the excess voltage is prevented from reaching its quiescent value, the ability of the SPAD to operate reliably and predictably is compromised. The particular behavior of a detection system, that, under high counting rates, does not generate output pulses, but rather restarts the reset time, is referred to as "paralyzable" [35]. Therefore, it becomes evident that minimizing the reset time, by choosing a suitable R_B, is of paramount importance for the sensor performance.
In [65], a resistor based quenching circuit ($C_P = 70\ fF$ and $R_B = 270\ K\Omega$) has been used for circular SPADs with a diameter of 7 μm, fabricated in a standard 800 nm CMOS technology, attaining a total dead time of 35 ns.

The quenching network depicted in Fig. 1.12 can be improved by replacing the ballast resistor with an NMOS transistor. The resulting circuit is shown in Fig. 1.13. The n-channel MOSFET used to quench the avalanche works in the linear region during the first part of the quenching phase, thus the current is determined by

$$I_{d-Mquench} = 2\mu_n C_{ox} \frac{W}{L} \left[(V_{gs} - V_t) - \frac{V_{ds}}{2} \right] V_{ds}, \tag{1.30}$$

where μ_n is the electron mobility, C_{ox} is the gate oxide capacitance, W/L is the ratio between the width and the length of the transistor channel, V_{gs} is the gate to source voltage, which, in this case, corresponds to the quenching voltage V_{quench}, V_t is the transistor threshold voltage and V_{ds} is the drain to source voltage, that during the quenching phase is charged from zero up to the excess voltage level. The approach proposed in Fig. 1.13 offers significant advantages in terms of dead time, even though achieving a proper functionality of this circuit entails a meticulous choice of the W/L factor. Fig. 1.13 illustrates various reset transients as a function of different W/L ratios. If the W/L of the NMOS transistor is too small, the time required for the anode discharge will be longer, reducing the advantages provided by the use of an active device. In addition, this circuit provides the capability of dynamically controlling the quenching current through the V_{gs} of the transistor. Changing the MOSFET current affects the total charge flowing through the SPAD, thus influencing the power dissipation, since the energy involved in the SPAD operation depends

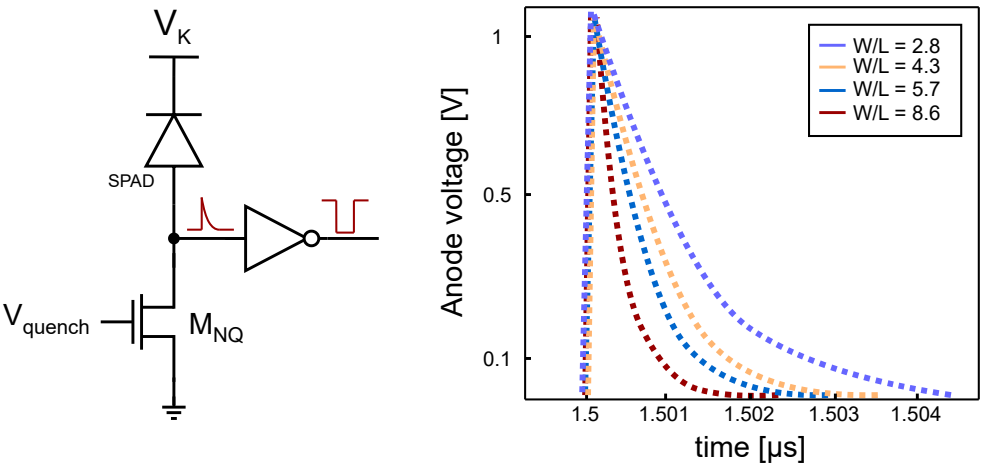


Figure 1.13: Passive quenching circuit with an NMOS transistor [34].

on the avalanche charge, and, in a moderate way, the afterpulsing probability. To summarize, passive quenching circuits offer simplicity, cost-effectiveness, and small area occupation. All of these features make the passive quenching approach suitable for the fabrication of large arrays with a high fill factor. However, it has some drawbacks, including a relatively slow reset transient which follows an exponential decay, the lack of a control system which is able to prevent the reset time from being reinitialized by random photons before the excess voltage is fully restored, and the inability to provide the SPAD with an effective mechanism to mitigate afterpulsing.

1.3.2.3 Active quenching circuits

AQCs are used to precisely control the quenching time, to reduce the reset period and to provide a solution against the afterpulsing phenomenon. The operating principle relies on the activation, upon the avalanche triggering, of an active feedback circuit, which regulates the bias voltage of the SPAD and recovers the idle state in very short time. Unlike passive quenching circuits, active ones allow the bias voltage to remain below the breakdown level for a specific duration, known as the hold-off time, which helps minimize afterpulsing. Additionally, the reset phase, following the hold-off phase, is carried out by using active devices, making the reset transient of active quenching circuits typically much faster compared to PQCs. In the case of passive circuits the reset transient follows an exponential decay, with a time constant depending on the ballast resistor. In AQCs, the SPAD bias voltage is restored by using active networks, which feature higher driving capability than a simple RC circuit. A faster reset transient reduces the likelihood of avalanches occurring during this phase, thus allowing higher maximum counting rates.
A first method of implementing an active quenching circuit is shown in Fig. 1.14. The circuit employs a feedback network consisting of a pulse generator controlling the SPAD bias voltage [64]. The output of the pulse generator, during the time interval when it is not active, can be considered as a high-impedance node. During the quenching process, the active network drives the bias voltage of the detector below the breakdown voltage, thus making the quenching time unaffected by the statistical behavior of the avalanche. As shown in the waveforms of Fig. 1.14, the current continues to flow at its peak value until the active network activates, as there is no high ballast resistor to perform the quenching operation (it can be worth specifying that the resistor shown in Fig. 1.14 has not the same scope of the quenching resistor R_B, used in Fig. 1.12, but it represents the internal resistance R_S of the sensor). To fully capitalize on the advantages brought about by the active quenching approach,

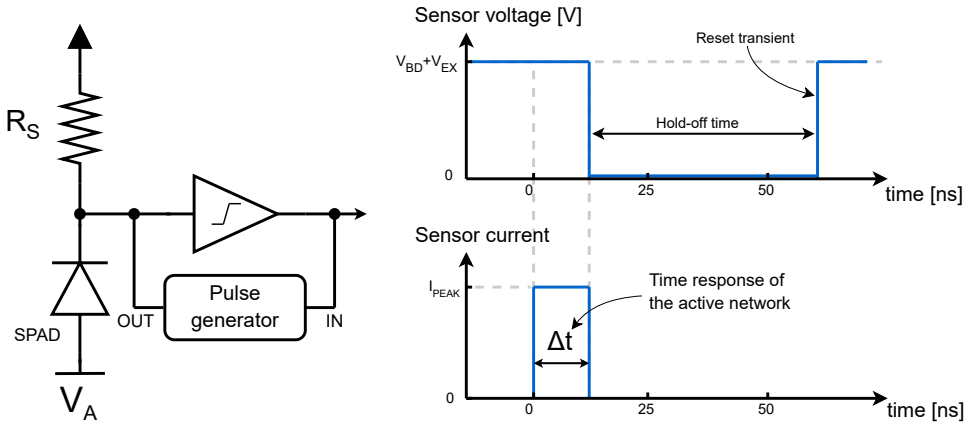


Figure 1.14: Active quenching circuit consisting of a feedback pulse generator (R_S represents the SPAD internal resistance).

the propagation time of the comparator and of the pulse generator should be as rapid as possible. If there is a significant delay between the triggering of the avalanche and the activation of the active devices, the quenching time may become longer compared to passive quenching circuits, thus allowing a large amount of charge flowing through the device. After the current starts flowing, the time interval needed to produce a quenching pulse is significantly influenced by the length of the interconnections between the anode and the comparator, as well as between the active devices and the SPAD. Minimizing the length of the interconnections can help in reducing the parasitic capacitances on the intermediate nodes, in order to ensure fast signal transmission and appropriate quenching pulses. By integrating the detector and the electronics on the same chip, the parasitic capacitance between the sensor and the quenching circuit can be effectively minimized. In general, providing smaller quenching time, if compared to PQCs, may not be trivial for active networks based on pulse generators. However, as shown in Fig. 1.14, precise hold-off time intervals and sharp reset transients can be achieved with this type of active quenching circuit.
An alternative way to quench the avalanche, with the active approach, is shown in Fig. 1.15. In this circuit, the feedback network is made of fast MOS switches (i.e. switches implemented by using MOSFETs), driven by a control logic circuit, which alternatively enables the pull-up or the pull-down branches. This circuit effectively addresses the quenching time issue of AQCs with pulse generator, by incorporating a ballast resistor R_B, which initiates

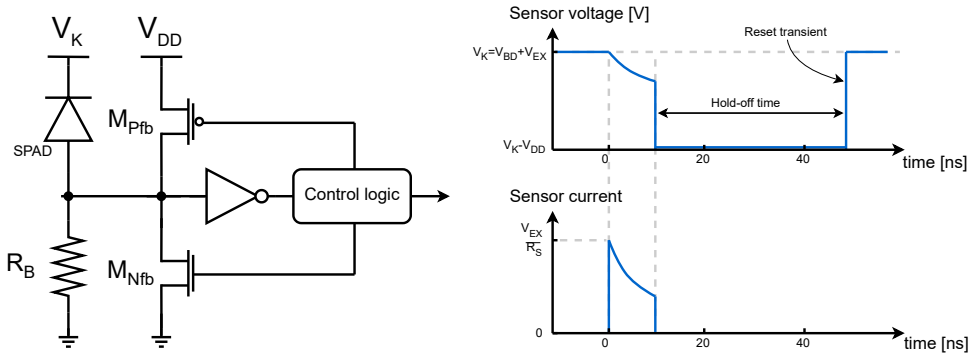


Figure 1.15: Active quenching circuit made of a ballast resistor and fast MOS switches.

the passive quenching of the avalanche even before the active feedback network is activated. While the resistor in Fig. 1.12 is in charge of entirely perform the quenching operation, the avalanche current flowing through the circuit of Fig. 1.15 is initially reduced by R_B in a passive way. The quenching transient is then completed by the PMOS feedback transistor, once it is activated. Therefore, the primary purpose of the ballast resistor is to start the discharge of the SPAD bias voltage, anticipating the triggering of the fast active network. As soon as the PMOS transistor is activated, the diode voltage drops to a voltage equal to $V_K - V_{DD}$, and the current is rapidly quenched. After the current shutdown, the PMOS transistor stays active for a determined amount of time, thus providing a hold-off time to the SPAD, in order to prevent any afterpulsing event from taking place. Eventually, the SPAD voltage is restored to its steady-state value through the NMOS transistor M_{Nfb}, which pulls down the anode voltage during a fast reset phase.
In Fig. 1.16, a modified version of the previous circuit, where the ballast resistor is replaced with an NMOS transistor, is shown. This configuration enables the M_{NQ} transistor to handle the entire quenching process, resulting in a reduced quenching time, as shown in the corresponding time diagram. The PMOS transistor serves only as a pull-up transistor, maintaining the SPAD at a bias voltage lower than the breakdown voltage for a programmable duration. Consequently, the active network has more relaxed requirements in terms of time response, as the feedback transistors are not directly involved in the quenching phase.
In summary, the choice of the quenching circuit depends on the specific application. Active quenching circuits provide fixed and short quenching times,

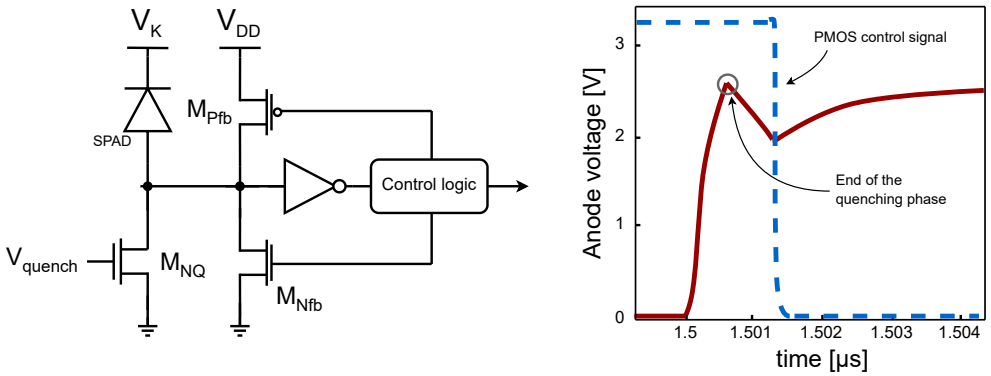


Figure 1.16: Active quenching circuit made of an NMOS transistor, entirely performing the quenching operation, and fast MOS switches.

reduced reset times, and the ability to keep the SPAD voltage below the breakdown level for a programmable time interval. However, they come with increased complexity and larger area occupation, which can negatively impact the fill factor in array configurations. Conversely, a passive quenching circuit provides a simpler and more compact solution, although it may have limitations in terms of quenching time and reset transient.

1.3.3 SPADs in CMOS technology

Over the last decades, substantial research activity has been focused on the fabrication of high performance APD sensors in CMOS technologies. The latter offer numerous advantages, motivating the fabrication of CMOS sensors, which can compete with detection systems produced in dedicated technologies. The integration of the sensing element and the readout electronics on the same chip, the possibility of creating high-density arrays with remarkable fill factors, the use of low voltages coupled with low power consumption have driven the development of SPAD sensors in well-established CMOS technologies. Despite a number of drawbacks coming from the adaptation of the SPAD structure to CMOS technologies, dense imaging systems based on CMOS SPADs have been successfully developed [66]. The main limitation is represented by the poor flexibility of the CMOS fabrication processes as far as the design of the sensing element is concerned. The strict constraints imposed by the foundries leave a relatively low degree of freedom to the designers, that may be forbidden to create particular SPAD structures, aimed at optimizing the detector performance [27]. Additionally, the PDE of CMOS SPADs is quite poor if

compared to the one which can be achieved with dedicated technologies, taking advantage of deep depletion layers and specific guidelines for detection optimization. SPADs in CMOS technology also suffer from high noise levels. The purity of the starting material and the cleanliness of the fabrication processes are critical in single-photon avalanche detectors, in order to minimize noise contributions. With the scope of obtaining extremely low density of defects and impurities, dedicated technologies employ specific gettering steps, close to the device active area, that cannot be implemented, with the same effectiveness, in CMOS technologies.
However, CMOS SPADs enable the fabrication of monolithic detection systems, which enclose the sensing elements and the processing electronics in the same chip, thus reducing the interconnection parasitic capacitances and, consequently, the involved avalanche charge. As a result, the equivalent time constant, related to the SPAD output node, is smaller, making simple passive quenching circuits feasible, since the exponential recovery after quenching is fast enough. Moreover, the compatibility with the array fabrication featured by SPADs in CMOS technology is an essential condition for various applications. Highly dense image sensors based on CMOS SPADs can be extensively found in the literature [67][68][69][70][71][72], while achieving the same results in dedicated technologies is practically impossible.
The state of the art of the CMOS SPAD fabrication can be examined by differentiating between three types of manufacturing processes [4]: *CMOS compatible processes*, aiming at producing high performance detectors with poor capabilities of array integration, *High Voltage (HV) CMOS processes*, which offer the possibility of a relatively low doping deep n-well used to ensure up to 50 V isolation from the substrate, *Standard CMOS processes*, allowing a very high scale of integration, low cost and high yield by trading off performance in terms of DCR, PDE and design freedom.
Therefore, even though technologies specifically developed for SPAD fabrication offer higher flexibility, allowing the optimization of device performance, they often suffer from low yield, high cost, and limited compatibility with array integration. In contrast, CMOS technologies have become more attractive due to their ability to meet the demands of the current market, which is gradually moving towards the large-scale production of cost-effective, dense sensing arrays.

1.3.4 Silicon Photomultipliers (SiPMs)

Silicon Photomultipliers (SiPMs) are solid state detectors consisting of arrays of SPADs. Depending on the specific architecture, hundreds or thousands of

photodiodes are somehow connected together in a pixelated structure with the aim of providing photon counting with high resolution and single photon sensitivity [23]. Since SiPMs rely on Geiger mode devices, the swift internal avalanche amplification enables the acquisition of precise timing information, regarding the arrival time of detected photons, within tens of picoseconds. Practically, the information provided by the detection system represents the number of simultaneously fired SPADs. If an adequate spatial distribution of the light particles is provided, the output signal of a SiPM is proportional to the number of impinging photons. The sensing elements integrated within a photomultiplier structure typically have active areas with dimensions ranging from few micrometers to several tens of micrometers. The pixel design is carefully carried out to optimize dark noise, resolution, and fill factor, depending on the specific applications. The size of the active area and the overall number of SPADs are chosen in order to reasonably trade off between PDE and linear range [73]. In applications where large signals are expected, the limited number of SPADs in a SiPM can result in signal saturation. Early commercial SiPMs had SPAD diameters around few hundreds of micrometers. Recent advancements in the production technologies have allowed for much smaller SPADs, as small as 10 μm, greatly improving the linearity of the detection system. However, if the active area is too small, the dead space, intended as sensor-surrounding regions not contributing to photon detection, may not scale with the same factor, thus leading to a decrease in the PDE.
Silicon photomultipliers find extensive applications across numerous fields, playing a pivotal role in various light detection systems. Typical applications involve light detection and ranging (LIDAR), functional optical spectroscopy, fluorescent light detection in physics and biology, quantum physics, quantum computing, nuclear medical imaging, gamma spectroscopy, time tagging of high energy particles and oncological diagnostics (TOF-PET). Further information about practical applications of SiPMs can be found in the following of this chapter.
For the sake of comparison between various SiPM structures, a first differentiation can be found between analog and digital architectures. In the case of analog SiPMs, the SPADs are connected in parallel, as shown in Fig. 1.17. When a single SPAD is triggered, it generates a current output signal. If multiple SPADs are triggered simultaneously, the output pulses overlap, resulting in a combined signal with an amplitude that is proportional to the number of triggered SPADs. Therefore, the photon count information is obtained by integrating the charge over time, since the latter is directly proportional to the number of detected photons. This operation can be done with a charge

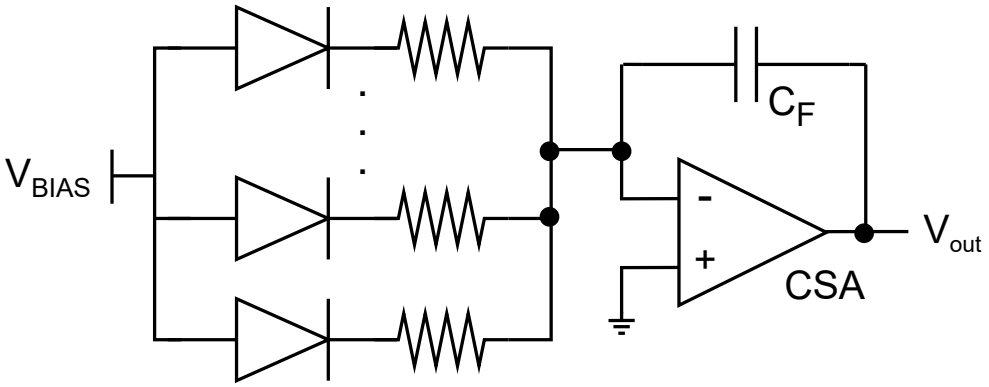


Figure 1.17: Schematic diagram of an analog SiPM.

sensitive amplifier (CSA). The analog architecture offers high-speed response, allowing for the detection of photons with significantly high temporal density. However, this approach brings about several drawbacks, mainly involving the high power consumption, associated with each analog circuit used in the readout channel. Indeed, even when no particle is detected, the analog circuits reading the SPAD output signal continue to consume power. Additionally, the design adaptability and scalability of the analog architecture is strongly limited, since each component of the readout chain needs to be specifically designed for the particular application. A custom design of the entire readout channel may result in higher development time and complexity, as well as reduced flexibility for different use cases or future upgrades. The SPAD output signal, after being treated by the analog circuits, is usually fed to an analog-to-digital converter (ADC), since most signal processing operations are performed in the digital domain. A great challenge to be faced during the design of an analog SiPM is represented by the parasitic capacitance of the not-triggered SPADs which are connected to the input terminals of the charge amplifier [74]. In the case of single photon detection, the non-triggered pixels introduce a non negligible parasitic capacitance (up to several hundreds of pF, depending on the SPAD dimension) at the input of the CSA, thus having a significant impact on the ability to effectively reveal the single photon.

1.3.4.1 Digital Silicon Photomultipliers (dSiPMs)

Despite analog SiPMs have represented the primary choice in applications demanding high frequency photon counting, most of the research effort has recently been focused on the design of digital silicon photomultipliers (dSiPMs),

that are deemed as the ultimate winner in the search for the ideal photomultiplier-based detection system [23] [75] [76]. In a dSiPM, each SPAD of the array is provided with a quenching network and a digital circuit implementing various logic functions. The output pulses generated by the fired pixels are processed by further levels of digital electronics, possibly located outside the SPAD cluster. Typical features that can be provided by dSiPMs concern pulse counting, threshold crossing and timestamp. Digital SiPMs offer high processing capabilities with reduced in-pixel readout channels, low static power consumption and different methods to keep DCR under control. These include the use of active circuits biasing the SPAD below the breakdown voltage for a determined time window or the possibility of selectively switching off SPADs with excessive noise levels. However a non-negligible delay in the readout response may be introduced, if compared to the analog counterpart. The propagation delay of cascaded logic gates and/or the discretized delay, intrinsically featured by clocked systems, may largely exceed the time response of an analog readout circuit made of a charge integrator. Anyway, since the nature of SPADs is inherently digital (the information provided by SPADs is of the boolean type), a fully digital circuit may represent the most suitable solution to collect and process data from the sensor.
Although analog and digital SiPMs adopt completely different reading approaches, single photon detection is the ultimate goal of both architectures. The detection performance deeply varies, depending on the design line which has been embraced. In analog SiPMs, every stage of the readout chain, including the ADC, contributes to electronic noise. The noise performance of the system can be significantly degraded by the parasitic capacitance loading the common node connecting all the SPADs together. In a good design, increasing the transconductance of the first transistor in the readout chain allows to reduce the noise impact on the detector performance [75]. In digital SiPMs, the electronic noise is not a significant problem since it is irrelevant to the ability of resolving single photons. The front-end electronics, integrated close to the sensor, acts as a discriminator enabling single photon resolution. Discrete voltage levels, differentiating between an avalanche event and the steady state condition, are generated as output of each frontend circuit, which has to deal with the sole capacitance provided by the relevant SPAD.
During the fabrication process, some SPADs can be subject to damage, resulting in a higher noise count rate. In a SPAD array, even a small number of faulty SPADs may lead to a significant overall DCR due to crosstalk effects, making the resolution of single photons highly challenging. In large detection systems, an analog SiPM may need to be definitely shut down, resulting in

the loss of a photosensitive area corresponding to the size of the SiPM. In the case of dSiPMs, noisy SPADs can be selectively deactivated. Through a SPAD masking operation, single-photon resolution is preserved, while reducing the total dynamic range only by the number of disabled SPADs [77]. Additionally, unlike the analog counterpart, digital SiPMs can take advantage of active quenching circuits, which limit the charge involved in the avalanche, so as reducing afterpulsing, and control the recharge time, allowing for better PDE uniformity [75].
Although the integration of SPADs and electronics in the same substrate represents a great advancement in the field of imaging sensors, the design constraints set by CMOS foundries represent a serious limitation to performance of state-of-the-art SPADs [14]. Additionally, monolithically integrated SiPMs have to trade-off between processing capabilities and photosensitive fill factor. Therefore, by following the concept of "More than Moore" [78], the 3D vertical integration of heterogeneous layers has been recently proposed to overcome the drawbacks associated to monolithic systems [79]. Dedicated layers for the sensing elements and the readout electronics are separately developed and eventually connected by means of bonding techniques. An FSI (Front-Side Illuminated) or BSI (Back-Side Illuminated) state-of-the-art SPAD array with a high fill-factor is implemented on the first tier, employing a dedicated SPAD process. Within the second tier, the electronic circuits, used to readout the SPAD array, can be integrated using a carefully selected CMOS process that aligns with the specific needs of the application. The 3D heterogeneous vertical integration strategy combines the strengths of both tiers, resulting in efficient photon detection and high processing capabilities. The price to pay is mainly represented by the speed of response, primarily limited by the connections between layers, the complexity of the system and, consequently, the high cost of production.
Following, some dSiPM implementations, or complete detection systems based on dSiPMs, will be discussed more in detail, with special focus on the circuit implementation. It is worth specifying that the following subsections do not claim to be a complete review of digital SiPMs, but a brief list of some architectures which have set the benchmark in the last decades.

The Digital Photon Counter (DPC) by Philips [80][81][76]. The DPC, produced by Philips Digital Photon Counting, represents one of the first examples of high density dSiPM (Fig. 1.18). Arranged inside a silicon die of $7.8 \times 7.2\ mm^2$, the system is composed by 4 sections (2×2), referred to as pixels, for a total of 25584 SPADs. Two Time to Digital Converters (TDCs)

and an acquisition controller are shared between the pixels. Each pixel consists of 3200 micro-cells, grouped in 4 blocks (subpixels), and a trigger network producing a signal upon the generation of avalanches taking place inside the subpixels. The main task of the trigger logic is to minimize the probability of generating trigger signals due to dark counts. Since, in a real event, the simultaneous detection of photons from multiple sub-pixels is very likely, the superposition of detection pulses is exploited by setting trigger thresholds, which enable a more efficient discrimination between true events and noise. Four threshold levels, as the number of subpixels, have been implemented by using a combinational circuit consisting of a reconfigurable tree of AND/OR gates. However, this approach assumes a uniform distribution of photons over the sensor area, since photons falling on sub-pixels that have already generated a trigger signal cannot produce a second one within relatively short time. A validation network is used to process the trigger signals and produce the "Event Valid" flag. If the validation threshold set by the user is not met, a rapid reset of the sensor is carried out to minimize the dead time. Each micro-cell ($118.8 \times 64\ \mu m^2$) is composed by a couple of SPADs and the relevant frontend electronics. Every single SPAD is equipped with an active quenching circuit, a 1-bit memory, to enable/disable the current SPAD, and various output data line drivers. The TDC network is used to produce pho-

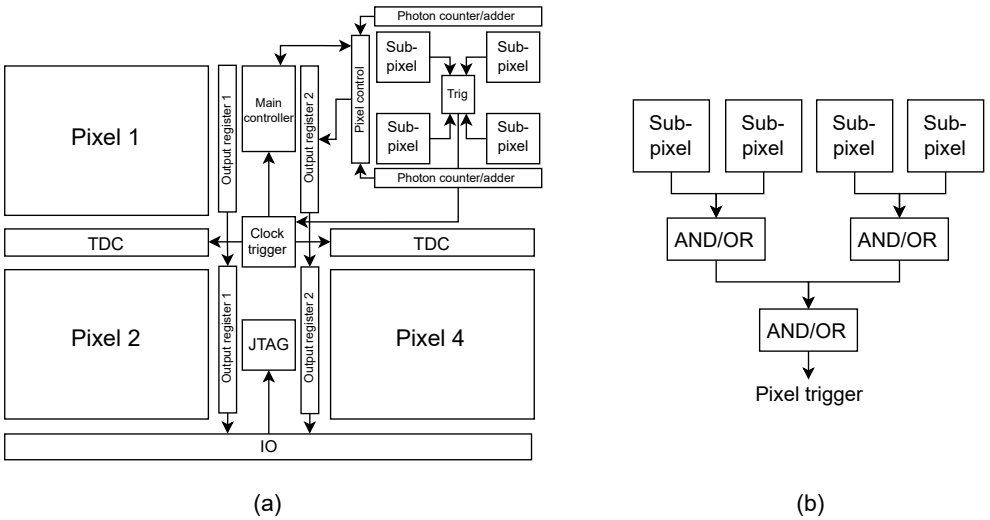


Figure 1.18: Schematic diagrams of the Philips Digital Photon Counter [76]: (a) chip floorplan, (b) combinational network implementing the threshold logic.

ton time stamps. The start signal is fed through the trigger signal coming from one of the subpixels, while the stop signal is generated from the TDC reference clock. Two TDCs are used to introduce redundancy in the system and produce at least one valid time stamp, since a basic problem of TDCs is metastability, which can occur if the stop signal arrives too close to the start one.

SPADnet [82][83][76]. SPADnet is a project aiming at the development of a networked, fully digital detection system in a standard 130 nm CMOS imaging process with TSV (Through-Silicon Via) technology. The structure is designed for the detection of gamma events and coincidence sorting in multi-modal time-of-flight PET. The detection system relies on the concept of deferred coincidence detection, where individual photonic modules are arranged in a single-ring or multi-ring configuration to detect gamma photons, generate the timestamp and the energy estimates for each event, and transmit data into the network for further processing and analysis. The sensor chip, located inside each photonic module, consists of an 8×16 array. Each of the 128 pixels contains 4 mini-SiPMs and two TDCs with 12-bits. A mini-SiPM consists of 180 SPADs, for a total count of 720 SPADs per pixel (with a pixel area occupation of $610.5 \times 571.2\ \mu m^2$). The single pixel is able to provide a timestamp and an energy measurement per event, with a rate of 100 Msamples/s. Within each mini-SiPM, spatial and time compression techniques are used to shrink the amount of information which is delivered to the higher levels of logic. For this purpose, SPADs are organized into groups of three using OR logic gates. Further compression is achieved by means of monostable multivibrators enabling additional output combination. The output signals coming from the four mini-SiPMs are sent to the two in-pixel TDCs which work in anti-parallel mode, in order to provide a continuity in the conversion. A continuous detection module is formed by assembling 5×5 sensor chips together. At each node in the network, gamma-ray timestamps and energy measurements are injected. A snooper is tasked with identifying coincidence pairs within the network, storing them and releasing them upon request.

The multi-channel digital SiPM (MD-SiPM) [84][85][86][76]. In the framework of the EndoTOFPET-US project, a detection probe has been designed for endoscopy and positron emission tomography. The detection system is based on a chip consisting of a 4×4 array of multi-channel digital SiPMs. A simplified block diagram of the chip floorplan is shown in Fig. 1.19. Inside the chip, a bank of 192 TDCs is positioned beneath the MD-SiPM core

array, enabling the time-stamping of up to 192 individual photons. For each MD-SiPM a maximum of 48 photons per event is fixed. The chip includes also a masking mechanism, to deactivate SPADs with excessive noise levels, a smart reset function, to mitigate the impact of dark noise pulses, and a high-voltage generator, to ensure a consistent excess bias voltage, featuring very small dependence on process, supply voltage and temperature (PVT) variations. After being stored in the in-pixel memory, the detection pulse is delivered to a column-parallel TDC, to extract the relevant photon time-stamp information. The time-to-digital conversion is achieved through an interpolated procedure, consisting of two steps. A 12-bit counter is used to perform a coarse conversion, while a VCO (Voltage Controlled Oscillator) provides a fine resolution by generating four different phases. When a candidate gamma event is detected, its energy is stored in the energy register, which contains the result of all TDCs. The number of TDCs that have fired indicates the number of microcells that have been triggered in the MD-SiPMs array. If a predefined threshold is not reached within a time frame, all the SPAD readout circuits and TDCs are reset in preparation for the detection of the next gamma event.

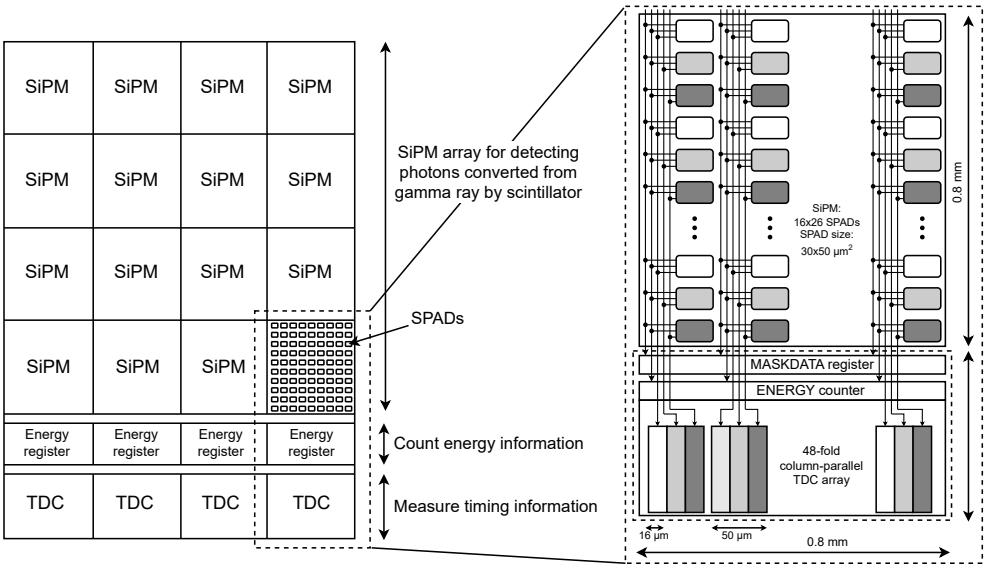


Figure 1.19: Schematic diagram of the chip developed for the endoscopic sensor used in the EndoTOFPET-US project [76].

dSiPMs with full-frame readout [87][88][76]. A two dimensional single-photon-sensitive detector has been developed in a 350 nm CMOS technology. The sensor chip, made of 88×88 square SPADs (with an area of 56.44 × 56.44 μm^2 per photodiode), is operated by means of a full-frame readout, which allows a readout speed up to 400k frames per second. A self-triggering logic, based on a selectable multiplicity of up to $\geq$ 4 hits, has been implemented. The time window used to manage the hit coincidence mechanism can be changed between preset values, by means of configuration bits. Indeed, such a SiPM enables the extraction of valuable information from single scintillation events. The enhanced capability proves especially beneficial when reading out high-resolution scintillation detectors that rely on complex light-sharing approaches or to investigate light propagation mechanisms in various crystal structures.

3D Photon to Digital Converter (3D-PDC) [75][89][90][91][79]. The dSiPM developed by the Sherbrooke group is one of the most representative example of 3D integrated structure. The digital detector, which is referred to as PDC (Photon to Digital Converter), is aimed at increasing the contrast of images in clinical and pre-clinical PET. This objective can be attained by minimizing the pixel-to-pixel skew, which is mainly originated due to the mismatching lines connecting the pixels and the system TDC. For this purpose, a two layered dSiPM has been developed. The sensing layer has been fabricated in a custom TDSI (Teledyne Dalsa Semiconductor Inc.) technology, while the readout layer consists of an ASIC produced in TSMC CMOS 65 nm technology. The readout array, implemented in the ASIC die, is made of 256 pixels (16×16). Each pixel, having an area of 65 × 65 μm^2, consists of the bonding pad connecting the front end network to the corresponding SPAD located in the sensing die, a quenching circuit, a TDC and a local counter. The 3D approach turned to be critical to allow the integration of a full TDC inside each pixel of the array, since the fabrication of SPADs in a dedicated layer, well separated from the one performing the readout, leaves more room to the design of complex processing circuits. Through the implementation of a digital SiPM with a one-to-one coupling between SPADs and TDCs, the pixel-to-pixel skew can be corrected, effectively reducing timing jitter in the system. Incorporating advanced features within the single pixel of a 2D digital SiPM may decrease the array FF, posing several limitations on the PDE. While the degradation of the optical performance might be affordable for applications like 3D imaging, it may turn to be less suitable for ToF-PET, where the detection efficiency is of paramount importance. The TDC developed for this

application relies on a ring oscillator-based Vernier architecture. An in-pixel counter, working in parallel with the TDC, is used to measure the energy at the output of the relevant quenching circuit. For all the events, each pixel provides the position information, the timestamp of the first photon and a count result describing how many times the SPAD was triggered within a determined integration time. In order to discriminate noise-generated events, a threshold logic based on the number of triggered columns has been implemented at the array level.

1.3.5 Main fields of application

In the following, some of the most important applications exploiting the SPAD detection capabilities will be briefly explored. The high number of fields, to which SPADs can give a beneficial contribution, highlights their impact on advancing technology and research.

- **Biomedical Imaging: PET [92][93][94][77][95][96][97].** Positron Emission Tomography (PET) is a nuclear imaging method that relies upon annihilation of gamma photons, generated after positron decay, to produce three-dimensional functional images of the human body. The primary applications encompass clinical oncology, preclinical research, and brain function analyses. Notably distinct from other body imaging techniques like Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), PET provides metabolic insights into the patient's body. To retrieve this information, emissions from radioactive compounds, also known as tracers, are leveraged to identify tissues where specific cell functions, such as the elevated glucose metabolism typical of cancer cells, occur. Fig. 1.20a shows a schematic representation of a PET system. A radioactive tracer is injected into the patient. When a radioactive atom of the tracer decays, it emits a positron, which travels a short distance (slightly smaller that 1 mm) before undergoing annihilation. During the annihilation process, the positron combines with an electron, resulting in the emission of a pair of 511-keV gamma photons in opposite directions. The PET scanner detects both the emitted photons to determine the line of response (LOR), which contains the annihilation location. By collecting millions of LORs, PET forms a detailed 3D tomographic image of the subject, providing valuable information about the distribution and concentration of the tracer in various tissues and organs. As shown in the figure, PET scanners are designed as a ring of detectors to facilitate the identification of photon pairs. Each detector in the ring is responsible for

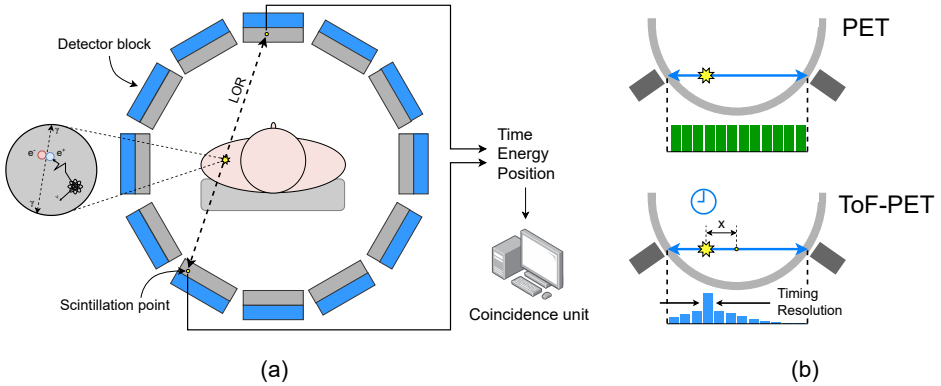


Figure 1.20: positron emission tomography: (a) ring scanner, (b) main difference between original PET and ToF-PET [92][97].

measuring the position, energy and time of arrival (ToA) of the gamma photons. The full set of information is processed by a coincidence unit, assessing whether the detected photons originate from a unique annihilation event. The detector unit, used for this application, is made of a scintillator, absorbing gamma photons and emitting light, coupled with a photosensor. For years, SPADs, arranged in SiPM structures, have been the sensors of choice for PET scanners, replacing PMTs, which suffer from bulkiness, fragility, high power consumption and high sensitivity to magnetic fields. In the case of analog SiPMs, a current signal proportional to the number of fired SPADs builds up as a response to a scintillation event. Recently, dSiPMs are gaining significant relevance in PET applications, mainly due to their remarkable performance in terms of electronic noise and processing capabilities. In a dSiPM, the timing information is usually provided by a cluster of TDCs, while a system of memories can be used to disable noisy SPADs, thus increasing performance and device reliability. Over the past twenty years, the remarkable progress experienced in the field of detection systems has significantly enhanced the quality of medical imaging. Among the latest advancements, the introduction of time-of-flight (ToF) PET technology stands out as a major breakthrough. In this case, using dedicated TDCs, the difference in the arrival times of the gamma photons is measured at the two detectors. The additional information is used to precisely pinpoint the location of the source (Fig. 1.20b), thus substantially improving the SNR (Signal-to-Noise-Ratio) and the contrast recovery of the acquired image.

- **3D imaging and LiDAR technology [98][99][100][101][102].** The demand for enhanced autonomy in modern devices, appliances, robots, and transportation means necessitates the integration of three-dimensional imaging as a fundamental technology. Through suitable sensors, autonomous systems can develop a comprehensive awareness of the surrounding environment, facilitating sophisticated decision-making processes. While various ranging techniques are available, optical ranging has emerged as the best choice in numerous fields, due to its remarkable depth and lateral resolution. LiDAR (Light Detection and Ranging) systems, can perform highly precise ranging measurements, outperforming non-optical methods such as ultrasonic and radar. 3D imaging techniques, relying on SPAD sensors, are currently used in smartphones, autonomous smart vehicles (hence contributing to the Advanced Driver Assistance Systems - ADAS), robots, security systems (to perform 3D face recognition) and also in navigation and landing devices for spacecrafts. LiDAR imagers are primarily made up of three key components: a light emission system employing a pulsed laser, an image sensor embedding SPAD arrays, and a control mainframe. In order to build the 3D image of an object and receive information about its relative position with respect to the detection system, the time-of-flight of the signal emitted from the pulsed laser, until it returns to the image sensor after being reflected by the object, is measured (Fig. 1.21). The accuracy, range, and frame of the LiDAR system are directly impacted by factors such as the pulse width, the peak power, and the frequency of the laser used in the emission process. The diverse applications of 3D imaging systems

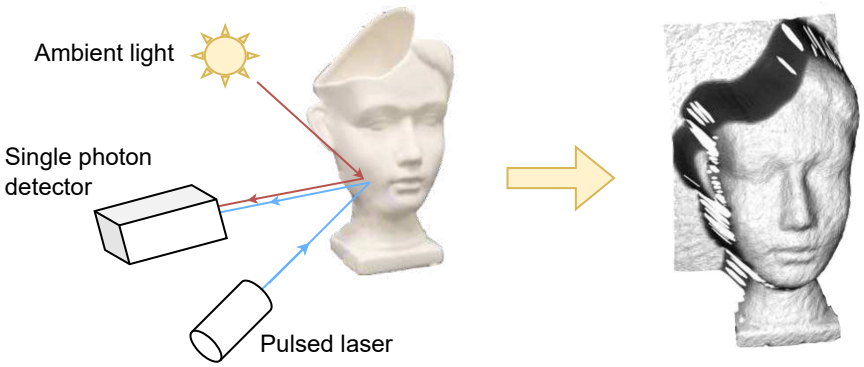


Figure 1.21: Simplified diagram of a ToF acquisition system [101].

result in varying requirements, involving factors like speed, resolution, cost, probing wavelength, optical ranging method, power consumption, and size.

- **Fluorescence Lifetime Imaging Microscopy (FLIM) [103][104][105] [106][107][108][109][110].** Fluorescence lifetime imaging microscopy (FLIM) is an imaging technique used to measure the lifetimes of fluorophores with microscopic spatial resolution. It serves as a valuable tool for cell biologists, enabling the detection, visualization, and investigation of the structure of biological systems at the subcellular level. Fluorescent dyes are used as markers to allow the visualization of complex molecular assemblies within single voxels (volumetric picture elements). FLIM can find practical applications in fields like blood flow imaging, study on drug uptake and investigation on neuronal activity. Despite conventional fluorescence microscopy, that performs spectrally-discriminated measurements of intensity, FLIM is based on the measurement of the rate of decay of a fluorophore, thus resulting in better image reconstruction. Exploiting time-correlated single photon counting (TCSPC), the arrival time of photons is compared with the time stamp of an excitation pulse, generally a pulsed laser, which is used to stimulate the fluorophore. The relative time measurements are repeated to extract the information about the intensity decay profile. PMTs have been for years the detectors of choice in FLIM systems. Nonetheless, due to the lack of an electrical gating mechanism, more effective than an optical one in suppressing background signals correlated with the excitation pulse, and the well known drawbacks brought about by photomultiplier tubes (bulkiness, fragility, sensitivity to magnetic fields, high power consumption), SPADs have been largely preferred to other detectors. The integration of SPADs in CMOS technology has created an innovative platform for FLIM imaging, which can take advantage of very dense arrays of sensors, offering significant space and time resolution.

- **Quantum Technology [111][112][113][114][115][116].** Quantum technologies leverage the SPAD ability to detect individual photons with high temporal resolution, in order to extract quantum information. The concept of harnessing quantum correlation to overcome classical limitations in data acquisition has sparked extensive research and development. Groundbreaking improvements in quantum technologies opened new possibilities in various fields of science. Despite not being the only detectors used for quantum purposes, SPADs significantly contributed

to advancements in the state-of-the-art quantum systems. If compared to usual detectors for quantum applications (superconducting nanowire single photon detectors, PMTs, transition-edge sensors), SPADs represent a reliable and compact alternative at room temperature, with remarkable timing performance and no dependence on the polarization of light. In quantum imaging, the gating time, which represents the effective sensitivity window allowing the acquisition of a single frame, is one of the major factor to consider. If the gating time exceeds the source coherence time, intensity fluctuations within each frame can be partially erased, making the reconstruction of correlation more challenging. SPAD arrays offer a unique combination of fast acquisition rates, short gating times, and low noise levels, thus being the ideal detectors for high-quality sampling of light statistics. In quantum cryptographic applications, SPADs are used to produce random number generators (RNGs) with high bit rate. Different quantum procedures in nature are used to generate non-deterministic signals, such as entangled state measurement, wave function collapse of single photon due to measurement, effects of vacuum fluctuation and quantum phase fluctuation. Generally, a quantum RNG uses an external light source coupled to a sensor with single photon capabilities. Recently, monolithic systems exploiting SPADs, both as intrinsic photon generators and sensing elements, have been developed, paving the way for highly compact devices integrated into miniaturized systems and smart packages. Additionally, SPAD capabilities are being explored for qubit (quantum bit) readout in the field of quantum computing, for the development of Quantum Key Distribution (QKD) protocols in quantum communication, and in view of quantum photonic experiments, to study quantum entanglement, quantum interference, and other quantum phenomena.

- **Charged particle physics [117][5][36][42][118][119][120][6].** Beyond photon specific applications, SPADs have been also investigated in view of particle radiation detection. In High Energy Physics (HEP) experiments at next-generation particle colliders and B-factories, SPADs can represent a valuable alternative to hybrid pixel technology, which is already used in Large Hadron Collider (LHC) experiments. Although designated for recent upgrades of the ATLAS and CMS trackers, hybrid pixels might not represent the optimal choice for tracking applications, due to their relatively large amount of material. The need for low power, highly granular and light detectors has fostered the design and characterization of SPAD sensors featuring reduced material budget, so as to

minimally interfere with the particle path. Since the overall detection mechanism takes place entirely inside the SPAD active volume, which is represented by the depleted region, the chip die can be thinned down to few microns, without undermining the sensor functionality. However, extremely thinned chips may result in mechanically fragile detection systems and modified bulk electrical properties. SPAD arrays provide significant advantages in terms of timing resolution, power consumption, and immunity to electromagnetic interference. In the field of nuclear medicine, SPAD performance is currently evaluated for applications to radioguided surgery and Particle Therapy (PT) treatments. In the first case, SPADs can be used to assist the medical staff during surgery by providing the exact location of a specific tissue which has been previously marked through beta emitters. In the case of PT treatments, which is a technique aiming at destroying tumor cells by using proton and carbon beams, SPADs represent a good candidate to carefully monitor the emission of secondary particles generated during the treatment. Indeed, an efficient detection mechanism is required in this case, since secondary particles may represent a non negligible risk for healthy tissues.

1.4 The APIX2/ASAP project

This thesis work, concerning the design and characterization of SPAD arrays in 110 nm CMOS technology, has been carried out in the framework of the APIX2/ASAP project, funded by the National (Italian) Institute for Nuclear Physics (INFN). This project brings together individuals from University of Pavia, INFN Pavia, University of Trento, the Trento Institute for Fundamental Physics and Applications (TIFPA), University of Siena, University of Padova, INFN Padova, and INFN Pisa. The primary objective is the development of next-generation stacked avalanche detectors to be used for charged particle detection. Special focus is placed on enhanced efficiency, minimization of noise, and low material budget. The latter is achieved by assuming that the substrate of the detector can be thinned down to a few microns, corresponding to the effective device active volume, without compromising the functionality of the sensors. Within the project, three test chips have been produced so far: APIXFAB0 [36][34], APIX prototype-I [121][122], and APIX2LF [5][34]. The initial chip, APIXFAB0, was a single-layer device manufactured using a 180 nm CMOS technology with HV (high voltage) option. The main purpose was to investigate the technology capabilities for SPAD device development. The APIX prototype-I chip, fabricated using a 150 nm CMOS standard tech-

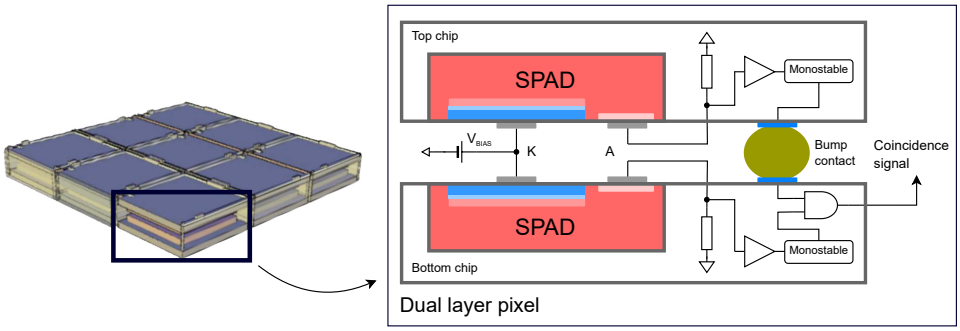


Figure 1.22: coincidence-based dual layer pixel in an array structure.

nology, was designed as the first-ever dual-layer SPAD array. The aim of the APIX2LF chip, produced in the same technology as the previous one, was the demonstration of the beneficial impact, as far as DCR is concerned, brought about by a dual layer structure in moderately large arrays of pixels. In addition, different architectures of readout circuits were implemented within the same chip, in order to study different ways of treating the SPAD output signal. In the context of this PhD thesis work, a new test chip, namely ASAP110LF, has been designed using a 110 nm CMOS technology. The scope of the newly fabricated test chip is to carry out a technology characterization specifically focused on SPAD performance. Moreover, novel structures, involving a new SiPM architecture and experimental SPAD junctions, have been integrated within the chip for test purposes. A description of the ASAP110LF chip will be provided in Chapter 2.

1.4.1 Description of the coincidence-based dual layer SPAD structure

The core concept of the APIX2/ASAP project relies on the coincidence-based layered approach. The dual layer position-sensitive detectors, developed in the framework of this project, leverage a vertical alignment configuration, consisting of two layers of CMOS SPADs. In the 3D structure, the sensing elements are arranged in pairs, positioned face to face, to form a basic cell, as shown in Fig. 1.22. When a charged particle passes through the two overlapping SPADs, a coincidence signal is generated. Since the two chip layers are exposed to radiation from the backside, the PDP of the final structure, strictly intended as the detection efficiency of the sensing system in response to upcoming photons, is virtually zero. However, if the sensor active volume is sufficiently thick, the two SPADs combined in coincidence are highly effi-

cient in revealing charged particles. The latter, travelling through the SPAD layer stack, produce clusters of electron-hole pairs while interacting with silicon, thus increasing the probability of triggering self-sustaining avalanches. In particular, for minimum ionizing particles (MIPs), clusters of carriers are generated in the sensor material with an average distance of about 0.21 μm [123].
The two SPAD layers are manufactured independently on separate chips and then interconnected using micro bump bonding techniques, that guarantee a yield close to 100% at a pitch larger than or comparable to 50 μm [122]. The effective distance between the active volumes of two overlapping SPADs depends on the vertical dimension of the bump pad, not exceeding 10 μm. Since the two sensors making a single detection cell are physically located on different chips, a mismatch on the breakdown voltage may occur. As a result, considering that the two SPAD layers are biased with the same SPAD voltage, a different V_{EX} may be applied to the two sensors. For applications involving particle detection, the V_{BD} mismatch does not represent a problem, since the detection efficiency at sufficiently high excess voltage is practically 1. However, detection pulses and DCR events can appear at the output of the bi-layered cell only if the applied bias voltage exceeds the higher between the breakdown values of the two cells. The resulting threshold voltage can in this case be considered as the equivalent V_{BD} of the dual layer pixel. Each individual SPAD element is equipped with dedicated readout electronics and quenching circuit. The coincidence signal, generated when two overlapping SPADs are simultaneously triggered, is achieved through an AND gate located in the bottom chip. The SPAD properties and readout electronics in both layers are identical, except for the AND gate producing the coincidence signal. The floorplan of the two layers is carried out in order to obtain two versions of the same chip. The layout views of the two tiers are mirrored each other, ensuring matching

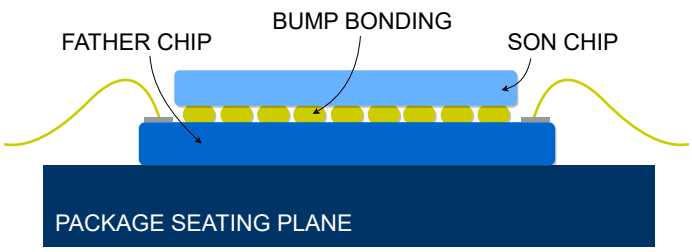


Figure 1.23: Section view of the dual layer detection system developed in the APIX2/ASAP project.

between SPADs intended to be in a coincidence configuration. With reference to Fig. 1.23, where a lateral view of the three-dimensional structure is shown, the bottom chip, which contains the AND gate, is referred to as the father chip while the top one, bonded onto the father SPAD layer, is referred to as the son chip. The latter has no bonding wires towards the package, since all the control signals are provided to the second layer, through the father die, by using the bump bonding connections. Therefore, the father chip is the only layer which is directly connected to the outside through the bonding wires.
In a dual layer structure consisting of overlapping SPADs, vertical crosstalk may happen between sensors belonging to different chip layers. In order to reduce this effect, the active area of the sensors can be covered with metal layers, preventing electro-luminescence photons from being absorbed in the active region of the other layer.
Commercially available CMOS technologies have been selected to minimize system complexity, enhance chip durability, and achieve a high yield, resulting in cost reduction. Furthermore, CMOS technology is well-suited for applications requiring the fabrication of pixelated detectors.
If compared to hybrid pixels or monolithic solutions consisting of relatively thick sensors (from few tens to few hundreds of microns), the use of a dual layer detection system based on overlapping SPADs represents a significant gain in terms of material budget, as the thickness of the entire structure, after the vertical connection of two chip layers and the removal of the passive substrate, may, in principle, be limited to less than 10 μm.

1.4.2 Dark noise mitigation in a coincidence-based structure

The dual-layer SPAD structure offers significant advantages in terms of noise performance. The coincidence-based approach provides the detector with the ability of differentiating between a signal generated by a radiation source and a noise pulse. The fundamental idea behind the dual-layer sensor is to leverage the extremely low probability of two simultaneous dark count events occurring in two overlapping pixels, thereby greatly reducing the impact of DCR on the detection process, compared to a single-layer detector. The AND gate implemented in the bottom chip allows only time-overlapped pulses, generated in both SPADs of a dual layer pixel, to be delivered to the output. In a pixel consisting of two overlapping SPADs, the DCR can be expressed as [121]

$$DCR_C = 2 \times \delta t \times DCR_T \times DCR_B, \tag{1.31}$$

where DCR_C, DCR_T and DCR_B are the dark count rate featured respectively by the dual layer pixel, the top layer SPAD and the bottom one, while δt

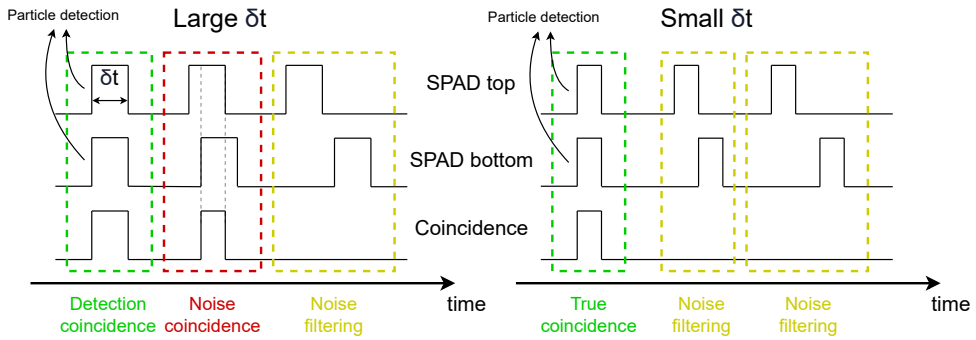


Figure 1.24: time diagram showing the effectiveness of small values of δt in filtering out noise pulses.

represents the coincidence time window, typically in the order of few nanoseconds. As depicted in the time diagrams of Fig. 1.24, when a particle passes simultaneously through both aligned detectors, a valid event is detected by the electronics, regardless of the coincidence window duration. To minimize the occurrence of random coincidence signals caused by dark count events, small values of δt should be used. If a large coincidence window is selected, dark pulses generated in the two SPADs may partially overlap, resulting in a noise-generated coincidence signal. Hence, smaller values of δt are preferred to reduce the overlapping probability, and, consequently, the impact of DCR. However, the time window duration may have a lower limit determined by the time jitter of the detector response and by the timing performance featured by the signal generated in the top chip, which could be affected by the parasitic capacitance introduced by the bump connection.
Conceptually, if more than two SPADs are read out in coincidence, (1.31) can be extended to the more general case

$$DCR_N = N \times \delta t^{N-1} \times \prod_{k=1}^{N} DCR_k, \tag{1.32}$$

where N is the number of sensors connected in coincidence. As apparent from (1.32), the noise performance of a detection system based on layered sensors is strongly enhanced when multiple SPADs are involved in the coincidence. Assuming $N = 3$, $\delta t = 10^{-9}$ s and a DCR of 2 kHz for a single SPAD, a coincidence DCR of 2.4×10^{-8} Hz is obtained.

Chapter 2

The ASAP chip

This chapter will discuss the design of the ASAP110LF chip, developed in the framework of the ASAP project. The main purpose of the chip is to provide a test platform supporting the characterization of CMOS SPADs fabricated in a 110 nm CIS (CMOS Image Sensor) technology. Almost all the circuits included in the chip were designed for the sake of the SPAD characterization. The core of the chip is made of SPAD arrays and periphery circuits, which can be accessed through suitable digital stimuli. A time to digital converter was included for inter-avalanche time interval measurements. Part of the chip is devoted to an array of digital SiPMs, whose readout relies on a novel architecture based on parallel counters.

2.1 Chip overview

The ASAP chip was designed to investigate the performance of SPADs fabricated in a 110 nm CIS technology. As compared to standard CMOS technologies, particular technologies targeting the production of image sensors apply specific modifications to the manufacturing processes, so as to mimimize the density of defects in the device sensitive region. In particular, the use of non-silicided, double-diffused source/drain implantation, hydrogen annealing and the employment of p-epitaxial substrates allow to mitigate the trap-related dark count and reduce the field enhancement effects [40]. However, the DCR of SPADs fabricated in CIS technologies may escalate with the scaling of the process node, due to the increase of the doping concentration in the space charge region of the device [47].
The structures that were integrated within the ASAP110LF chip are aimed at supporting the characterization of SPADs manufactured by using a 110 nm

CIS technology. Arrays with relatively large dimensions were purposely designed, in order to allow the collection of statistically meaningful sets of data. In addition, the matrix arrangement makes it possible to study the performance of SPADs integrated in a complete detection system. Various readout approaches and multiple analysis tools were included within the chip, thus offering a wide range of measurement options. Dark count rate, crosstalk between pixels, the nature of the dark count events taking place in the active region of the sensors, time resolution of SPADs and radiation hardening, are some of the main research topics that can be explored with the structures integrated within the ASAP110LF chip. The almost fully digital nature of the circuits contained inside the chip brings about several advantages in terms of automation of the readout procedures and immunity against electronic noise. Beside supporting the characterization of sensors in the aforementioned technology, the ASAP110LF chip also integrates the first prototype of a digital SiPM that was designed by using a novel architecture based on parallel counters [124]. The new structure, developed in different variants, was arranged in

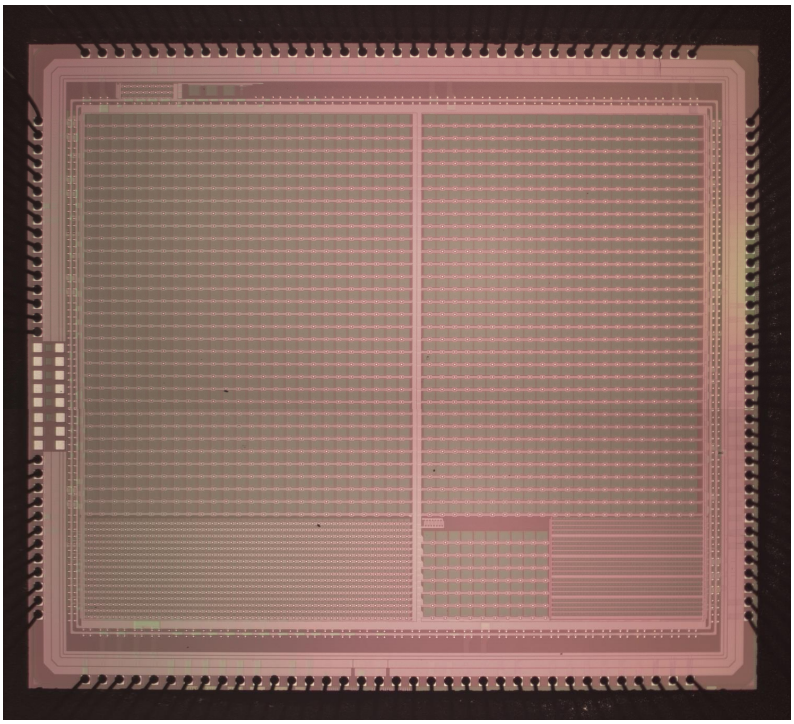


Figure 2.1: Microscope photograph of the ASAP110LF chip.

an array accommodating 60 SiPM samples. A description of the SiPM will be provided in the second part of this chapter, while the results from a preliminary characterization will be shown in chapter 3.
A microscope photograph of the ASAP110LF chip is shown in Fig. 2.1. The design, from the schematic to the layout phase, was carried out in the Cadence Virtuoso environment. The chip, with a die area of $5.1 \times 5.8\ mm^2$, was designed by using four metal layers. The layout design was carried out without using any automatic place and route tools, so as to optimize the delay performance of specific critical paths. The nominal core supply voltage, as well as the nominal voltage used for the IO structures, is $1.2\ V$. A reference voltage of $3.3\ V$ is used in dedicated circuits consisting of thick oxide transistors. The padring, running all around the chip core, is made of 136 pads, distributed between supplies, analog and digital pads. Pass through pads were selected for the input digital signals, in place of digital ones, in order to allow a proper signal distribution in the case of a vertical connection between two chip layers, according to the APIX2/ASAP approach [5]. If digital pads were used on both layers, the common bus transmitting stimuli from the father chip to the son one may find itself in a multi-driven condition, thus resulting in the input digital signals being set to a non-deterministic logic value. The SPAD cathode voltages are provided through five pads (VKi, with i= $1, 2, ..., 5$), which were deprived of the clamping diodes, so as to enable the application of SPAD bias voltages well above the IO supply voltage.
The sensing structures integrated in the chip can be distinguished between single sensors located in the chip periphery and core circuits. Most of the die surface is dedicated to the SPAD arrays implemented in the chip core, which contains a significantly higher number of SPADs as compared to the single device section. In the following section, all the structures in the chip, as well as all the operating procedures for chip characterization, will be discussed.

2.2 Single SPADs and linear APDs

For test purposes, some APDs, with different junction characteristics, were included in a rectangular area, located in the left side of the padring. Each sensor is provided with two custom pads, representing the anode and the cathode terminals of the structure. The I-V characteristics, as well as the main electrical parameters featured by the sensors, can be extracted through the use of microprobes, since no bonding wires are connected to the anode/cathode pads. Six individual APDs, each implementing a different type of structure, are included in this area:

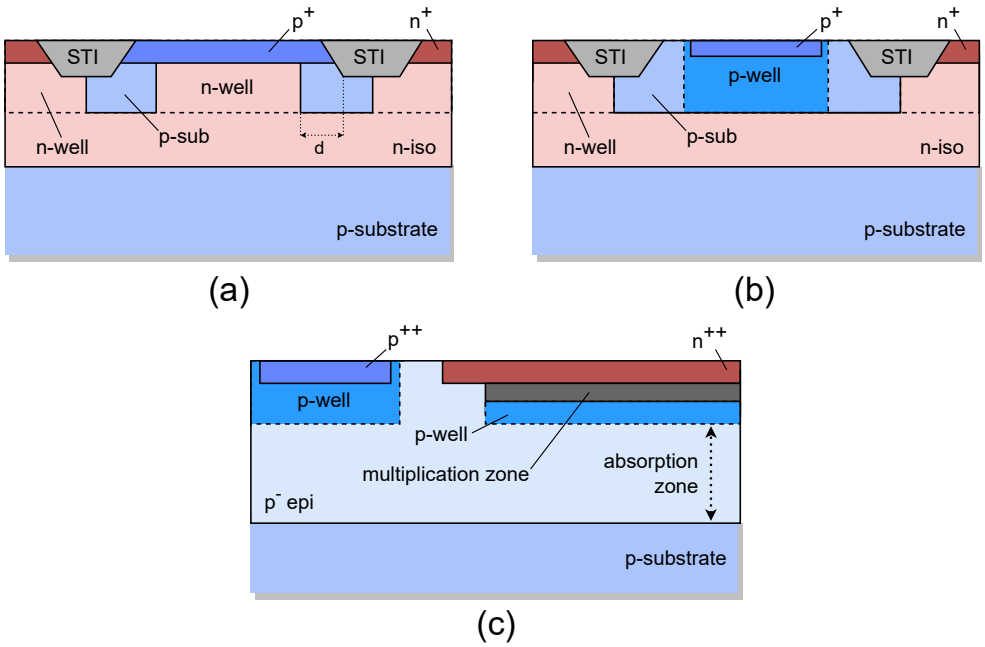


Figure 2.2: Cross-section of the different SPAD structures included in the single-sensors area (left side of the padring).

- SPAD 1 (active area = $50 \times 50\ \mu m^2$). A cross section of the sensor structure is shown in Fig. 2.2a. This sensor is based on a p+/n-well junction [42][125][126], with an active volume less than 2 μm thick. The isolation from the substrate is obtained by means of a deep nwell, which allows the chip die to be thinned down to a few micron without undermining the sensor functionality. The guard ring is achieved by blocking the n-well, at the border of the junction, through a low doped ring sorrounding the active area, in order to avoid premature edge breakdown. Shallow trench isolations (STI) were implemented to reduce electrical crosstalk effects between adjacent pixels. The distance (d), under a Non-Disclosing Agreement, between the edge of the n-well and the inner side of the STI is a key parameter affecting the noise performance of the junction, since the STI fabrication process may increase the density of deep-level carrier generation centers at the trench interface [127].

- SPAD 2 (active area = $50 \times 50\ \mu m^2$). The structure used for SPAD 1 was replicated in this sensor, with the exception of the STI layer, which was omitted in order to study the effect of shallow trench isolations on

the noise performance of the sensor.

- SPAD 3 (active area = 48×48 μm^2). The structure employed in this sensor closely resembles the one found in SPAD 1, with the sole variation being the increment of distance d by 1 μm ($d_{SPAD3} = d_{SPAD1} + 1$ μm).
- SPAD 4 (active area = 46×46 μm^2). The structure employed here closely resembles the one found in SPAD 1, with the sole variation being the increment of distance d by 2 μm ($d_{SPAD4} = d_{SPAD1} + 2$ μm).
- SPAD 5 (active area = 50×50 μm^2). The cross section of this sensor is shown in Fig. 2.2b. The structure relies on a p-well/deep n-well junction, which results in an active volume located further beneath the sensor surface, as compared to SPAD 1.
- Linear APD (active area = 49×49 μm^2). The cross section of the linear APD is shown in Fig. 2.2c. The sensor is based on a n+/p-well/p$^-$epi structure. The main difference between a p+/n-well junction and this sensor lies in the thick absorption layer, represented by the p-epi layer, and in the multiplication zone located at the n++/p-well junction. This separation allows a substantial increase of the drift region thickness, where photons are absorbed [128].

All the sensors included in the core arrays are based on the SPAD 1 structure, except for sensors in the ASSTI array (see section 2.3) that implement the same junction exploited in SPAD 2.

2.3 SPAD arrays and core circuits

A schematic diagram of the chip core is shown in Fig. 2.3. The core structures, involving both SPAD arrays and processing circuits, are arranged as follows (a detailed description of all the structures will be provided later on in this chapter):

- Array 1 (A1): 32×26 pixels, single SPAD per pixel, pitch of 100 μm, SPAD active area of 80×84 μm^2, fill factor (FF) of 67%, front-end circuits with passive quenching, in-pixel readout electronics with 1-bit memory, triggered-read option and availability of a non-latched output (monostable output).
- Array 2 (A2): 32×22 pixels, single SPAD per pixel, pitch of 100 μm, SPAD active area of 90×72 μm^2, FF of 64%, front-end circuits with passive quenching, in-pixel readout electronics with a 10-bit counter.

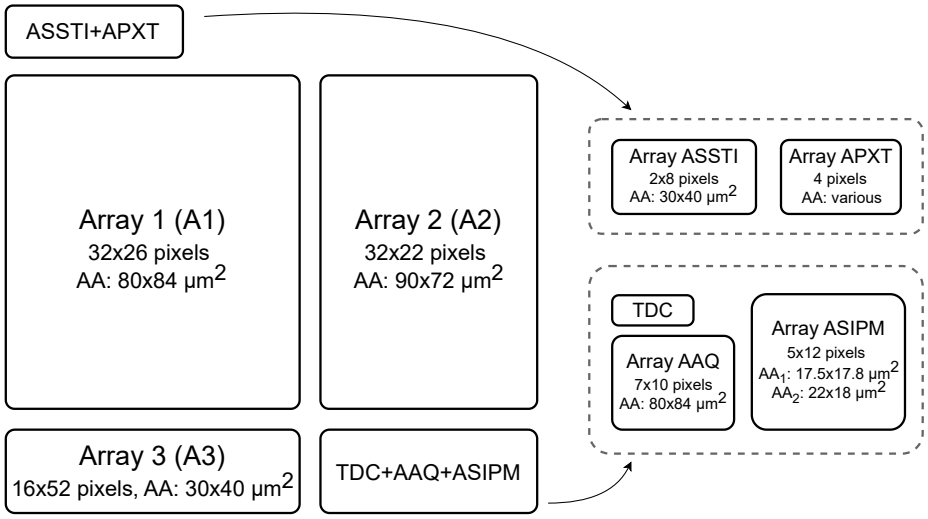


Figure 2.3: Schematic diagram of the chip core.

- Array 3 (A3): 16×52 pixels, single SPAD per pixel, pitch of 50 μm, SPAD active area of 30×40 μm^2, FF of 48%, front-end circuits with passive quenching, in-pixel readout electronics with 1-bit memory, triggered-read option and availability of a non-latched output (monostable output).

- Array ACT-Q (AAQ): 7×10 pixels, single SPAD per pixel, pitch of 100 μm, SPAD active area of 80×84 μm^2, FF of 67%, front-end circuits with active quenching, in-pixel readout electronics with 1-bit memory and availability of a non-latched output (monostable output).

- Array SiPM (ASIPM): 5×12 pixels, 16 SPADs per pixel (4 × 4) arranged in a SiPM sensor, vertical pitch of 162 μm and horizontal pitch of 100 μm, single SPAD active area of 17.5×17.8 μm^2, $FF_{subpixel}$ (considering the single SPAD and the relevant front-end electronics) of 34% and FF_{pixel} (i.e., SiPM FF) of 30%, front-end circuit (at the subpixel level) with passive quenching, in-pixel digital readout with a 16-to-5 parallel counter, a 5-bit memory and a Signal Over Threshold (SOT) logic.

- Array SEPARATED-STI (ASSTI): 2×8 pixels, single SPAD per pixel (sensors with the structure indicated as SPAD 2 in section 2.2), pitch of

50 μm, SPAD active area of 30×40 μm^2, FF of 48%, front-end circuits with passive quenching, in-pixel readout electronics with a 1-bit memory.

- Array PIXEL-TEST (APXT): 4 spaced pixels, single SPAD per pixel, differentiated as follows:
  - 3 squared pixels with dimension of 100 × 100 μm^2, SPAD active area of 80×84 μm^2 and front-end circuits with passive quenching, like in array 1.
  - 1 pixel with area of 37×25 μm^2, SPAD active area of 17.5×17.8 μm^2 and front-end circuits with passive quenching, like a subpixel of the SiPM structures.
- TDC: programmable time to digital converter, working with 10 bits (continuous-time mode) or 20 bits, connected to the non-latched outputs of A1, A3 and AAQ. Internal (from a programmable ring oscillator) or external clock sources can be selected. Minimum LSB of 4 ns.

The ASAP110LF chip represents a fully digital detection system. All the digital circuits within the core were designed by employing a mix between custom blocks and standard cells. A summary of the digital signals controlling the operation of the ASAP110LF chip is provided in Table 2.1. The input digital signals can be distinguished between dynamic signals and configuration signals. The dynamic, or control, signals consist of digital waveforms evolving during the regular chip operation. These signals, which can be provided by an external microcontroller or an FPGA (Field Programmable Gate Array), are used to control the logic circuits within the arrays and manage the operation of the ancillary circuits. Dynamic signals are involved in pixel selection, in the enable procedures, in the reset phases, and in the data readout. Some control signals are shared between multiple structures in the core, thus enabling concurrent measurements across different arrays.
Configuration, or static, signals are used to select the operating mode of specific circuits, to enable/disable circuit blocks, to define digital parameters and for multiplexing purposes. In the ASAP110LF chip, there is a total of 16 configuration bits, each provided with a dedicated input pad.

2.3.1 Row and column selection registers

In order to allow the generation of SPAD detection pulses, pixels need to be enabled before the measurement phase. The enabling procedure, performed through the assertion of the EN signal, is effective only for pixels that are

currently selected. A specific pixel is said to be selected if both the internal row and column bits, specific for that pixel, are driven to the low logic value. Row lines and column lines, arranged in multiple signal grids running over the arrays, are used to select the pixels in the different structures of the core. A selecting grid for each array of the core was designed, so as to allow the parallel selection of pixels belonging to different arrays. For each matrix structure, the relevant row and column lines are driven by a pixel selection circuit, like the one shown in Fig. 2.4. The selection network, consisting of a row shift register (RSR) and a column shift register (CSR), is made of library flip-flops. After the initialization procedure, achieved through a positive pulse of the INIT signal, the first output line of each register is set to 0, so as to perform the selection of the first pixel in the array. A single pulse of the INIT signal is sufficient to reset both the row and column shift registers. A single

Table 2.1: Input digital signals.

SIGNAL	PIN NO.	TYPE	STRUCTURE	SIGNAL DESCRIPTION
SPADOFF1	78	STAT	AAQ	selection of hold-off time [b1]
SPADOFF0	79	STAT	AAQ	selection of hold-off time [b0]
THRESHSEL1	80	STAT	ASIPM	selection of SiPM threshold [b1]
THRESHSEL0	81	STAT	ASIPM	selection of SiPM threshold [b0]
TDCSRCSEL1	82	STAT	TDC	selection of TDC source [b1]
TDCSRCSEL0	83	STAT	TDC	selection of TDC source [b0]
TDC20BIT	86	STAT	TDC	selection of TDC working mode
TDCCKSEL	87	STAT	TDC	selection of TDC clock
OSCCKSEL1	89	STAT	TDC	selection of int. osc. freq. [b1]
OSCCKSEL0	90	STAT	TDC	selection of int. osc. freq. [b0]
CELLRESMOD	113	STAT	A1/A2/A3/AAQ/AMSTI	selection of reset type (glob./loc.)
S0	114	STAT	A1/A2/A3/AAQ/AMSTI/ACHT	selection of monost. dur. [b0]
S1	115	STAT	A1/A2/AAQ/ACHT	selection of monost. dur. [b1]
VCONN	116	STAT	A1/A2/A3/AAQ	enabling of the coincidence logic
CONF0	131	STAT	A1/A2	pixel selection mode [b0]
CONF1	132	STAT	A1/A2	pixel selection mode [b1]

SIGNAL	PIN NO.	TYPE	STRUCTURE	SIGNAL DESCRIPTION
INITDOWN	42	DYN	A3/AAQ/ASIPM	reset of selection regs
CKROWACT	43	DYN	AAQ	ck control for row selection
TDCRESET_N	44	DYN	TDC	reset signal of TDC
TDCCKSHIFT	45	DYN	TDC	readout clock of TDC
TDCCKEXT	46	DYN	TDC	external clock of TDC
CKCOLA3	47	DYN	A3	ck control for col selection
CKROWA3	49	DYN	A3	ck control for row selection
TDCIN4	57	DYN	TDC	external input of TDC
TDCLOCK	58	DYN	TDC	freeze signal of TDC
CKROWSIPM	62	DYN	ASIPM	ck control for row selection
SIPMRESET_N	70	DYN	ASIPM	reset signal of SiPM
CKCOLACT	73	DYN	AAQ	ck control for col selection
CKCOLSIPM	74	DYN	ASIPM	ck control for col selection
OSCEN	88	DYN	TDC	enable of int. osc.
INTEG	91	DYN	A2	integration window defin.
RCNT_N	92	DYN	A2	counter reset in A2
CKCOLA2	117	DYN	A2	ck control for col selection
CKROWA1A2	118	DYN	A1/A2	ck control for row selection
OUTRES_N	121	DYN	A1/A3/AAQ/AMSTI	reset of second latch
ENRESGLOB_N	122	DYN	A1/A2/A3/AAQ/ASIPM/AMSTI	pixel disable signal
TRIGRES_N	127	DYN	A1/A3/AAQ/AMSTI	reset of first latch
TEST_N	128	DYN	*All arrays	channel test signal
DATATX	129	DYN	A1/A2/A3/AAQ/AMSTI	data transmission signal
EN	130	DYN	*All arrays	pixel enable signal
CKCOLA1	133	DYN	A1	ck control for col selection
INITUP	134	DYN	A1/A2	reset of selection regs
ROWSTI/SPADEN	140	DYN	AMSTI/ACHT	row selector/pixel enable

*All arrays: A1/A2/A3/AAQ/ASIPM/AMSTI/ACHT

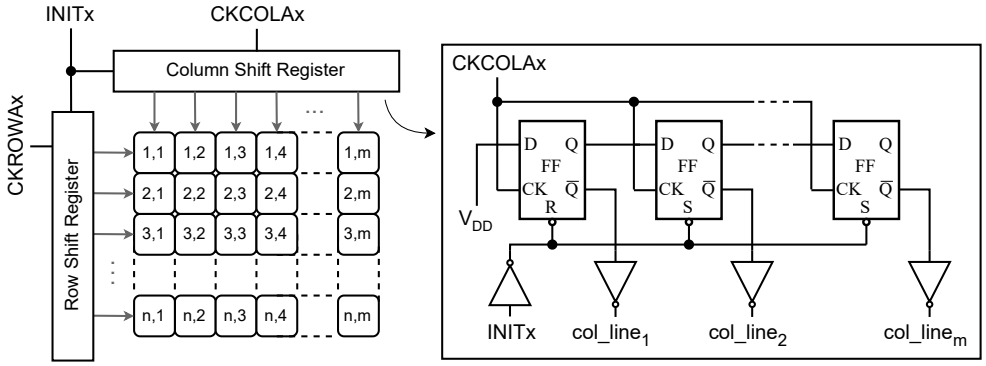


Figure 2.4: Schematic diagram of a pixel selection network based on shift registers.

INIT signal for the reset of both the row and column shift registers was used due to pad limitations. In order to let the selecting 0 move through one of the shift registers, a number of positive pulses must be provided on the clock line of the corresponding circuit. Starting from the reset condition, a pixel located in the (n,m) position (where n and m are the row and column indexes respectively) can be selected by providing $n-1$ row clock pulses and $m-1$ column clock pulses. The shift registers can move the selecting 0s only in one direction. If the number of clock pulses exceeds the number of flip flops in the register, no pixel is selected, as no output line of the shift register is asserted. Except for A1 and A2, which share the same row selecting network, each array, having number of rows higher than 2, is equipped with exclusive row and column shift registers. While dedicated clock signals are used to control the shift registers of each array (CKROWAx and CKCOLAx, where x identifies the array), only two initialization signals (INITUP and INITDN), connected to multiple arrays, are used to reset all the selecting networks in the core. The two rows of ASSTI are selected by means of a 1-bit input signal (ROWSTI/SPADEN), which is also used as pixel enable bit in APXT. In the latter, a row selection circuit was not implemented, since SPADs are arranged in a single line. The columns of AMSTI and APXT are addressed through the column shift register of A1.
The RSRs of A1, A2 and A3, which represent the largest arrays in the chip as far as the number of pixels is concerned, can be programmed to assert multiple row lines at a time, through the CONF0 and CONF1 configuration bits. Three alternative operating modes were developed to speed-up the enabling procedure and allow the parallel reading of pixels located in the same array.

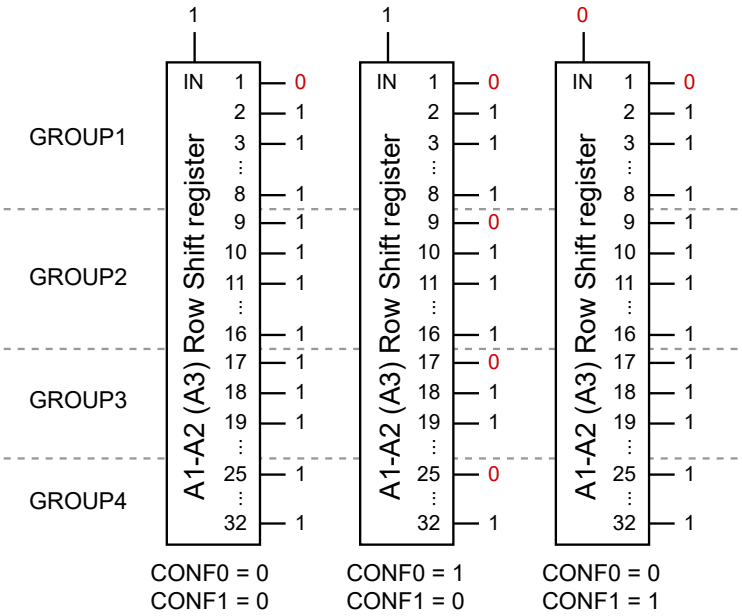


Figure 2.5: Output values of the RSRs of A1, A2 and A3, under reset condition, in the three different operating modes (the RSR of A3 is equipped with only 16 output ports).

A schematic representation of the different operating modes, as a function of the CONF0 and CONF1 configuration bits, is shown in Fig 2.5.

- CONF0 = 0, CONF1 = 0. Parallel assertion of the row lines is disabled and rows are selected serially. A number of clock (CKROWAx) cycles equal to $n_{MAX} - 1$ is needed to assert all the row lines, one at a time, in the array (n_{MAX} identifies the number of rows in the array).

- CONF0 = 1 and CONF1 = 0. Upon a reset (INIT) pulse, a row line every eight is asserted. This operating mode was developed to perform a parallel readout of A1 and A3, which are internally subdivided into 8-rows blocks (see sections 2.3.2 and 2.3.4). Since each block is provided with independent output pins, the data produced by the enabled cells can be read out in parallel, thus reducing the acquisition time. In this operating mode, all the rows in the array are reached within seven clock cycles, thus resulting in a substantial speed up of the readout operation.

- CONF0 = 0 and CONF1 = 1. In this configuration, a logic 0 is fed as input to the shift register. At each clock cycle, a new selecting 0

enters the shift register, and the number of simultaneously selected rows increases linearly with the number of clock pulses. In order to fill all the RSR with zeroes, starting from the reset condition, a number of clock pulses equal to $n_{MAX}-1$ must be provided. This operating condition can be used to reduce the time needed to enable all the pixels in a column. However, after the enabling operation, a different selection mode must be used for pixel readout, in order to prevent the readout combinational network of each array from working in a non-deterministic condition due to the high number of pixels concurrently connected to the same output common bus (see section 2.3.2).

Fast selection networks, with multiple options for parallel pixel enabling, help reduce the complexity of the external readout algorithms, and promotes the acquisition of data at high speed, thus significantly reducing the measurement time. In particular, some measurements may require the alternate selection of different pixels, as well as the re-iterated parsing of all the pixels in the array. The availability of a global enable signal, simultaneously provided to all the cells in a chip or in a single array, was not taken into account during the design of the pixel selection mechanism. Only the capability of enabling a single cell at a time was considered, as it is required by most of the measurements of interest. In the case of arrays with a relatively high number of cells, the time needed for pixel enabling is strongly affected by the operating mode of the selection circuits. Fig. 2.6 shows the time diagrams of the waveforms

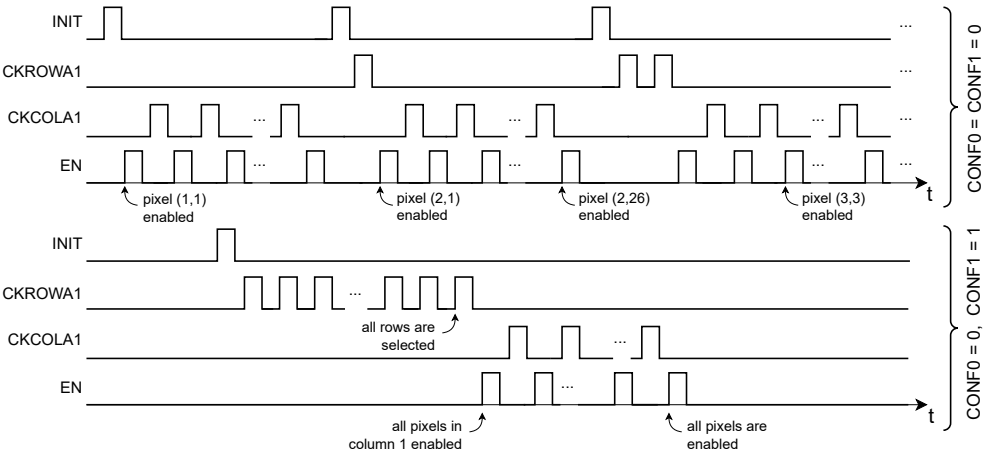


Figure 2.6: Time diagrams of the array enabling procedure in the two cases CONF0=CONF1=0 and CONF0=0, CONF1=1.

involved during the pixel enabling procedure of A1. Two different selection modes are described in the diagrams. If none of the parallel operating modes is used (CONF0=CONF1=0), 2.6 signal pulses per pixel must be provided, on average, to enable all the cells of the array (2159 pulses, subdivided in 800 CKCOLA1 pulses, 31 INIT pulses, 496 CKROWA1A2 pulses and 832 EN pulses). By using CONF0=0 and CONF1=1, only 82 pulses (31 CKROWA1A2 pulses + 25 CKCOLA1 pulses + 26 EN pulse) are needed to transmit the enable command to all the pixels in A1, starting from the reset condition. Therefore, if a high number of pixels must be enabled, parallel pixel selection should be considered, since it represents an effective way to reduce the measurement time and increase the acquisition rate.

2.3.2 Array 1 (A1)

Array 1 consists of 832 pixels arranged in a 32×26 matrix structure. The SPADs in the array, based on a p+/nwell junction, have an active area of $80 \times 84\ \mu m^2$. The cathodes of all the sensors are connected to the VK1 pad, which is distributed across the pixel rows, through horizontal lines. The array is divided into four equivalent subarrays, or groups, as shown in Fig. 2.7. Each group, made of eight rows of pixels, is provided with dedicated output pads (TRIGx and OUTx, with x=1, 2, 3, 4), so as to allow the parallel reading of data. Rows and columns are selected by using the shift register approach described in section 2.3.1.
In Fig. 2.7, the layout view of a single pixel is also included. Pixels are square-

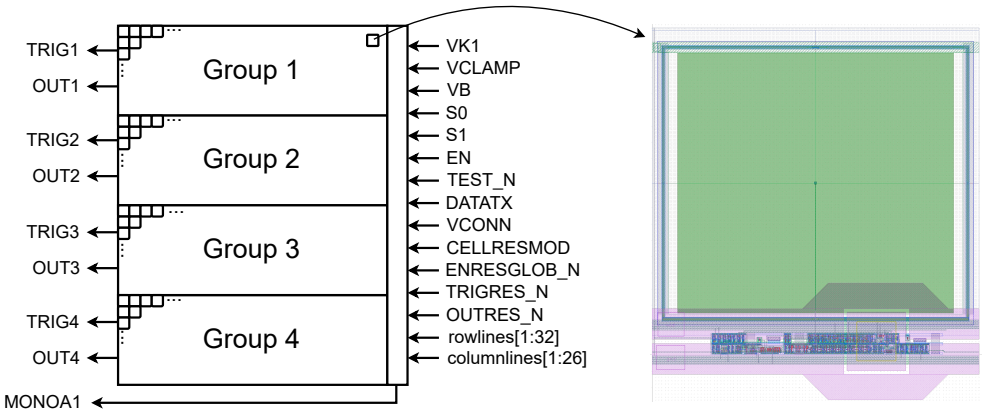


Figure 2.7: Schematic representation of A1, with the layout view of a pixel in the array.

shaped, with an area of $100 \times 100\ \mu m^2$. The size of the SPADs in this array, together with the size of the SPADs in array A3 ($50 \times 50\ \mu m^2$, to be described later on), was chosen with the purpose of studying possible dependence of the main parameters on the device dimensions. In the pixel plane, separated regions are dedicated to the SPAD sensor and to the processing electronics. A FF of 67% was attained for pixels in this array. Most part of the cell is occupied by the sensor, which is located above the in-pixel electronics. The cathode contact ring encloses the active area, while the anode terminal is connected to the readout circuit through an array of multiple contacts located in the center of the squared surface. It is worth specifying that the sensor representation depicted in the figure represents a placeholder, since the actual stack of layers, used for the design of the SPADs, is under a non disclosure agreement. Even though not visible in figure, the corners of the sensor are cut with 45 degree angles, in order to reduce non-linear effects which may come due to a non-uniform distribution of the electric field. The pixel surface corresponding to the active area was not covered with passivation layers, so as not to jeopardize the photon detection capability of the structure.

2.3.2.1 A1 in-pixel electronics

The in-pixel electronics of cells in A1 can be divided into enable network and readout network. A circuit diagram of the enable network is shown in Fig. 2.8. This circuit is used to manage the internal signal (*en_inside*) in charge of activating the quenching circuit and the pixel readout network. The enable

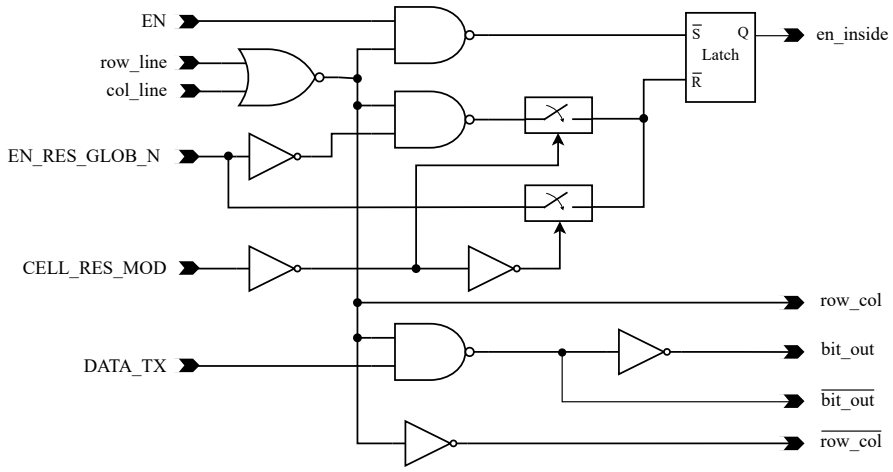


Figure 2.8: Circuit diagram of the enable network for pixels in A1.

EN	EN_RES_GLOB_N	CELL_RES_MOD	en_inside
positive pulse	1 (not active)	X (don't care)	1 (locally)
0 (not active)	negative pulse	0	0 (locally)
0 (not active)	negative pulse	1	0 (globally)

Table 2.2: Enable and disable procedures for the selected pixel.

network consists of a combinational logic, driven by dynamic and configuration signals, and a set-reset (SR) latch used to store the digital value of the enable bit. Once the pixel is selected ($row\_line = col\_line = 0$), the internal signal *row_col* is set to 1, thus making the memory element sensitive to a positive pulse of the EN signal. As the SR latch is set, the pixel is enabled to produce a detection pulse. EN pulses with a duration $\geq 20\ ns$ are needed to set the SR latch. In order to reset the *en_inside* bit, thus performing the pixel disable operation, a high-to-low transition of the EN_RES_GLOB_N signal must be provided, while keeping EN = 0. The pixel disabling can be performed globally (in all the pixels of the array) or locally (only in the selected pixel), according to the CELL_RES_MOD configuration bit, as shown in Table 2.2.
The readout circuit occupies most of the area dedicated to the pixel electronics. The circuit diagram of the readout network is shown in Fig. 2.9. All the transistors involved in the circuit are core devices (supporting the maximum operating voltage of $1.2\ V$), apart from the MOSFETs used in the quenching circuit. Three main blocks, connected in cascade, can be identified into the readout electronics: the front-end circuit, the coincidence section and the 1-bit memory. All these sections will be separately described in the following.

Front-end circuit. In the front-end circuit, the current pulse generated by the SPAD undergoes a series of processing steps aimed at producing an output pulse with standard duration. The interface between the SPAD sensor and the readout network is represented by the passive quenching circuit, consisting of an NMOS transistor (M1) driven by the gate voltage VB. The current flowing through the device, during the quenching time, can be tuned by changing the quenching gate voltage. This makes it possible to control the amount of the avalanche charge, which may impact on the dark noise produced by the SPAD [64]. The quenching circuit is enabled, upon the assertion of the *en_inside* signal, by means of the NMOS transistor (M2), in series with the quenching current source. If the gate of the M2 transistor is driven to the low logic value (by the network in Fig. 2.8), the avalanche current is prevented from

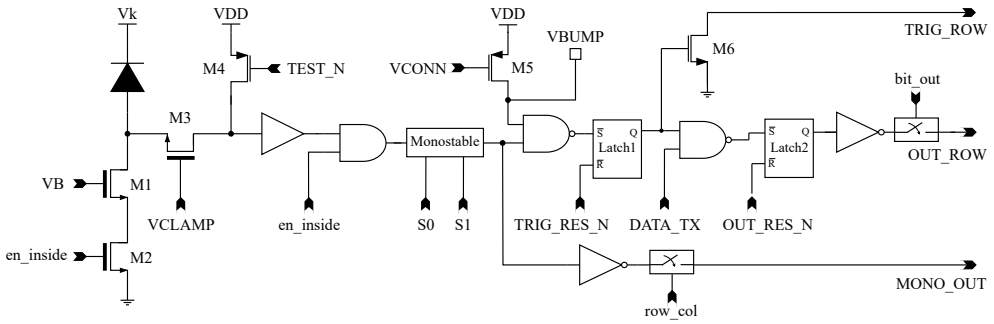


Figure 2.9: Circuit diagram of the readout network in a pixel of A1.

flowing through the SPAD. Therefore, as long as the *en_inside* signal is at 0, no detection pulse can be generated at the anode on the sensor. Thick oxide devices, supporting an operating voltage of 3.3 V, were used for the implementation of the quenching circuit, so as to enable the use of SPAD excess voltages above the core supply voltage. In order to prevent permanent damage to the front-end circuit, excess voltages not higher than 3.3 V should be applied to the SPAD. Transistor M3 is used to adapt the amplitude of the SPAD output pulse to the voltage range supported by the subsequent electronics ($0 - 1.2\ V$). When an avalanche is triggered, the amplitude of the pulse generated at the anode terminal is equal to the excess voltage (V_{EX}), which can be higher than the nominal operating voltage of the core readout electronics. Transistor M3, implemented by means of a thick oxide device, is used to clamp to 1.2 V the voltage pulse provided to the input of the subsequent digital buffer. In order to accomplish the clamping operation, the gate voltage of M3 (VCLAMP) should be equal to $1.2 + V_{th}$, where V_{th} is the transistor threshold voltage. From simulations, VCLAMP values around 1.6 V were obtained. A digital buffer, with a switching threshold around 500 mV, is used to sharpen the edges of the detection pulse, after the clamping operation. The input terminal of the digital buffer is connected to a pull-up transistor (M4). This device, driven by the TEST_N signal, is used to test the readout electronics, by simulating a SPAD firing event. A TEST_N negative pulse must be provided to assess the working condition of the readout circuit. However, it is worth specifying that the quenching circuit is excluded from the readout circuit test. The detection pulse, generated by the digital buffer, is gated by means of the *en_inside* signal. The monostable circuit sets the duration (δt) of the detection pulse, according to the values shown in Table 2.3. The circuit diagram of the monostable is shown in Fig. 2.10. In transparent mode, no

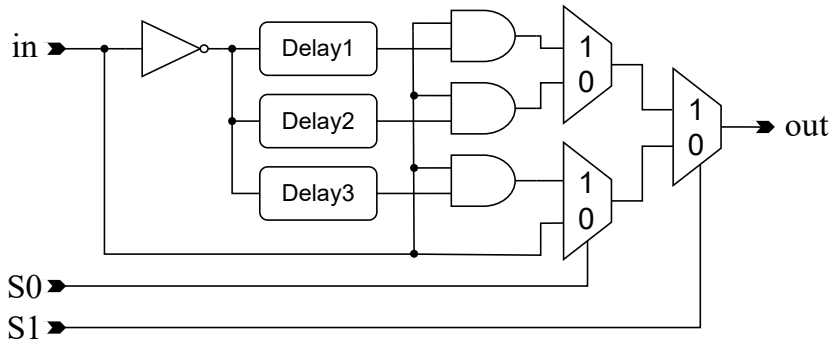


Figure 2.10: Circuit diagram of the monostable circuit used in pixels of A1.

S1	S0	pulse duration [ps]
0	0	transparent mode
0	1	400
1	0	800
1	1	2000

Table 2.3: different pulse duration as a function of the S0 and S1 bits in A1 (values obtained from post-layout simulations in nominal conditions). Actual values may differ by ±30% due to process parameter variations.

change is applied to the duration of the SPAD pulse, which depends mainly on the parasitic capacitance loading the anode terminal and on the equivalent resistance featured by the quenching network. The output signal produced by the monostable circuit of the selected pixel is sent to the MONO_A1 output pad, through the combinational network shown in Fig. 2.11. Pixels of the same row are divided into four groups of 6 or 7 cells. Within each group, pixels share the same MONO_OUT common bus, which is kept at the high logic value by means of a weak pull-up transistor. Each pixel is connected to the common bus through a tristate inverter, which is controlled by the *row_col* signal, generated in the enabling circuit. If a pixel is selected, the corresponding tristate inverter takes control of the MONO_OUT common bus. It can be worth specifying that the MONO_OUT common bus can never result in a multi-driven condition, since a single pixel at a time is allowed to take control of the bus. This is guaranteed by the column selection procedure, that does not allow for parallel column selection.

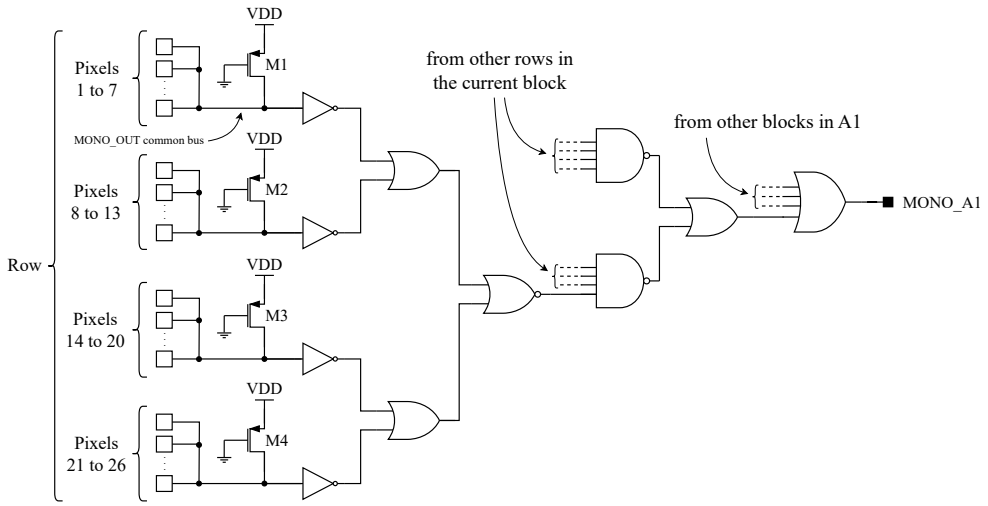


Figure 2.11: Combinational circuit connecting the pixels in A1 to the MONO_A1 output pad.

Coincidence section. The coincidence network consists of a NAND gate, a pull-up transistor (M5) and a bump bonding pad. The aim of this section is to reveal the coincidence of pulses generated in two different SPAD layers. According to the dual layer approach described in chapter 1, the signal produced by a pixel in the son chip is connected, through bump bonding pad, to the coincidence network of the corresponding pixel located in the father chip. The two signals, generated by the monostable circuits of the two layers, are fed to the NAND gate. If a coincidence of pulses is detected, the output of the NAND gate is set to 0. The duration of a coincidence pulse, generated as a consequence of a particle detection, corresponds to δt. In the case of dark count events, the duration of the NAND output pulse is not fixed, as the two monostable signals may overlap only partially. The bump bonding contact between the two chip layers does not provide a significant parasitic capacitance to the vertical connection, thus marginally affecting the duration, as well as the propagation delay, of the signal generated in the son layer. Measurements performed on the readout channels of the APIX2LF chip, aiming at testing the effectiveness of a similar coincidence network, revealed that the coincidence signal is correctly generated, with a yield of 100%, for monostable nominal durations equal to or larger than 400 ps.
The weak pull-up transistor (M5), which can be disabled through the configuration bit VCONN, keeps at high logic value the NAND input port connected

VCONN	Operating mode
0	Single layer
1	Dual layer

Table 2.4: Operating modes of the coincidence network depending on the VCONN configuration bit.

to the bump bonding pad, so as to allow the separate test of single layer chips. Table 2.4 shows the two operating modes of the coincidence section as a function of the VCONN signal.

1-bit memory. As shown in Fig. 2.9, the in-pixel memory is made of two SR latches connected in cascade through a NAND logic gate. The detection pulse, generated by the coincidence network, is used to set the output of the first latch, which can be reset through the TRIG_RES_N signal (active low). Simulations, performed at all the process corners, revealed that the SR latch can correctly set its output for all the monostable configurations. However, in a dual layer pixel, the pulse width fed to the input of the latch can be significantly shorter than 400 ps, as the two monostable pulses may overlap only partially, since they result in general from uncorrelated noise events taking place in the two layers. In this case, the minimum pulse width that can trigger the latch is, nominally, 130 ps. The reset of the first memory element is a global operation, affecting all the pixels in the array. The output signal of the first memory block drives the gate of an NMOS transistor (M6). The latter is attached to the TRIG_ROW common bus, which is shared by all the pixels in a row, as shown in Fig. 2.12a. If a cell in a row produces a detection pulse, the TRIG_ROW bus is asserted and the information bit is sent to the output pad through an OR-based combinational network. Each block of eight rows has a dedicated TRIGx pad. Before being stored in the second memory element, the output bit from the first latch is gated by the DATA_TX signal. When DATA_TX is asserted, the NAND gate acts as an inverter, thus making the second memory element store the same bit data as the first one. The output signal of the second latch is connected to the OUT net by means of a tristate inverter. The latter is controlled by the *bit_out* signal, which is set by the enabling circuit upon the assertion of the DATA_TX signal. The combinational network connecting the OUT port of each pixel to the output pad is shown in Fig. 2.12b. The OUT_ROW common bus, used by all the pixels in a row, is kept at the high logic value by a weak pull-up transistor,

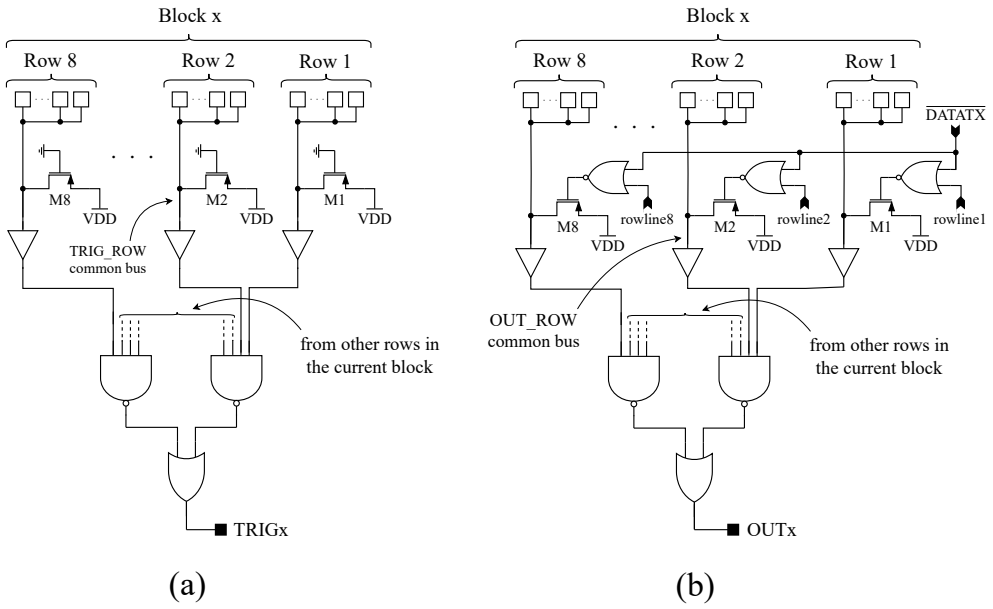


Figure 2.12: Combinational network connecting the pixels in A1 to the TRIGx and OUTx output pads (rowlinex signals are active low).

which is disabled if the DATA_TX signal is 1. An OUTx pad for each block of eight rows is available. The bit stored in the second memory block is reset in a global way, by means of the OUT_RES_N signal (active low).
The implementation of two cascaded memory blocks allows the acquisition of data, generated upon actual detection events, through a triggered read mode. An example of this reading mode is shown in Fig. 2.13. In triggered read, after enabling all the pixels in the array, the system can be set into an idle condition, waiting for a detection event. Upon a SPAD firing, the TRIGx of the corresponding block is set to 1, indicating that at least one SPAD in the 8-rows group has produced a detection pulse. Even though the TRIGx signal is representative for all the 208 pixels contained in a block, only the first latch of the fired cells can be found to store a data different from 0 (noise events are not taken into account in this discussion). The exact location of the fired pixels can be addressed by parsing all the cells in the block featuring TRIGx = 1, looking for the row and column couple returning OUTx = 1, while DATA_TX is asserted.

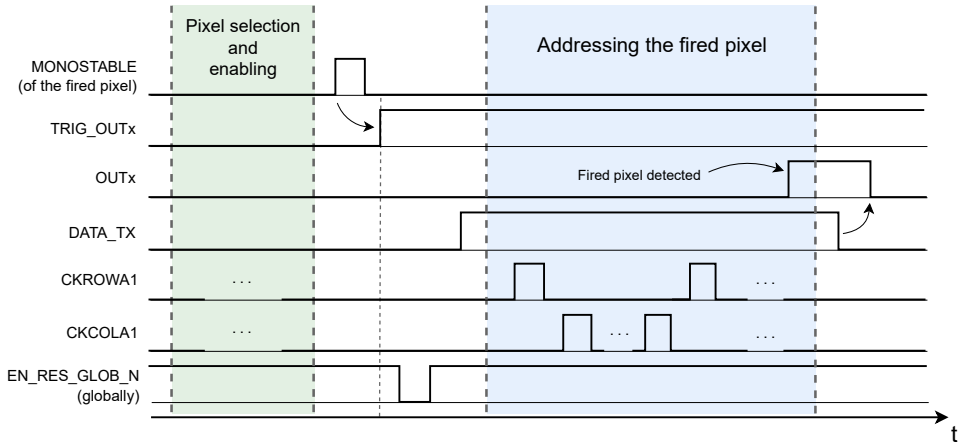


Figure 2.13: Time diagram for triggered read mode.

2.3.3 Array 2 (A2)

Array 2 is made of 32×22 pixels, arranged in a unique block. The input/output signals involved in the A2 operation are shown in Fig. 2.14. While most of the input signals are shared with other structures in the core, INTEG and RCNT_N are specific for A2. The output signals from this array consist of ten bits (OUTBITXx), representing the result of a binary count. The row lines in A2 can be addressed by using the same row shift register of A1, thus allowing parallel pixel enabling. The layout of pixels in A2 is included in Fig. 2.14. The active area of the in-pixel sensor, which receive the cathode voltage from the VK2 pad, is 90×72 μm^2. The layout structure of pixels in A2, featuring a FF of 64%, is similar to the one already discussed for pixels of A1.

2.3.3.1 A2 in-pixel electronics

As for A1 cells, the electronics of the A2 pixels can be divided in enable circuits and readout network. All pixels in A2 implement the same enable circuits, front-end network and coincidence section already discussed for A1. The circuit diagram of the readout network used in pixels of A2 is shown in Fig. 2.15. Each cell is provided with a 10-bit asynchronous counter, which can be controlled through dedicated input signals. The pulse generated by the monostable circuit of these pixels is not available on the output pads. Therefore, the ten bits of the asynchronous counter represent the only output signals produced by the readout networks of this array. The in-pixel counter is used to count the number of detection (or dark noise) pulses, generated within the

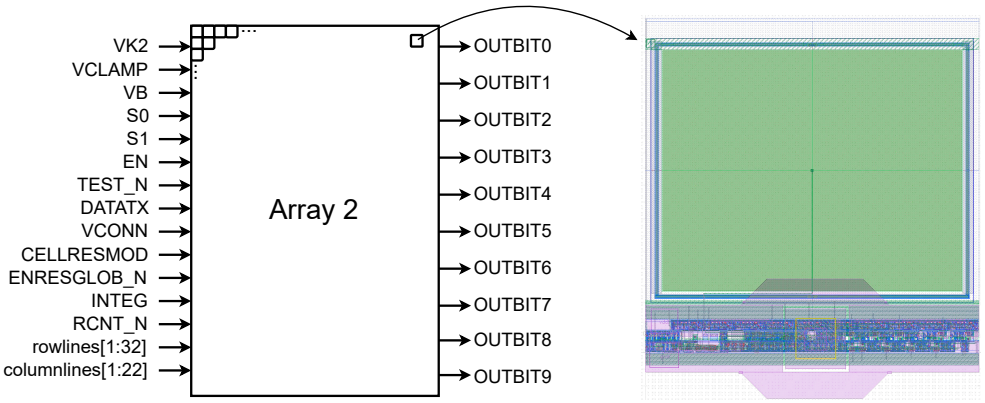


Figure 2.14: Schematic representation of A2, with the layout view of a pixel in the array.

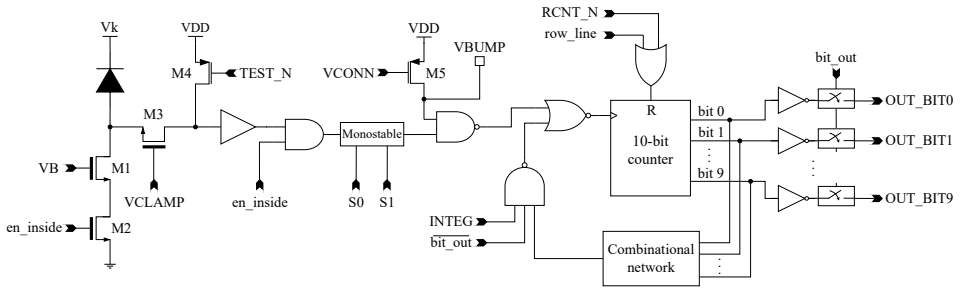


Figure 2.15: Circuit diagram of the readout network in a pixel of A2.

front-end network. The asynchronous architecture has been adopted, in place of a synchronous one, to optimize the fill factor of the pixel. The toggling latency featured by the different flip-flops of the counting chain does not represent a significant problem for the pixel functionality, since the expected time interval between adjacent pulses is significantly large, in average, if compared to the settling time of the ten elements of the circuit. In the worst case, occurring when toggling from "'b1111111110" to "'b0000000001", the time needed by all the output bits of the counter to settle is 2.3 ns (obtained through post-layout simulations), which is even lower than the SPAD reset time. Potential alterations of the settling time due to process variations, taken into account during the circuit design, should not undermine the functionality of the structure, since the expected counting rate is well below the maximum one allowed by the circuit. The counter reset is performed at the row level,

through the RCNT_N input signal (active low). The INTEG signal (active high) determines the integration window. As long as INTEG is asserted, the 10-bits counter updates the count result upon each pulse generated by the coincidence network. The counter is prevented from running into overflow by a combinational circuit, blocking the propagation of avalanche pulses if the count result reaches 1020. In such a condition, a RCNT_N pulse is needed to reset the counter and start a new count. 1020 was chosen, in place of $2^{10} - 1$, as full scale count value, in order to reduce the area occupation of the gating network used to prevent overflow.
When the DATATX signal is asserted (then also bit_out is asserted), each bit of the count result is connected to the corresponding routing network, through tristate inverters. Under this condition, the counting operation is inhibited, so as to prevent the counter from updating during the reading procedure. One of the ten routing networks connecting a single output bit of the counter to the dedicated output pad is shown in Fig. 2.16. Each bit of the in-pixel counter is connected to a common bus, which is shared by all the pixels in a row. If a row is not selected, the corresponding common bus is kept at the high logic value through a pull-up transistor, which is controlled by the DATATX signal. For the sake of the signal routing, the 32 common buses are divided

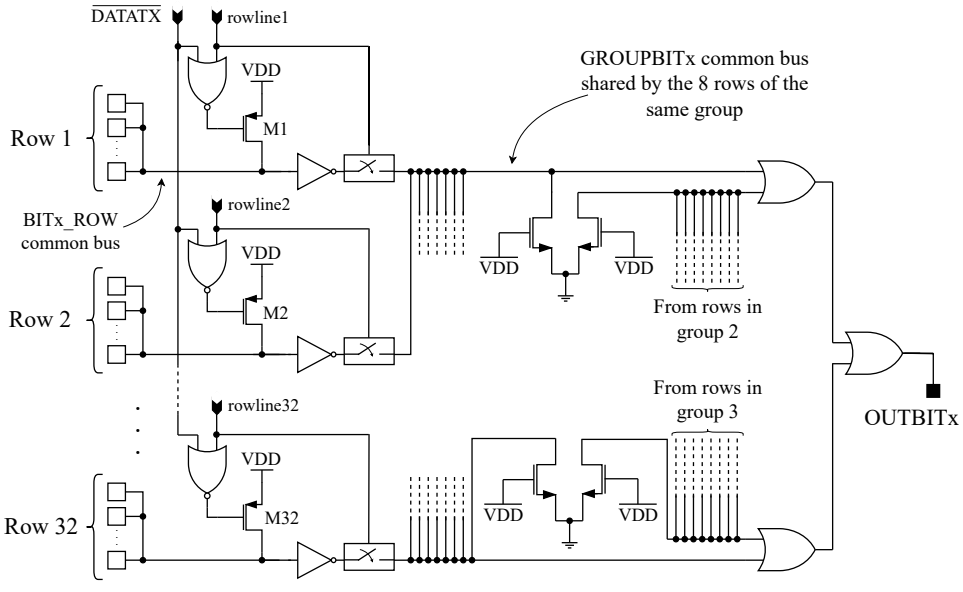


Figure 2.16: Routing network connecting a count bit, with x binary weight, to the corresponding output pad (rowlinex signals are active low).

into four groups. All the common buses in a group are connected to the same GROUPBITx net, which is, in turn, a shared bus. The tristate inverters, connecting the BITx_ROW buses to the shared GROUPBITx net, are controlled by the rowline signals. Through an OR-based network, the count bits reach the output pads. It is worth specifying that ten copies of the network shown in Fig. 2.16 are implemented, so as to connect all the count bits to the dedicated output pads. Pixels in A2 can be enabled in parallel, by exploiting the capabilities offered by the A1-A2 row shift register. However, due to the architecture of the routing network shown in Fig. 2.16, parallel reading of multiple pixels is not allowed. Therefore, during the bit reading operation, CONF0 = CONF1 = 0 should be used, in order to let a single pixel at a time to be connected to the output pads.

2.3.4 Array 3 (A3)

Array 3 consists of 16×52 squared pixels. Each cell is 50×50 μm^2, with SPADs having an active area of 30 × 40 μm^2. A schematic diagram of the array is shown in Fig. 2.17. The structure is divided into two blocks, or groups, of 8 rows. Each block is equipped with two output signals, which can be read on dedicated output pads (TRIGx and OUTx, with x=5,6). The pixel layout is included in Fig. 2.17. With the layout design shown in figure, a FF of 48% was attained. The SPAD cathode voltage is fed into the structure through the VK3 pad. Due to the small pixel size, the metal tracks used to distribute the supply voltages overlap with a significant portion of the pixel active area. However, in view of applications involving particle detection, the fill factor of

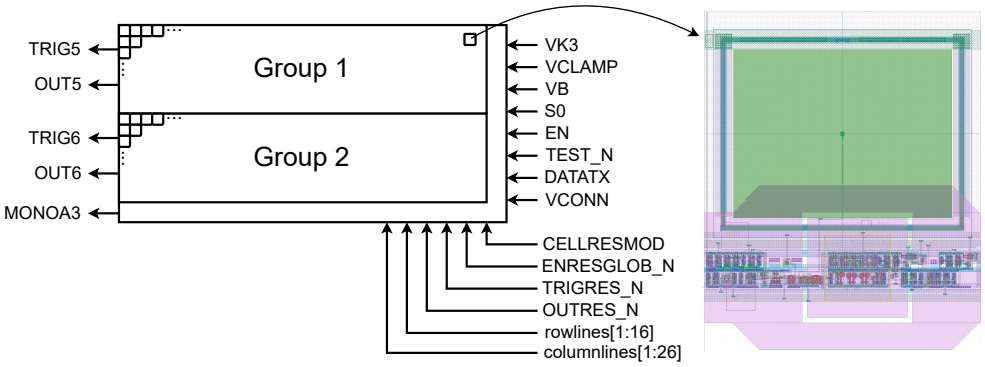


Figure 2.17: Schematic representation of A3, with the layout view of a pixel in the array.

S0	pulse duration [ps]
0	transparent mode
1	2000

Table 2.5: Different pulse duration as a function of the S0 configuration bit in A3.

the structure is not affected by the use of large metal tracks, that allow a good power distribution while negligibly interfere with the particle path.

2.3.4.1 A3 in-pixel electronics

The A3 pixels implement an in-pixel electronics similar to the one shown in Fig. 2.8 and Fig. 2.9. The main difference lies in the monostable circuit, which, in the case of A3, is controlled by a single configuration bit (S0), to save area. Therefore, only two values are available for the duration of the monostable pulse, as shown in Table 2.5. The signal generated by the monostable circuit can be accessed on the MONOA3 output pad. Since the row shift register of A3 can be configured to simultaneously select multiple cells, the parallel reading of pixels, located in different blocks, is allowed.

2.3.5 Array ACT-Q (AAQ)

The array ACT-Q (active quenching) is made of 7×10 squared pixels. A schematic diagram of the array is shown in Fig. 2.18. Pixels in this array are equipped with active quenching circuits, which are managed through the configuration bits SPADOFF0 and SPADOFF1. The 3.3 V voltage reference, provided at the chip core through a dedicated pad, is used for the generation of the waveforms involved in the active quenching operation. Due to the moderate number of rows, the array was not subdivided into groups, thus multiple pixel reading is not allowed. The output signals, generated by pixels in AAQ, are available on dedicated pads (MONOACT, OUTACT). The layout of a pixel in AAQ is included in Fig. 2.18. The SPAD cathode voltage is provided by means of the VK4 pad. Each cell, having a size of 100×100 μm^2, is equipped with the same sensor active area used in A1 (80×84 μm^2). A FF of 67% was obtained with the layout structure shown in the figure.

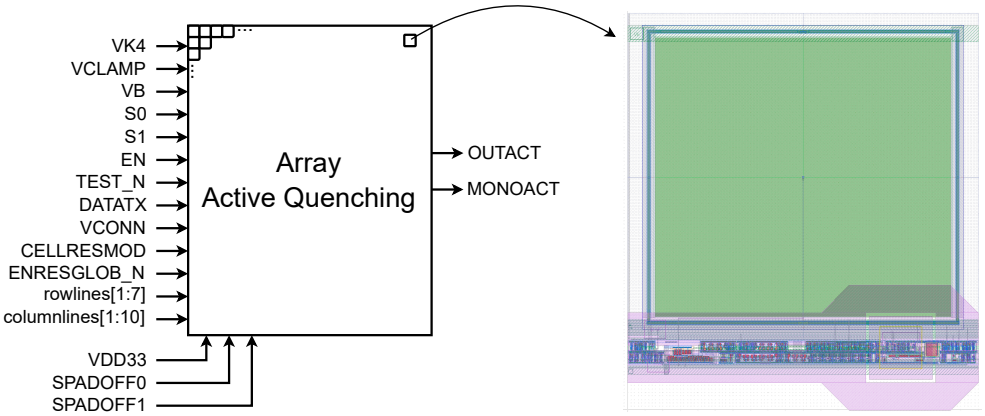


Figure 2.18: Schematic representation of AAQ, with the layout view of a pixel in the array.

2.3.5.1 AAQ in-pixel electronics

The in-pixel electronics implemented within cells of AAQ is made of the enable circuit depicted in Fig. 2.8 and a readout network. The circuit diagram of the readout network is shown in Fig. 2.19. In the AAQ pixels, the quenching operation and the reset phase are performed by using an active quenching circuit, consisting of a couple of thick oxide transistors (MP and MN) and a feedback network made of two monostable circuits. When the avalanche is generated, the quenching is started in a passive way by the M1 transistor, which is controlled by the reference voltage VB. The signal produced by the front-end circuit is fed to a feedback network, which enables, in different time windows, the quenching transistor MP and the reset transistor MN. A schematic time diagram showing the output signals generated within the active feedback network is depicted in Fig. 2.20. As the monostable output is set, thus indicating the SPAD is firing, a pulse signal, with a programmable duration, is generated by monostable 2. This signal is used to complete the quenching operation in a fast way, by enabling the PMOS transistor. It can be worth specifying that, if provided with a suitable VB voltage level, M1 may completely attain the quenching operation before the activation of MP. After the quenching phase, MP is kept enabled for a fixed amount of time, so as to apply a hold-off time to the sensor. The duration of the pulse generated by monostable 2, determining the hold-off time, can be programmed through the SPADOFF0 and SPADOFF1 configuration bits, according to the values show in Table 2.6. MP can be also used to simulate a SPAD firing event, allow-

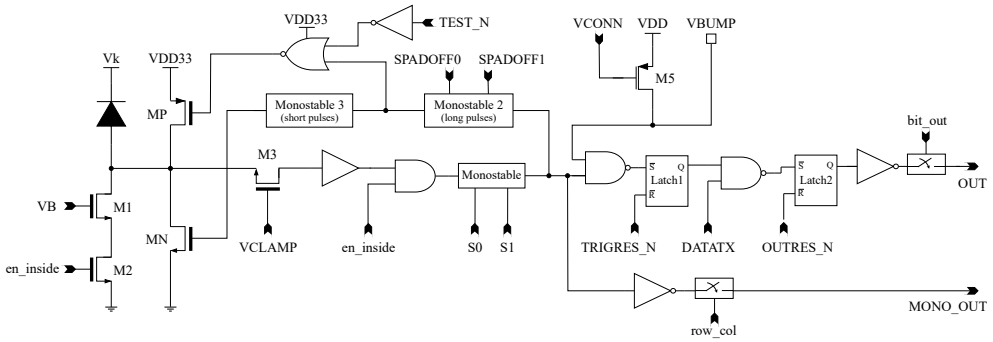


Figure 2.19: Circuit diagram of the readout network in a pixel of AAQ.

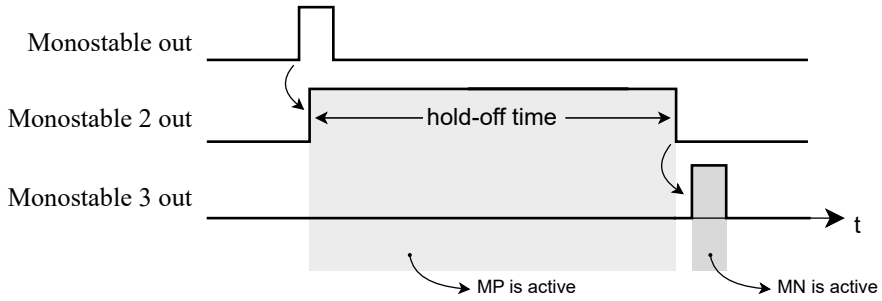


Figure 2.20: Schematic time diagram showing the output signals generated by the different monostable circuits implemented in pixels of AAQ (pulse durations are not to scale).

ing the test of the readout electronics. The logic gate, performing the NOR function between the output signal of monostable 2 and the TEST_N signal, is made of thick oxide transistors, since it performs the level shifting between 1.2 V and 3.3 V. After the hold-off time, the falling edge of the signal generated by monostable 2, triggers the activation of monostable 3, which enables the MN transistor for a small amount of time. During this time interval, the anode voltage is discharged, with a fast transient, through transistors MN and M1. When the reset phase is complete, the starting bias condition of the SPAD is recovered.
The diagram of the circuit implementing the monostable 2 function, generating long pulses, is shown in Fig. 2.21. Since the target of this circuit is the generation of hold-off time intervals in the order of 10^{-7} s, an architecture

SPADOFF1	SPADOFF0	hold-off time [ns]
0	0	30
0	1	70
1	0	110
1	1	150

Table 2.6: Nominal hold-off time as a function of the SPADOFF0 and SPADOFF1 bits (values obtained from post-layout simulations). Actual values may differ by ±40% due to process parameter variations.

based on a charging capacitance was preferred to one using cascaded logic gates, for the sake of the area occupation. The capacitor involved in the pulse generation was achieved by exploiting the gate oxide capacitance of a MOSFET. When the input signal is zero, the capacitor is charged to VDD, thus keeping the output latch in the reset condition. As soon as the detection signal is generated, the NMOS transistor, which has higher driving capability if compared to the pull-up branches, discharges the capacitor, with a fast transient (lasting a few hundreds of picoseconds). Concurrently, the latch is set by the input signal. In order to avoid a condition where both the input ports of the latch are in the active state, a delay is interposed between the input signal and the S pin of the memory block. When the input signal takes the low logic value, the NMOS transistor is switched off and the MOSFET capacitance enters the charge phase. The PMOS transistors having the drain terminal connected to one of the capacitor plates (M2, M4 and M6) are used as MOS switches, selectively enabling the different pull-up networks. Depending on the SPADOFF0 and SPADOFF1 configuration bits, the voltage across the capacitor is discharged with different time constants, since M1, M3 and M5 have different channel lengths. The gates of the driving PMOS transistors are controlled by the node to be charged, so as to slow down the charging phase. When the voltage across the capacitor reaches the switching threshold of the following inverter, the output latch is reset, determining the falling edge of the hold-off pulse. With this mechanism, relatively large hold-off time intervals can be obtained, with reduced area occupation. The nominal values of the available hold-off time intervals are shown in Table 2.6. Depending on possible fluctuations in the process parameters, the hold-off time may depart from the nominal values. However, the ratio between the four time intervals is expected to be at least roughly preserved, since it depends on the ratio between the dimensions of the PMOS transistors.

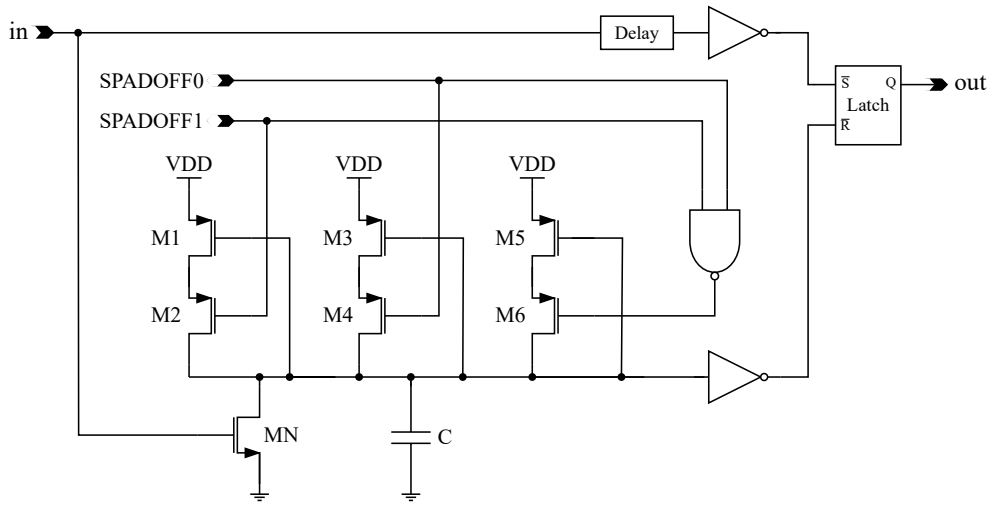


Figure 2.21: Circuit diagram of the monostable circuit determining the hold-off time.

The coincidence section and the data storing network, based on a 1-bit memory, are similar to the ones integrated in pixels of A1. Since the output of the first memory element is not available on the output pads, pixels in AAQ cannot be read in triggered mode. The two output signals, produced by the pixels in this array, are routed towards the output pads (MONOACT and OUTACT) through combinational networks similar to the ones used in A1.

2.3.6 Array SiPM (ASIPM)

ASiPM, located in the bottom right corner of the chip, is made of 5×12 pixels, each of them representing an independent digital SiPM. A simplified diagram describing the array structure is shown in Fig 2.22. Each dSiPM ($162 \times 100\ \mu m^2$), representing a single pixel, consists of 16 SPADs, arranged in a 4×4 array structure, and a readout logic, located beneath the 16 cells. The output of each SiPM consists of a 5-bits binary word, representing the number of simultaneously fired SPADs, and a signal over threshold (SOT) bit, indicating whether the number of generated pulses overcomes a fixed threshold value. The 5-bits count result and the SOT flag are provided respectively on the SIPMBITx (with x = 0, 1, 2, 3, 4) and SIPMSOT output pads. The output signals are routed to the output pads through a combinational network, similar to the one shown in Fig 2.12. Within an independent SiPM, the 16 cells building up the structure, are referred to as subpixels. All the subpixels

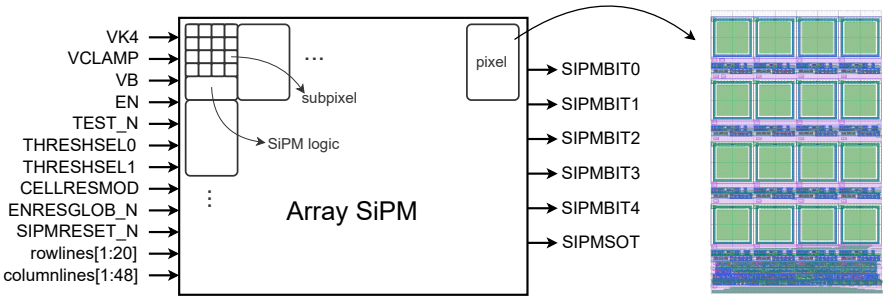


Figure 2.22: Schematic representation of ASIPM, with the layout view of a pixel in the array.

contain a squared sensor, with an active area of $17.5 \times 17.8\ \mu m^2$, and a reduced front-end network connected to the SiPM readout circuit. Totally, 960 SPADs (20×48) are included in the structure. The row and column selection can be performed at the subpixel level, through dedicated selection shift registers controlled by the INITDN, CKROWSIPM and CKCOLSIPM dynamic signals. As a result, the subpixels in the array can be independently enabled, thus allowing to selectively mask high noise SPADs. The output signals of a particular SiPM are available on the output pads if one of the corresponding 16 subpixels is selected. In this case, the SiPM is said to be selected as well. Since the ASIPM selection shift registers do not allow parallel assertion of the row and column lines, a single SiPM at a time can be selected.
The FF of the subpixel, considering only the single SPAD and the front-end circuit, is 34%, while the FF of the entire SiPM, adding the area contribution given by the relevant readout electronics, attains 30%. A group of 12 SiPMs (rows 1 to 4, columns 10 to 12) was designed with SPADs sharing the lateral cathode contacts, thus creating larger active areas ($22 \times 18\ \mu m^2$), while keeping the total area of the SiPM unvaried. For these pixels, a significant gain, in terms of fill factor ($FF_{subpixel} = 41\%$ and $FF_{pixel} = 37\%$), was achieved. The cathode of all the SPADs in this array are connected to the VK4 pad.

2.3.6.1 SiPM electronics

The digital SiPMs developed in the framework of this Ph.D. thesis are based on a novel architecture exploiting the fast response of parallel counters [124]. The structures integrated in ASIPM represent first prototypes of this architecture, aiming at studying the circuit feasibility while identifying the most critical design challenges.

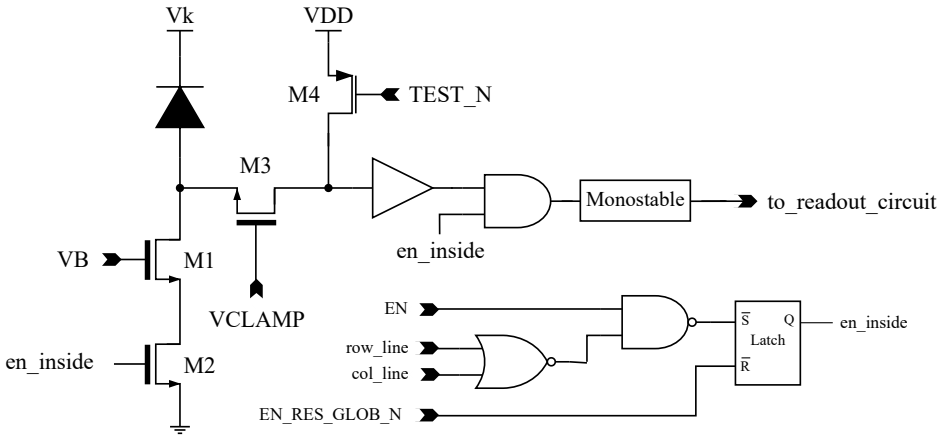


Figure 2.23: Front-end circuit integrated in the subpixels of ASPIM.

Each subpixel of the structure is equipped with the front-end circuit shown in Fig. 2.23. The network represents a simplified version of the front-end circuit used in A1 pixels. The monostable circuit is used to produce a pulse signal, with a fixed duration (2 ns nominally), that is fed as input of the SiPM readout electronics. The enable procedure is managed through a circuit made of a couple of digital gates and a memory block. The latter, set by the EN signal, is connected to the quenching transistor and to the AND cell performing pulse gating. By skipping the enable procedure of specific subpixels, particularly noisy SPADs can be kept silent, therefore removing their contribution to the DCR of the SiPM.
The SiPM readout network, used to process the data generated by the 16 subpixels, is shown in Fig 2.24. The circuit consists of a parallel counter, used to compress the number of simultaneously fired SPADs into a 5-bit binary word, and a memory network, storing the data produced by the parallel counter. The readout network also implements functionalities like the generation of a SOT signal, the rejection of individual noise events and the filtering of residual count glitches. The different parts constituting the readout circuit are separately addressed in the following.

Parallel counter. Before delving into the description of the parallel counter integrated in the SiPMs of the ASAP110LF chip, some general concepts about this digital circuit will be discussed.
Parallel counters are combinational circuits providing a binary word, expressed on q bits, representing the number of concurrent 1's (signals at the high logic

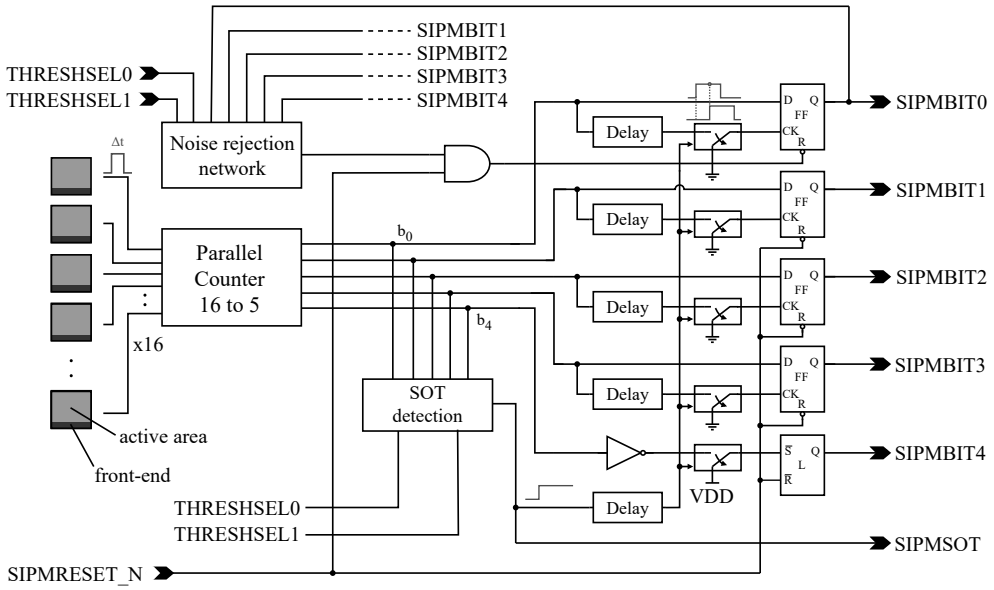


Figure 2.24: SiPM readout circuit.

level) occurring among the $p \leq 2^q - 1$ input signals [129][130]. Given the combinational nature of these circuits, the output result changes, after a relatively short propagation delay (or latency time), upon every variation of the input signals. Parallel counters are usually designed with the scope of effectively perform the sum operation between bits having the same binary weight. Ripple carry adders and carry lookahead networks can fulfil the same task of a parallel counter, even though with a non negligible expense in terms of delay and area occupation.

Digital compressors are the basic building blocks of parallel counters. A digital compressor (simply referred to as compressor in the following) is a circuit that can be used to collapse the information contained in equally weighted input bits into a more compact binary word. The ratio between the number of the input bits and the number of the output bits indicates the order of the compressor. As a matter of a fact, the parallel counter itself can be considered as a (p,q) compressor (intended as a compressor processing p input bits to produce a binary word on q bits), having higher order as compared to its building blocks. In the compression process, the information about the location of the triggered bits is lost. For example, in a (7,3) compressor, the output binary word 'b010 can be generated indiscriminately as a response to 'b0100100 or 'b0011000. Parallel counters can be implemented by using compressors of dif-

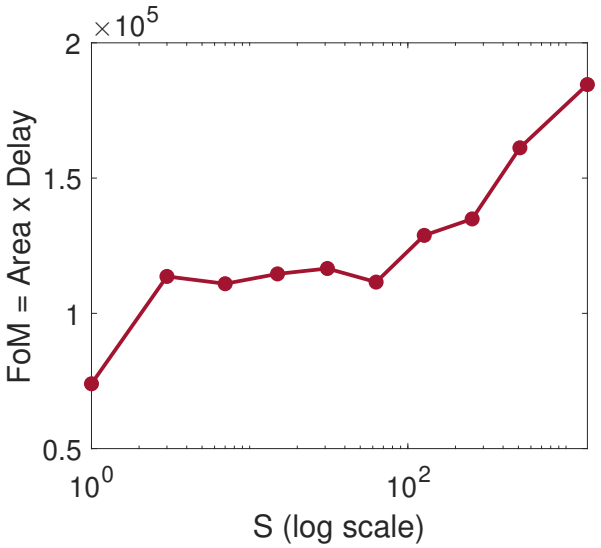


Figure 2.25: Area-delay figure of merit as a function of the number of sections (S) constituting a parallel counter with 4096-bit input.

ferent orders. If the optimization of the propagation delay is the main target of the design, a parallel counter made of only full adders often represents the best choice, even though negatively affecting the area occupation [131].
Depending on the number of the input signals and on the relative distance between the respective driver circuits, some implementations may require to split a parallel counter into equal smaller sections (S). Each section of the structure represents a local parallel counter itself, producing an output binary word consisting of a number of bits that decreases as S becomes larger. The binary words produced by the S local parallel counters are processed by a final network, that can be implemented as a usual binary adder circuit (ripple carry adder, carry lookahead) or, again, as a parallel counter (identified in the following of this section as "final parallel counter"), thus becoming the $(S+1)^{th}$ parallel counter of the entire structure. It can be worth specifying that the final parallel counter requires a different architecture as compared to the local ones, since its input bits have not equal binary weights. The parameter S deeply affects the performance, in terms of area and propagation delay, of a parallel counter. Fig. 2.25 shows the area-delay figure of merit (FOM) as a function of S, for a 4096-bit input structure. The FOM was calculated as the product between the area occupation of the circuit and the propagation delay. In the case of $S > 1$, corresponding to an actual splitting of the parallel

counter into smaller sections, the final sum of the binary output words produced by the small counters is performed with a final parallel counter. The best FOM is obtained when $S = 1$, thus indicating that a full parallel counter, without sectional division, features the best trade off in terms of area and delay. As the number of sections increases, the delay and area occupation follow opposite trends, with the former gradually decreasing. Before $S = 10^2$, the two parameters roughly compensate one each other. At higher values of S, the area occupation diverges, thus compromising the overall FOM. The maximum number of sections was set to No. of inputs/3 ($S_{MAX} = 4096/3 = 1365$), since the smallest local parallel counter possible, represented by the basic full adder, is reached when the number of inputs to each section is equal to 3 (this is the number of inputs of a full adder). Values higher than S_{MAX} would make little sense to the FOM calculation, as the employed full adders, each of them representing an individual parallel counter, would not be completely exploited.
From a conceptual point of view, a parallel counter can be considered as a three-dimensional structure made of a number of parallel planes, as shown

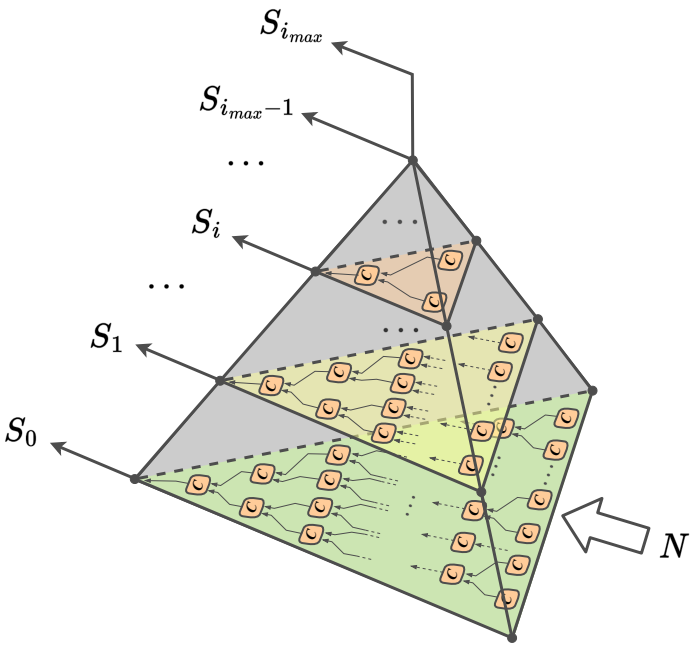


Figure 2.26: Three-dimensional structure representing the parallel counter internal division into different computing stages.

in Fig. 2.26. Each plane of the structure, representing a computing stage, contains all the basic compressors concurrently working to extract one of the binary weights of the resulting output word. The number of compressors included in each plane decreases while approaching the peak of the 3D-shape. The last plane of the structure, in charge of computing the highest binary weight, can be reduced to an actual point, since it is made of a single compressor. All the planes work in parallel to attain the output binary word, and the partial results produced in lower planes are fed as input of higher ones. The input bits, representing the information to be compressed, are provided only to the lowest plane. The intrinsic 3D nature of the parallel counters brings about significant advantages in terms of latency time, since the different binary weights are computed in a parallel fashion. The n^{th} computing stage starts working before the $(n-1)^{th}$ bit is ready, thus drastically reducing the time needed to generate a valid output word. However the time needed to produce a valid output data may vary depending on the configuration of the input bits. In addition, the implementation of a three-dimensional structure in a planar technology may result highly challenging, due to the high number of cross connections between different computing stages. This drawback represents one of the most limiting factor in the design of high order parallel counters. The routing of several signals through different computing stages may introduce non negligible mismatches in the internal delay of the counter, thus leading to the increase of the latency time and to the generation of count glitches.

The performance of parallel counters is strictly related to the number of the input bits. Fig. 2.27 shows the propagation delay and the area occupation featured by a parallel counter, designed by using only full adders, as a function of the number of input signals p. The data shown in the figure were extracted from a custom Matlab model, calculating the delay and the area of a parallel counter, starting from the number of input signals. The area occupation is expressed in terms of the total number of full adders required to build up the structure. The delay varies logarithmically with respect to the number of the input signals [132]. The delay function depends on the number of computing stages, which can only take integer values and increases for specific values of p i.e., when p is a power of 2, hence the logarithmic shape. The number of full adders, used to accomplish the compression procedure, varies linearly with the number of the input bits, as shown in Fig. 2.27b. However, according to the technology used to fabricate the structure, the intricate net of connections, which may take place in high order parallel counters, may induce to a non linear increase of the occupied area with respect to the number of compressors.

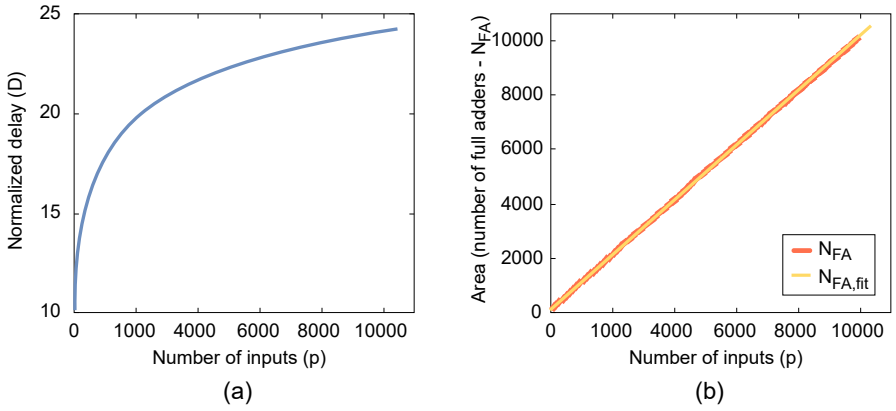


Figure 2.27: Performance of a parallel counter, as a function of the number of input signals [132]: (a) propagation delay, normalized with respect to the average input-to-output delay of a single full adder, (b) number of full adders, determining the total area occupied by the structure.

Especially in technologies featuring a low number of metal layers, the layout restrictions may require the basic compressors to be placed at a considerable distance from each other, in order to attain feasible connections between all the blocks.
Usually employed in digital multipliers, to perform the sum of partial products, parallel counters may be good candidates for the design of fast digital SiPMs. In applications involving the detection of photons with high temporal density, a relatively high number of cells may be triggered in a very small time window (hundreds of picoseconds). Parallel counters can be used as high-speed combinational networks, providing, with very low latency, the count result of simultaneously fired SPADs. A simplified scheme showing the integration of a parallel counter into a digital SiPM is depicted in Fig. 2.28. The output signals generated in the individual pixels represent the inputs of the parallel counter, which provides the number of simultaneously triggered sensors almost in real time. Besides the usual trade-offs associated with the design of electronic circuits, involving the optimization of parameters like area occupation and power consumption, the major challenge faced during the design of a parallel counter, intended to be used in non-clocked systems, is the generation of count glitches (or spurious pulses). Within the time interval between the assertion of the input signals and the production of a valid output result, a number of spurious pulses, with different voltage amplitudes, can be produced

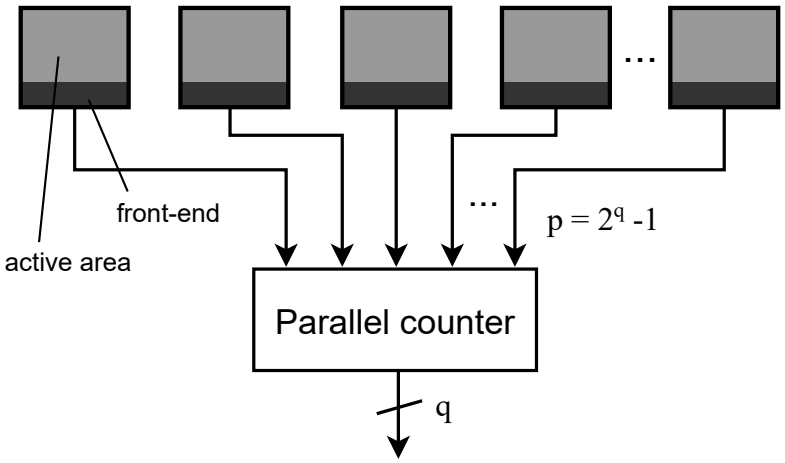


Figure 2.28: Integration of a parallel counter into a digital SiPM.

on the output bits of the parallel counter. The characteristics of the spurious pulses (voltage amplitude, time duration) are not affected by the duration of the digital signals that are fed as input to the parallel counter. Depending on the processing circuit used to store the compressed binary word, spurious pulses may cause false triggering in the readout networks. Count glitches can be generated due to the normal counting procedure taking place inside the parallel counter or as a consequence of mismatched delays between the internal paths. Mitigating the generation of counting glitches, by balancing the internal propagation delays of a parallel counter, helps reduce the power consumption and the occurrence of false triggering events, even though negatively affecting the time needed to produce valid output data. However, even though a good matching between the internal paths is obtained, some residual glitches can be generated due the regular evolution of signals within the structure. In this case, the amplitude of residual glitches should be kept, by design, well below the switching threshold of the subsequent readout electronics, so as to reduce the probability of false triggering events.
Conventional techniques for the design of digital SiPMs, providing information about the energy of the input radiation, exploit binary counters to estimate the total number of photons reaching the photodetector surface [98][119][133]. The output value of a binary counter is updated upon each positive edge of its input clock signal. In conventional architectures, the clock of the photon counter is directly connected to the output of an OR-tree, that is in charge of compressing and serializing all the SPAD pulses generated in the front-end

circuits of the relevant group of pixels. This approach poses several limitations to the rate of photons that can be successfully counted, since multiple photons occurring with not sufficient time delay between each other may cause a single triggering of the binary counter. In addition, even though separate pulses are generated by the OR-tree network, in response to different SPAD firings, the binary counter itself features a maximum counting frequency that, in the case of a 110 nm CIS technology (taken as example), can be around a few GHz. By using a parallel counter, each limitation on the photon temporal density becomes meaningless, since the SPAD pulses generated in the SiPM front-end circuits are processed in a parallel way. In principle, a parallel counter can efficiently replace both the OR-tree network and the binary counter of a conventional digital SiPM, since it has the capability of providing, with low latency time, the count value of simultaneously triggered cells. However, given the combinational nature of the parallel counter, dedicated registers need to be placed at the input or the output terminals of the circuit, so as to generate a registered output signal that can be acquired by the following readout network.

The parallel counter, designed for the SiPMs of the ASAP110LF chip, is shown in Fig. 2.29. The circuit, performing a (16,5) compression, processes the signals generated by the 16 sensing elements, to produce in real-time a binary word representing the number of concurrently triggered pulses. The counting result is conveyed through a 5-bit binary word which is computed after $6t_{FullAdder} + t_{OR}$ (where $t_{FullAdder}$ and t_{OR} are respectively the propagation delay of a full adder and of an OR gate). An overall propagation delay around 2.1 ns was obtained through post-layout simulations. The structure, based on the carry shower architecture [134], consists of 21 compressors, accomplished by using 20 full adders and an OR gate. In Fig. 2.29, the compressors are vertically aligned by reduction levels (RLs). At each reduction level, the number of internal signals is reduced by a factor equal to 1.5 (the ground signals are considered as actual signals for the sake of the reduction factor calculation). It can be worth specifying that the internal subdivision in reduction levels does not correspond to the identification of computing stages, which, instead, is a grouping criterion based on the binary weight produced by each compressor. In particular, full adders generating sums with the same binary weight constitute a computing stage (or plane). Fig. 2.30 shows the distribution of the compressors in the five different computing stages building up the structure. As suggested by the conceptual abstraction represented by the three-dimensional shape, the number of compressors involved in each computing stage decreases as approaching high level planes.

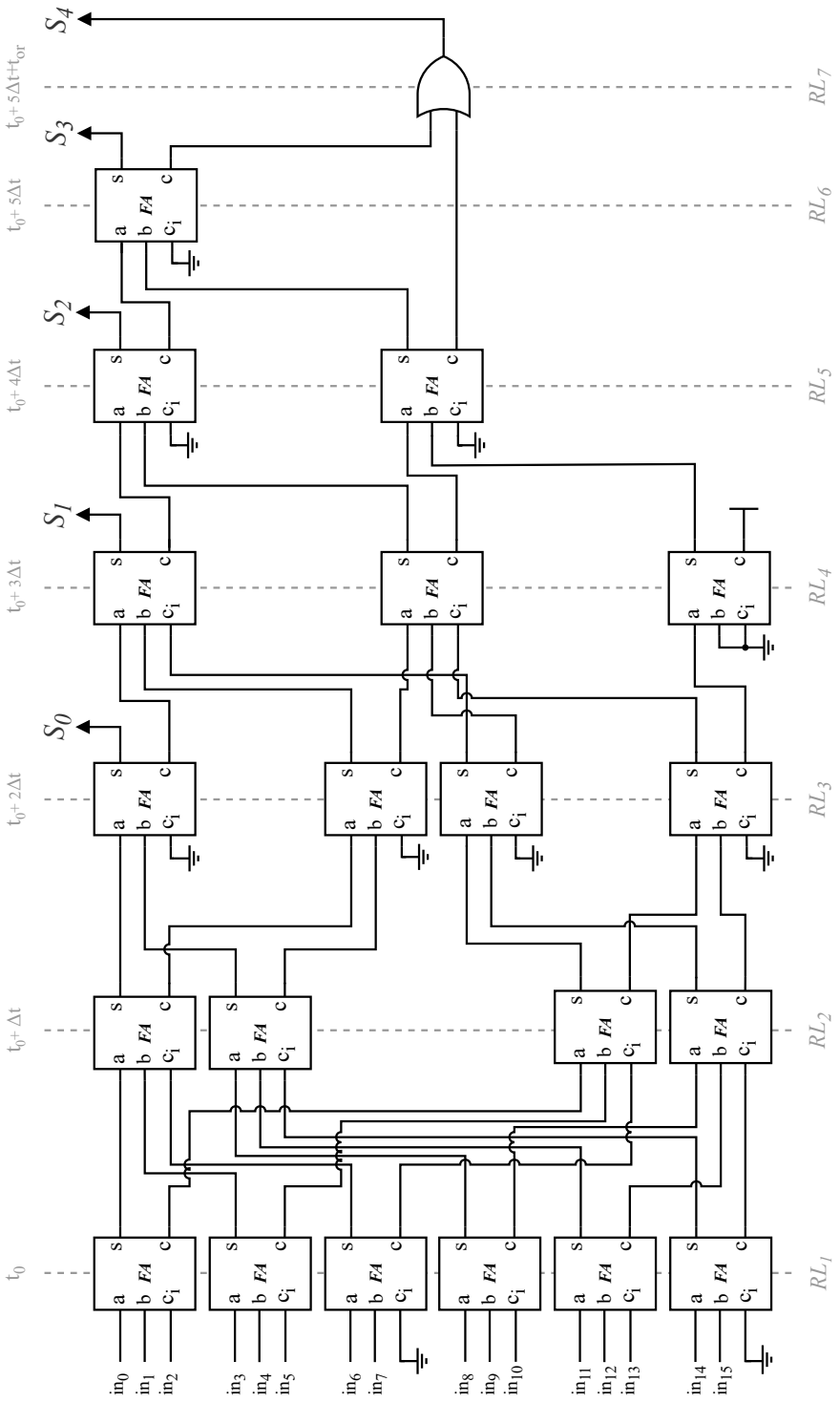


Figure 2.29: 16-bit input parallel counter integrated in the dSiPMs of the ASAP110LF chip.

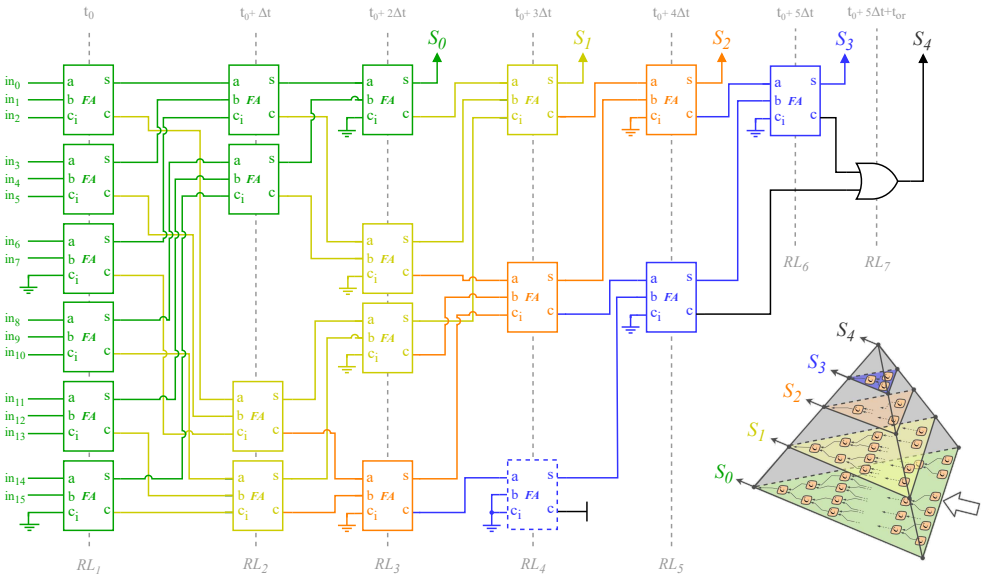


Figure 2.30: Subdivision of the 16-bit input parallel counter into the different computing stages making up the structure (colors representing the computing stages match with the ones used for the internal planes of the 3D shape).

The schematic of the counter was obtained through an automatic design algorithm [132], developed in the framework of the ASAP project. The algorithm, starting from the number of the input signals, generates the structure of a parallel counter in the form of a mathematical matrix, containing the information about the number of basic compressors and the connections between them. The scope of the algorithm, which is independent of the technology used, is to help the design of parallel counters featuring the lowest propagation delay possible, while keeping the generation of spurious pulses under control. Given the number of the input signals, the algorithm generates a single parallel counter structure. To fulfil the requirements in terms of latency and glitch generation, the algorithm was programmed to generate parallel counters based on the carry shower architecture. According to this architecture, all the sum and carry-out signals generated by each reduction level have the same distance, in terms of number of compressors, from the first reduction level. To reduce the occurrence of spurious pulses on the output bits, all the internal signals are prevented from skipping reduction levels. This requirement is attained by adding dummy full adders (like the one identified with a

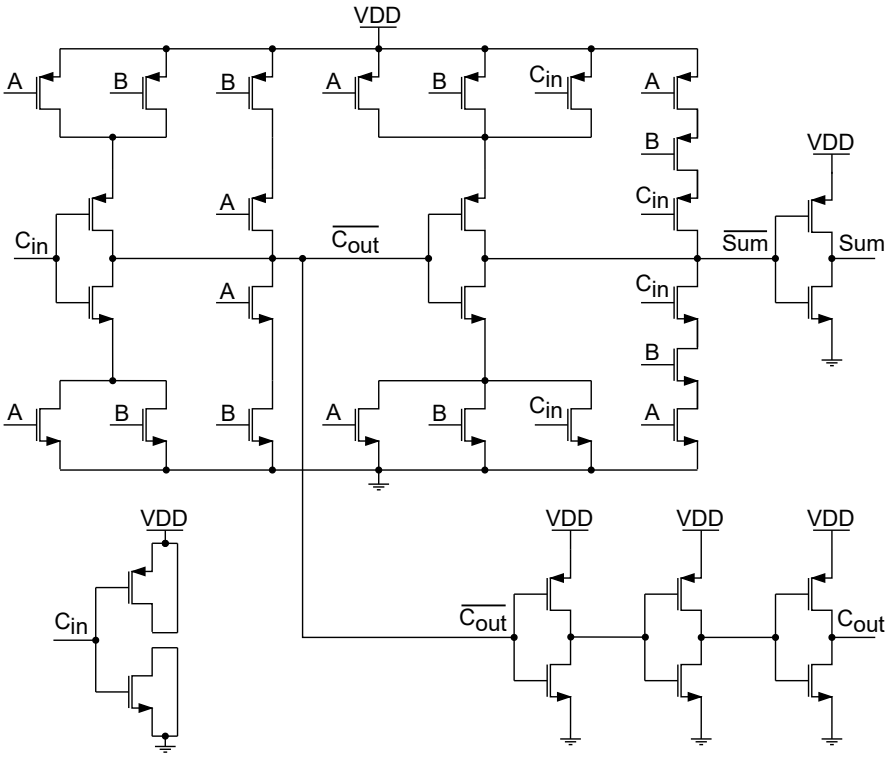


Figure 2.31: Schematic diagram of the custom full adder integrated within the parallel counters of ASIPM.

dashed outline in Fig. 2.30), blocking the propagation of signals between non adjacent RLs. Since the area saving is not the primary target, half adders were never used in place of full adders processing less than three input signals. Only the full adder belonging to the highest computing stage, poorly affecting the glitch generation, was manually replaced with an OR gate, preserving the logic function of the network.

A further compensation of the propagation delays taking place inside the parallel counter was accomplished through the development of a custom full adder. The schematic diagram of this block is shown in Fig. 2.31. The design of the new cell started from the mirrored full adder architecture, which intrinsically features relatively small area occupation and symmetric rising and falling edges of the output signals. In addition to the fundamental circuit, a network made of buffers and MOSFET capacitors was implemented, so as to equalize the input-to-output delays. In the resulting compressor, referred to as

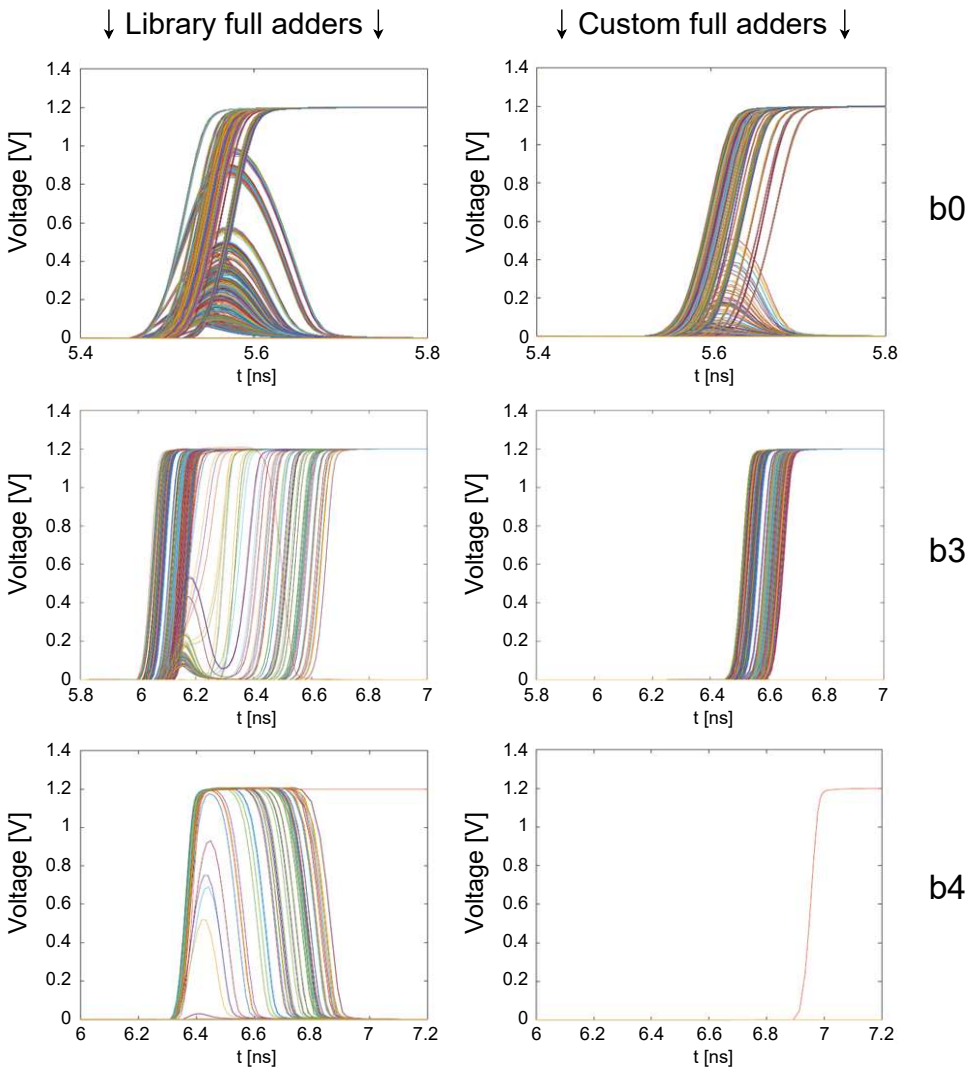


Figure 2.32: Count glitch performance of two parallel counters, separately integrating library full adders and custom full adders.

balanced symmetric full adder, the Input-to-Sum and Input-to-Cout propagation delays, as well as the 0-to-1 and the 1-to-0 transition times, are almost equal, thus strongly reducing the full adder contribution to the parallel counter propagation mismatches. In the final structure, all the signals processed by a specific reduction level experience the same propagation delay, regardless of the polarity of the signals or the relative binary weight. The main drawbacks

associated with the use of balanced symmetric full adders, in place of library ones, are concerned with the area occupation and the absolute propagation delay. The layout of a custom full adder, developed in 110 nm CIS technology, turned out to be 30% larger than the layout of a library one. The additional area required by parallel counters integrating custom full adders may represent a real challenge for the design of complex detection systems. Moreover, in order to obtain balanced delays between the input and the output ports of a custom full adder, the average propagation delay increased by 10%.
The glitch performance of two parallel counters, separately employing library full adders and custom full adders is shown in Fig. 2.32. These data were obtained after schematic simulations performed on structures based on the circuit diagram depicted in Fig 2.29. To extract the curves shown in the figure, the parallel counters were stimulated with a relatively large set of input configurations. The use of balanced symmetric full adders proved remarkably effective in mitigating both the quantity and the energy of count glitches, which were totally removed for some output bits. In addition, due to the use of balanced symmetric full adders, the distribution of the triggering times was found considerably narrower around the corresponding mean value, which, however, is slightly shifted a few hundreds of picoseconds afterwards.

Memory network. When a detection event is triggered, the binary word representing the number of simultaneously fired SPADs remains available at the output of the parallel counter for a specific time window, referred to as active window. Each output bit of the parallel counter features a different active window. However, the relative difference between the active windows of the five output bits does not exceed few hundreds of picoseconds. The active window depends mainly on the duration of the signals generated by the front-end circuits and, only marginally, on the number of fired cells as well as on their relative position in the SiPM. Depending on which input bit is set, different paths, with slightly different propagation times, can lead to the assertion of the n^{th} output signal. Typical values of active window, extracted through post-layout simulations for different process corners, may range between 2 and 5 ns. Within the active window, the network built around the parallel counter (Fig 2.24) automatically stores the five bits of the count result. Four positive edge-triggered flip-flops and an SR latch are used as memory elements. In order to suppress the effect of few residual glitches, which may sporadically occur even in the case of a parallel counter with balanced internal paths, the storing operation is triggered automatically once a valid data is generated. As shown in Fig 2.33, the auto-triggering mechanism, implemented on the first

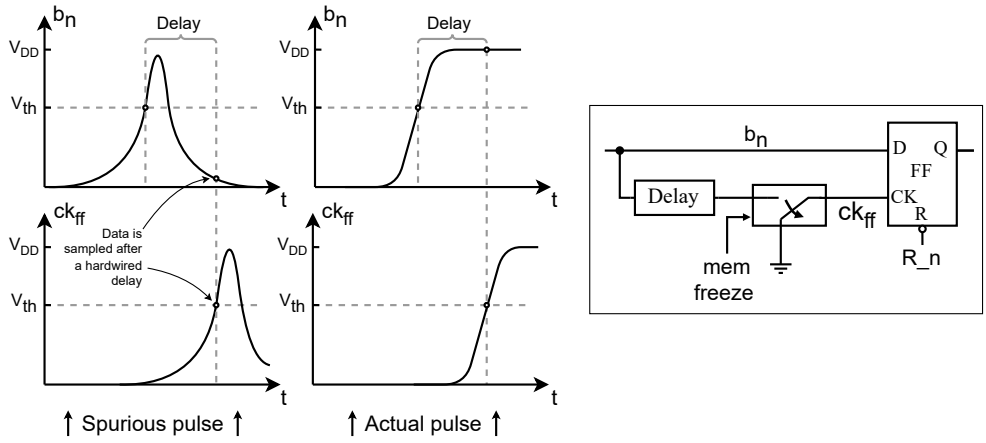


Figure 2.33: Auto-triggering mechanism filtering out residual count glitches generated by the parallel counter.

four memory blocks, consists of a digital delay line, connecting the D and CK input ports of each flip-flop. Only actual count pulses, with a duration equal to the active window, can be stored within the memory blocks. Count glitches, spending few hundreds of picoseconds above the digital threshold of the subsequent logic gates (the ones in the input stage of the memory element), are filtered out by the auto-triggering network, since the output signals of the parallel counter are sampled after a fixed time with respect to the rising edge of the spurious pulse. It is worth specifying that glitches with an amplitude smaller than the memory switching voltage are not detected by the readout circuit, thus not representing a significant problem. The propagation time of the delay blocks was carefully selected (800 *ps* nominally), after post-layout simulations in different process corners, to perform an effective filtering operation of all the potential glitches produced by the parallel counter. The choice of the delay block is crucial for the effectiveness of the structure. If the delay time is not sufficiently large, energetic glitches, with amplitude exceeding the following digital threshold, may be sampled by the memory elements. On the opposite, delay values larger than the active window may lead to data loss, since the output of the parallel counter may be reset before the memory is triggered.

As shown in Fig 2.24, the flip-flop storing the MSB has been replaced with an SR latch, since the output b_4 of the parallel counter was found to be completely glitch-free from post layout simulations.

The threshold circuit sets the SOT bit if the number of concurrent firing

SPADs is greater than or equal to a determined value, which can be selected through the THRESHSEL0 and THRESHSEL1 configuration bits. The threshold values, as a function of the two configuration signals, are listed in Table 2.7. The threshold bit, calculated starting from the 5 bit word of the parallel counter, triggers the activation of five switches, disconnecting the CK port of the flip flops (and the S port of the latch) from the output of the delay blocks. Therefore, upon the assertion of the threshold bit, the memory elements become insensitive to any other signal produced by the parallel counter. A reset signal must be provided to the circuit in order to start a new acquisition. The SOT signal is available on the SIPMSOT output pad.
The SiPM electronics is equipped with a self reset circuit, which is used to get rid of individual contributions of dark noise. A combinational circuit, referred to as noise-rejection feedback network (NRFN), is connected to the output bits of the five memory elements. The flip-flop storing the LSB is automatically reset, if the binary word 'b00001 is read at the output of the readout network. The noise rejection feedback network is disabled when a threshold equal to 1 is selected, so as to allow the recording of single events.
The time diagram describing a typical circuit operation is shown in Fig 2.34. The waveforms have been obtained from post-layout simulations, with nominal process corner and a supply voltage of 1.2 V. Three scenarios, each involving a different number of concurrently fired SPADs, have been reproduced in simulation, by directly driving the input ports of the parallel counter. A SiPM threshold equal to 2 was used during the simulation leading to the results shown in the figure. The first event, generated after the rising edge of the reset signal, occurs at 4 ns. In this case, the assertion of a single input signal, randomly chosen among the 16 input ports of the parallel counter, is simulated. Upon the rising edge of the input bit, the parallel counter initiates the counting process, which is concluded after a time delay of about 1 ns. The

THRESHSEL1	THRESHSEL0	threshold value
0	0	1
0	1	2
1	0	3
1	1	4

Table 2.7: SiPM threshold values as a function of the THRESHSELx configuration bits.

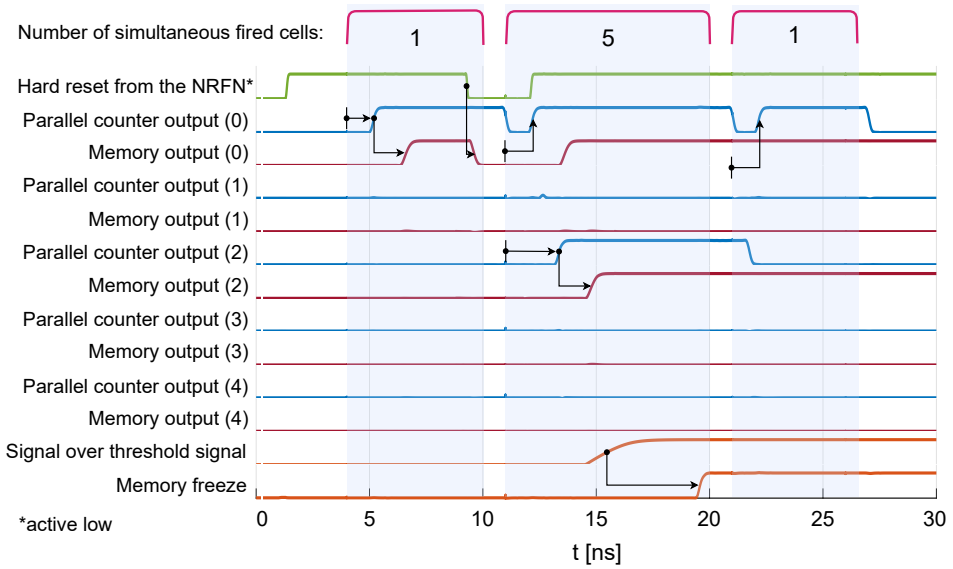


Figure 2.34: Time diagram showing a typical circuit operation of a SiPM integrated in the ASAP110LF chip.

content of the memory blocks is updated with the new count result ('b00001) after a determined time window depending on the latency time featured by the auto-triggering network. Since a threshold value larger than 1 is selected, the noise-rejection feedback network is triggered, thus resetting the output of the flip-flop storing the LSB. Given the threshold value, the SOT bit is not asserted, and the circuit is ready for the acquisition of the next detection event. In the second event, a five-photon detection is simulated. For the sake of simplicity, the five input signals were set at the same time. As the binary word corresponding to decimal 5 ('b00101) is provided at the output of the parallel counter, the signal over threshold bit is asserted. After the storing operation, the output of the memory elements is frozen, so as to prevent the readout network from updating the output result. All the events following the assertion of the SOT signal are ignored by the circuit, which requires a manual reset to allow the acquisition of a new detection event.

2.3.6.2 SiPM electronics - alternative design

The SiPMs located in row 5 of ASIPM are equipped with the alternative readout circuit shown in Fig. 2.35. In addition to the circuit depicted in Fig 2.23, the subpixel front-end circuit is provided with a memory block storing

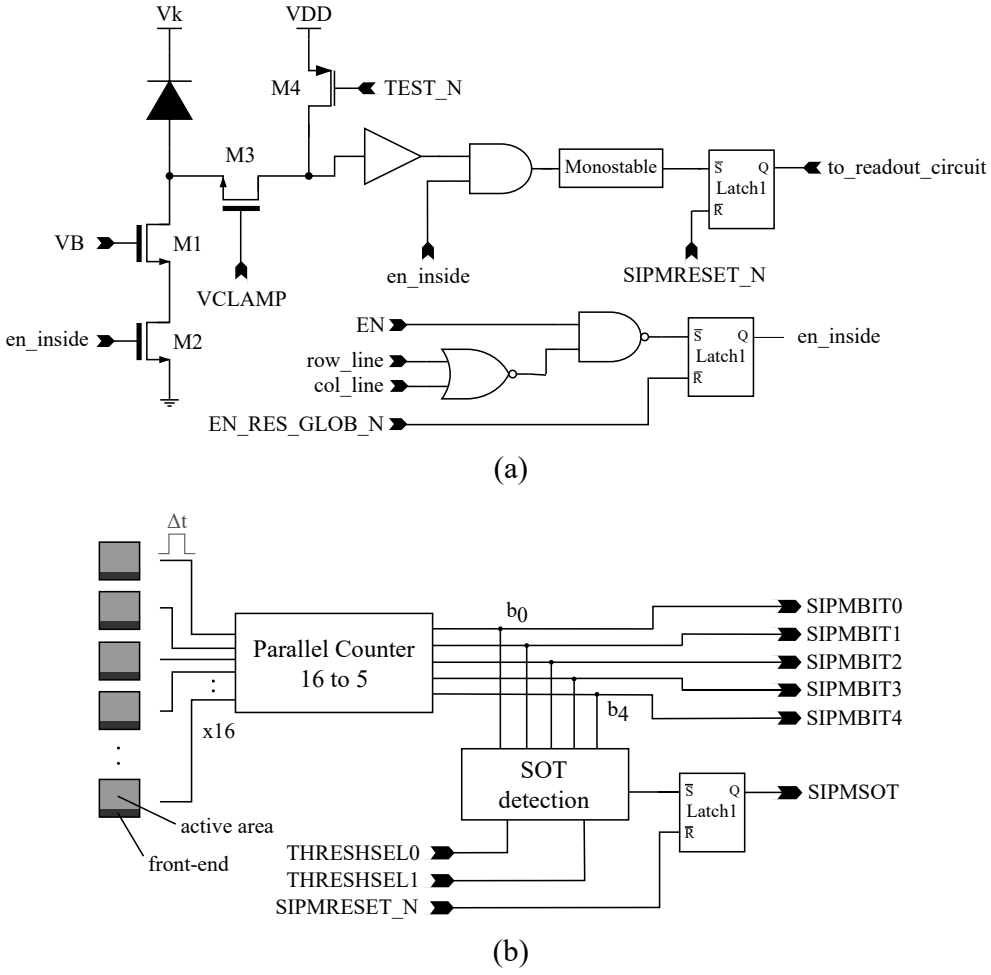


Figure 2.35: Alternative SiPM readout network employed in the SiPMs of the last row.

the bit information relevant to the SPAD firing. The signals generated by the subpixels are processed by the same parallel counter as in Fig. 2.29, which, in this case, is directly connected to the output pads. The active window of these SiPMs is potentially unbounded, since the bits fed as input of the parallel counter are latched. A SOT flag is also generated by the SiPMs of this row. Between adjacent acquisitions, a SIPMRESET_N pulse must be provided, to accomplish the reset of the subpixel memories and of the SOT signal.

2.3.7 Array SEPARATED-STI (ASSTI)

The array SEPARATED-STI consists of 2×8 cells, located in the gap between the pad ring and the top side of A1. Pixels in ASSTI have the same dimension and readout electronics as cells in A3. The 16 sensors integrated within this array are based on the SPAD structure indicated as SPAD 2 in Fig. 2.2. This array has been designed with the scope of studying the noise performance of SPAD 2 devices in the frame of a matrix structure. For the row selection procedure, ASSTI is not provided with a dedicated row selection shift register. The active row can be selected by setting the ROWSTI/SPADEN input signal, according to the values shown in Table 2.8. The column selection is carried out through the column lines of A1, as shown in Fig. 2.36. The array is provided with a single output pad (OUTSTIMOD), which is connected to the pixels through a combinational circuit similar to the one shown in Fig. 2.12b. The signal generated by the in-pixel monostable and the data stored in the first latch of the readout network are not available on the output pads. The cathode voltage is supplied through the VK5 input pad, which is specific for this array.

2.3.8 Array PIXEL-TEST (APXT)

The PIXEL-TEST array is made of four pixels, arranged in a single row. This array has been designed to support the measurement of specific sensor parameters, like the SPAD timing resolution and the SPAD pulse duration. To this scope, the number of components integrated within the readout circuits

		AMSTI	ACHT
ROWSTI/SPADEN	0	up row selected	disabled
	1	down row selected	enabled

Table 2.8: Effect of the ROWSTI/SPADEN signal on ASSTI and APXT.

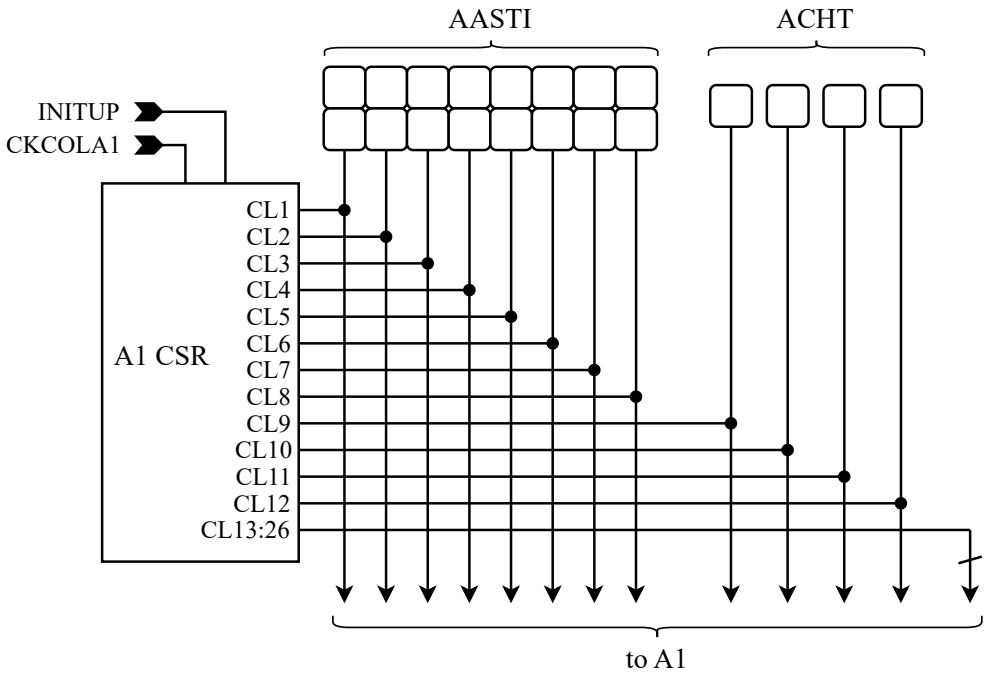


Figure 2.36: Column selection mechanism for pixels in ASSTI and APXT.

was reduced to a minimum, so as to negligibly affect the detection pulse timing jitter. Unlike the other arrays in the core, SPADs in APXT does not share the same pwell. A gap of 40 μm was added between adjacent pixels, representing independent structures that were grouped in a small array only for signal distribution purposes. The single row line can be activated by setting the ROWSTI/SPADEN input signal according to the values shown in Table 2.8. The columns are selected through the A1 column shift register, according to the diagram shown in Fig. 2.36. To avoid the use of combinational circuits performing the routing of the output signals, each pixel was equipped with a dedicated output pad (OUTSPADTESTx, with x = 0, 1, 2, 3). All the SPADs within APXT are connected to the VK1 cathode voltage. The readout circuits implemented in three pixels of APXT is shown in Fig 2.37. The SPADs contained in these cells have the same properties of A1 sensors, including the size of the active area. The electronics consists of different versions of the front-end circuit used in pixels of A1. In particular, for the first two pixels, the monostable circuit was completely removed, so as to reduce the number of logic gates potentially affecting the timing performance of the signal generated

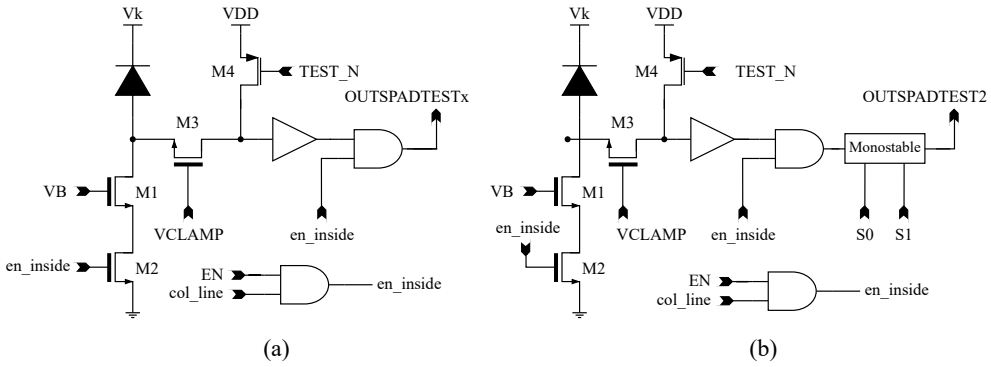


Figure 2.37: Readout circuits of three pixels in APXT: (a) pixels 1 and 2, (b) pixel 3.

by the sensors. The third pixel can be used to measure the actual duration of the monostable pulse in all the four configurations. Eventually, the remaining pixel, which is smaller if compared to the other ones, contains a copy of an ASIPM subpixel.
It is worth specifying that the output of each pixel is connected to the relevant output pad through a digital inverter, not shown in the figure.

2.3.9 Time to digital converter (TDC)

The ASAP110LF chip is equipped with an internal 10-bit continuous mode/20-bit time to digital converter. This structure can be programmed to continuously acquire and convert adjacent time frames between subsequent pulses (10-bit continuous mode) or individual frames separated by a readout time window (20-bit mode). The TDC can be used to measure the time intervals between adjacent pulses, taking place in the SPADs of the core arrays. This piece of information can be used to evaluate the after-pulsing features of SPADs or as an alternative way to measure DCR. The time information is converted into a binary word, that can be serially read out on a dedicated output pad (TDCOUT). The input of the TDC can be internally connected to the monostable outputs of A1, A3 and AAQ. Through the TDCIN4 input pad, an off-chip source can be fed to the TDC. Even though pulses may be simultaneously generated in multiple input sources, a single signal at a time can be processed by the circuit. The time to digital conversion is achieved through binary counters, which are controlled by sequential networks managing the count evolution. The count clock can be selected between internal or external signals, according to a dedicated configuration bit. The reading op-

eration, which is performed serially, is carried out by using an external clock, that can be provided on the TDCCKSHIFT input pad. Depending on the selected operating mode, specific time windows are allocated to the reading procedure. The TDCMODE signal, generated by the TDC, and available on an output pad, can be used to trigger the external acquisition system.
With an on board TDC, the DCR of single SPADs can be measured by extracting the average value of the time intervals between adjacent dark pulses. With this method, precise DCR measurements, getting rid of the pile-up error (described in section 3.2.4.1), can be performed. The analysis of the time intervals occurring between dark pulses may also reveal the presence of afterpulsing phenomena taking place in the junction.
The electronics making up the TDC can be divided into three different blocks, which will be separately described in the following: the pulse splitting network, the TDC core structure and the clock generation circuit.

Pulse splitting network. The pulse splitting network, shown in Fig 2.38, manipulates the input pulses occurring in the selected source, in order to generate the suitable control signals that are used to trigger the TDC core structures. The input signal to be processed can be selected through the TDCSRCSEL0 and TDCSRCSEL1 bits, according to the values in Table 2.9. The output of the source selection multiplexer is gated by the TDCLOCK signal, which can be used, during the TDC reading procedure, to freeze the configuration of the TDC internal structures. When the TDCLOCK signal is asserted, the pulses produced by the selected source are prevented from

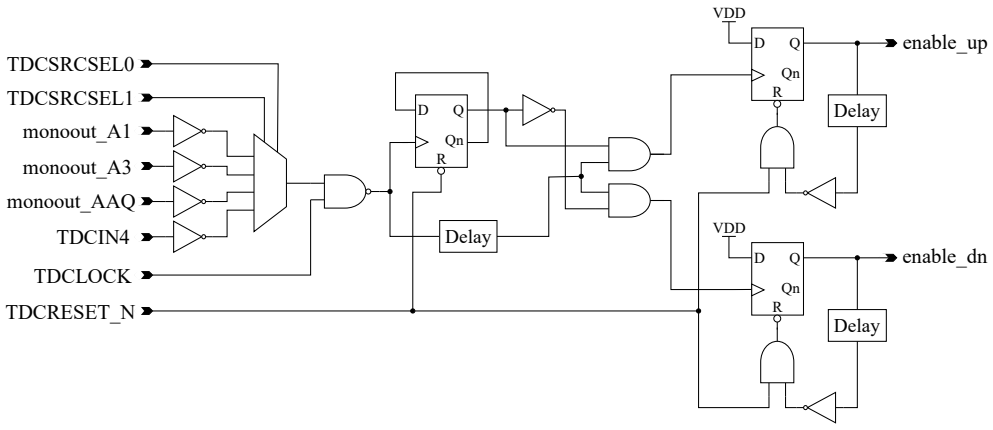


Figure 2.38: Pulse splitting network of the in-chip time to digital converter.

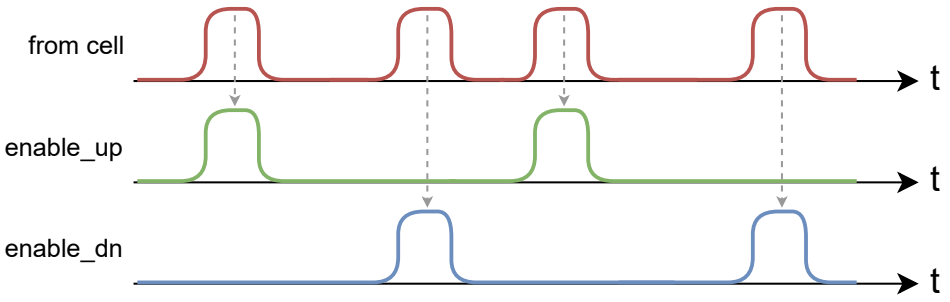


Figure 2.39: Signals generated by the pulse splitting network (the waveforms shown here are only a schematic representation and were not obtained from simulations or actual measurements).

TDCSRCSEL1	TDCSRCSEL0	input source
0	0	A1
0	1	A3
1	0	AAQ
1	1	TDCIN4 (external)

Table 2.9: Input source selection for the TDC circuit.

reaching the TDC core, thus not affecting the current state of the core counters. After the pulse gating block, the input pulses are splitted on two different lines by means of a circuit made of a toggle flip flop and a simple combinational network. Upon each pulse, the D flip-flop enables, alternately, one of the two following AND gates. The delay block, placed between the NAND cell and the AND gates, was designed with a propagation delay larger than $t_{CKQ} + t_{INVERTER}$. Under these assumption, adjacent pulses are splitted on two different nets, as required by the TDC core circuits. Each line is equipped with an auto-reset sequential circuit, setting a standard duration to the pulses reaching the TDC core (*enable_up* and *enable_dn*). A time diagram, showing the waveforms of the signals generated by the pulse splitting network, as a function of the source signal, is shown in Fig. 2.39.

TDC core structure. The circuit diagram of the TDC core is shown in Fig. 2.40. The structure consists of two 10-bit synchronous binary counters, which can work in interleaved mode or as a single 20-bit structure. The con-

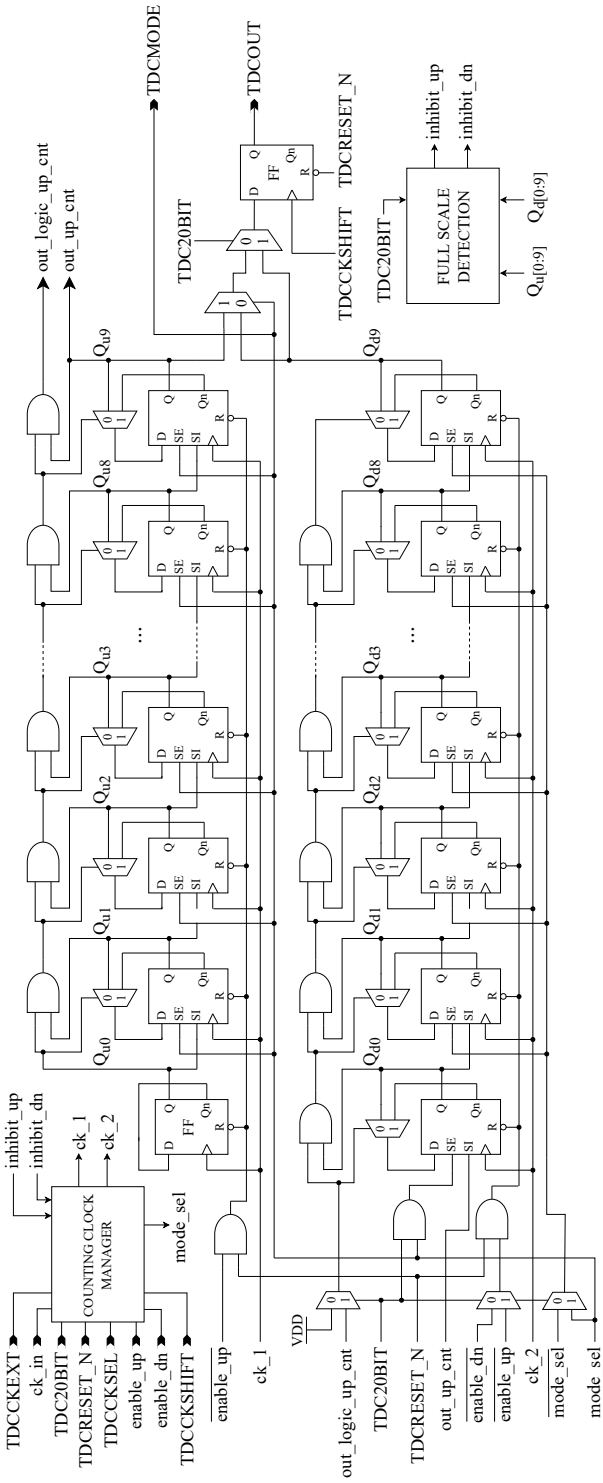


Figure 2.40: Circuit diagram of the TDC core.

version of the time interval into a binary word is performed by counting the number of clock periods occurring between two adjacent pulses of the input signal. The clock for the counting operation can be provided by an external source through the TDCCKEXT input pad. Alternatively, an internal clock signal, generated by the clock generation circuit, can be used for the counting operation. The selection between internal and external clock source can be performed through the TDCCKSEL configuration bit, according to the values in Table 2.10. The basic block of the counter circuits is represented by the scan flip flop (SFF), which, unlike to usual D flip-flops, is provided with two extra input ports (Scan-In, SI, and Scan-En, SE). Depending on the logic value at the SE port, the Q output of the SFF is updated, upon the positive edge of the reference clock, with either the logic value at the D or SI port. The scan flip-flops, which are usually employed for DFT (design for testability) purposes, are used here to easily rearrange the TDC architecture during the normal circuit operation.
Based on the TDC20BIT configuration signal, the TDC can work in two different operating modes:

- 10-bit continuous-time mode. The counting operation is alternately performed on both the 10-bit counters. During the time interval between two adjacent pulses, one of the 10-bit counters works properly, converting the time information into the digital domain. Concurrently, the other set of ten SFFs is set as a 10-bit shift register, so as to allow the serial reading of the binary word, that was acquired in the previous time frame. Upon every pulse of the input signal, the two 10-bit structures switch operating mode, alternately changing from counters to shift registers. In the 10-bit continuous-time mode, all the time intervals are measured and converted into a binary word. The counting result, achieved by one of the two counters, is available for reading when the same counter switches to the shift register configuration. Each binary word is generated and sent to the output pad by the same ten SFFs, dynamically rearranging their inter-block connections. In the 10-bit continuous-time mode, the entire time span is recorded, as long as pulses are generated by the selected input source. If a counter reaches the full scale value, a clock-gating network, inhibiting the clock distribution, prevents the circuit from going into overflow, thus freezing the result. In order to save area and reduce the activation time of the inhibition network, the full scale value was set to 1020.

- 20-bit mode. The counting operation is performed through a single 20-bit counter, which is obtained by connecting in cascade the two 10-bit

TDCCKSEL	0	internal clock
	1	external clock (TDCCKEXT)
TDC20BIT	0	10-bit continuous-time mode
	1	20-bit alternate mode

Table 2.10: Different configuration modes of the TDC.

structures. After the measurement of a time interval, the 20-bit counter is rearranged as a shift register, thus making the count result, that was acquired during the previous time frame, available to the output pad. When all the SFFs work as a shift register, no structure accomplishing the count operation is available. As a result, when the TDC is in the shift register configuration, no time interval measurement can be performed. The counting operation can start again upon the next input pulse. The full scale value of the 20 bit counter is set to $2^{20} - 8191 = 1040385$. A clock-gating network inhibits the counting operation, if the full scale value is reached. In the 20-bit mode, the variety of time intervals that can be acquired without saturating the counting capabilities is much larger. Time intervals with remarkable duration can be measured even by using high frequency clocks, thus significantly improving the dynamic range of the TDC. However, as already mentioned, the entire time span cannot be continuously acquired, since, after the time to digital conversion, a time frame must be dedicated to the reading operation.

The TDC operating mode is managed through the TDC20BIT configuration signal, according to the values in Table 2.10. In both the operating modes, the switch between the counting configuration and the shift register arrangement is triggered by the *enable_up* and *enable_dn* signals, which are generated by the pulse splitting network. These two signals are used to provide the start and stop prompts to the counters, thus defining the integration time. The *enable_up* and *enable_dn* signals are used also to reset the counters before the counting operation starts. The arrangement of the two 10-bit structures and the counting operation are managed by a sequential circuit, generating the count clocks of the two counters (*ck_1* and *ck_2*) and the *mode_sel* bit, that governs the transition between the counter and shift register configurations. The *mode_sel* signal is available on an output pad (TDCMODE). A summary of the TDC arrangement, as a function of the TDCMODE signal, is shown in Table 2.11.
The readout of the binary word is performed serially on the TDCOUT output

		TDC20BIT	
		0	1
TDCMODE	0	counting on $Q_{up[0:9]}$ reading $Q_{dn[0:9]}$	counting on 20 bits
	1	reading $Q_{up[0:9]}$ counting on $Q_{dn[0:9]}$	reading 20 bits

Table 2.11: TDC arrangement, as a function of the TDCMODE signal, in the 10-bit and 20-bit operating modes.

pad. The bits of the output binary word are serially sent to the output pad upon the rising edge of the TDCCKSHIFT clock signal. A number of clock pulses equal to the number of bits that have to be extracted must be provided. Data exit the pad starting from the MSB.
Fig 2.41 shows the waveforms involved in the reading procedure. Two different approaches can be used, depending on the application. If the TDCLOCK signal is not used (Fig 2.41a), the TDCMODE signal can be observed to manage the reading procedure of the 10-bit structures. During a time window occurring between adjacent pulses, ten rising edges of the TDCCKSHIFT signal must be provided to successfully extract all the bits stored within the 10-bit structure currently arranged as shift register. The maximum frequency of the reading clock that is supported by the register is 100 MHz. If an input pulse occurs before the reading operation is complete, all the LSBs which were not extracted cannot be retrieved, since the shift register toggles into the counter configuration, thus disconnecting from the output flip-flop. The probability of loosing some LSBs depends on the frequency of the input pulses and on the frequency of the clock which is used for the readout operation. The occurring of a new input pulse, causing the counter to switch its operating mode, is reported by the TDCMODE signal, that can be used to check the partial validity of the read out word.
The internal configuration of the TDC can be locked, during all the operating phases, through the TDCLOCK signal. As shown in Fig 2.41b, as long as the TDCLOCK is asserted, the counter configuration is not affected by the upcoming input pulses, which are ignored. By applying the lock condition, the reading window can be manually defined, preventing any kind of bit loss from happening. However, in this reading mode, some time intervals between avalanches may be lost, since all the time frames initiated with an ignored

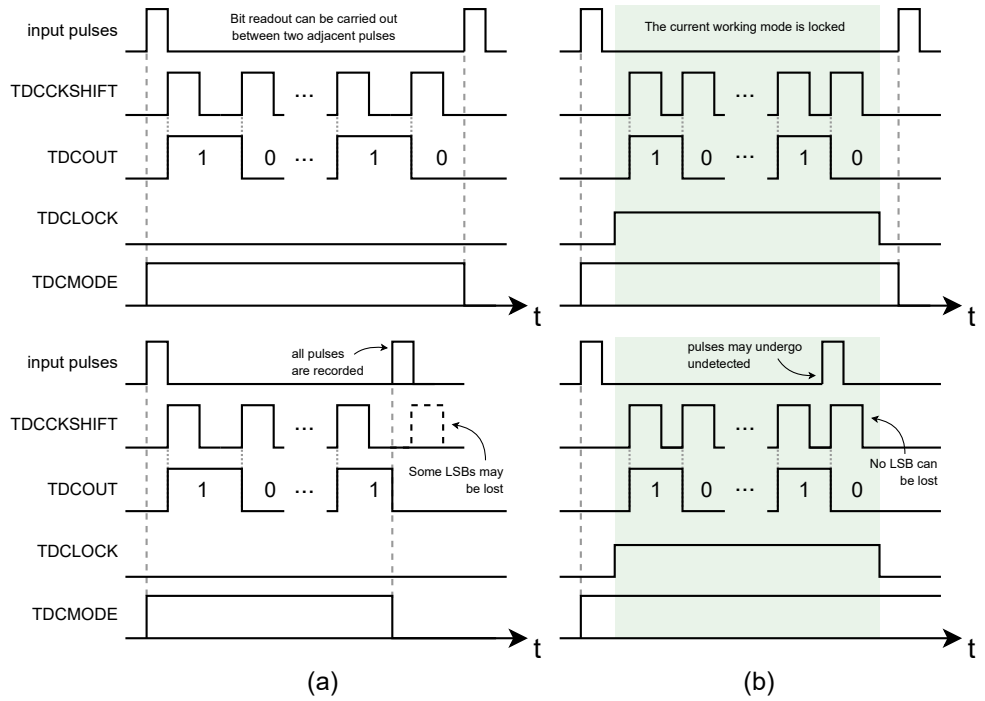


Figure 2.41: Reading procedure: (a) the TDCMODE signal is used to trigger the external reading system, (b) the TDCLOCK signal is used to define a fixed reading window.

pulse cannot be processed by the TDC. The first input pulse following the falling edge of the TDCLOCK signal restarts the time frame acquisition. The amount of time intervals which cannot be converted by the TDC depends on the frequency of the input pulses and on the duration of the TDCLOCK signal.

Clock generation circuit. The circuit providing the internal clock to the TDC is shown in Fig. 2.42. The internal clock is generated by means of a ring oscillator, exploiting the intrinsic propagation delay featured by logic gates. If the internal clock source is not used, the oscillation can be inhibited by setting the OSCEN signal to the low logic value, so as to reduce the power consumption. The nominal frequency of the clock signal generated by the ring oscillator is 500 MHz. Due to critical paths not supporting such a clock frequency in different process corners, the signal generated by the oscillator is not directly applied to the TDC. Before the clock reaches the TDC, the oscillation frequency is reduced through a programmable frequency divider.

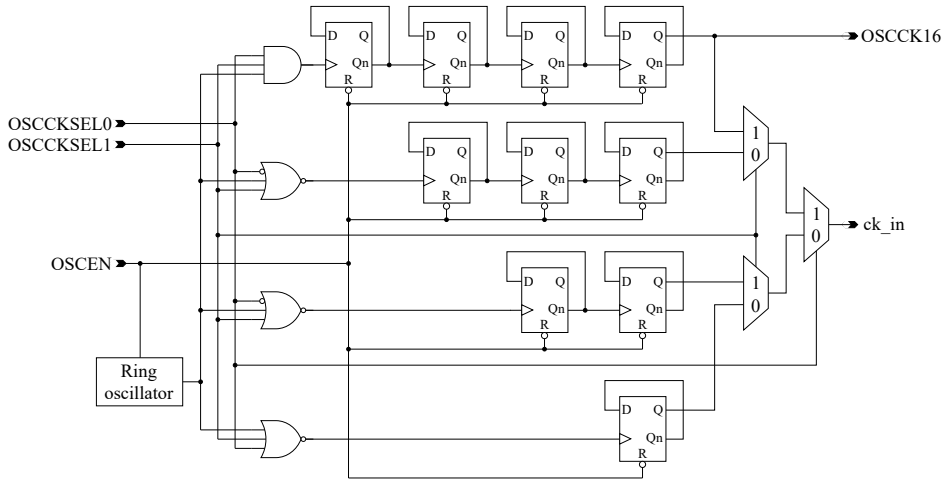


Figure 2.42: Clock generation circuit producing the internal clock provided to the TDC.

As shown in Table 2.12, four different ratios can be selected through the OSCCKSEL1 and OSCCKSEL0 configuration signals. As the frequency divider is based on binary counters, the resulting frequency is obtained through power-of-2 integer division. The clock signal with the lowest frequency ($f_{osc}/16$) is made available on the OSCCK16 output pad. This test point can be used to measure the actual oscillation frequency, which may depart from the nominal value due to process variations. The highest internal frequency that can be provided to the TDC is 250 MHz, thus achieving a maximum resolution of 4 ns.

OSCCKSEL1	OSCCKSEL0	ck_in frequency
0	0	$f_{osc}/2$
1	0	$f_{osc}/4$
0	1	$f_{osc}/8$
1	1	$f_{osc}/16$

Table 2.12: Clock frequencies provided to the TDC as a function of the OSCCKSEL0 and OSCCKSEL1 configuration bits ($f_{OSC} = 500\ MHz$).

Chapter 3

Measurement setup and characterization results

3.1 Measurement setup

To accomplish the characterization of the structures in the ASAP110LF chip, an FPGA based measurement setup has been designed. The core of the system, which is schematically represented in Fig. 3.1, consists of an FPGA, used to generate the digital signals needed to control the circuits in the device under test (DUT), an USART-USB digital interface, transmitting data from and to the FPGA, two custom boards, connecting the DUT to the measurement setup, and a personal computer, which performs data processing. For specific measurements, the core setup depicted in the figure has been extended with additional equipment like a climatic chamber and a position controlled semiconductor laser. Three types of signals, represented with different colors in Fig. 3.1, travel through the components of the measurement setup, as explained in the following:

- Stimuli represent a set of digital signals controlling the circuits inside the ASAP110LF chip. These data, basically consisting of sequences of pulses fed to the chip, are produced by the FPGA, which has been set in order to have dedicated IO pins for the stimuli generation.

- Communication signals, consisting of data packets made of at least one byte (8 bits), are generated by the PC to communicate instructions to the power supply or the FPGA. A communication packet may contain specific information to manage the different steps of a measurement, or a set of instructions to change the voltage provided by the power supply

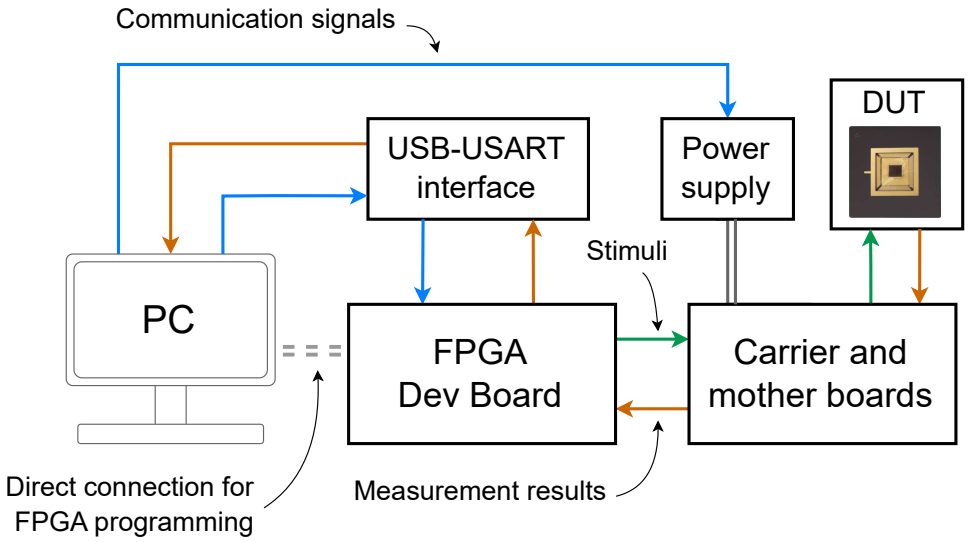


Figure 3.1: The core of the measurement system which was developed for the characterization of the ASAP110LF chip.

to the DUT. Through communication signals, the computer can control also external instrumentation (e.g., a climatic chamber, as already mentioned), enlarging the functionality of the core measurement setup.

- Measurement data are the product of a measurement procedure. Digital signals are made available at the output pads of the ASAP110LF chip as a result of a measurement phase. The FPGA collects all the digital bits produced by the DUT, and performs a preliminary step of signal processing. The measurement data, after being managed by the FPGA, are eventually organised into ASCII characters, which are sent, through the USART-USB interface, to the computer, where they can be further processed.

The measurement algorithms, managing both the generation of the stimuli sent to the chip and the transmission of data to the computer, are integrated within the FPGA, that implements multiple finite state machines. The number of states, as well as the transition conditions, depends on the type of measurement which has to be performed. Generally, measurements evolve between an idle state and a set of states performing chip configuration, stimuli generation, data collection and USART transmission. The hardware description programs, managing the different measurement phases, have been developed

in the Quartus II environment, by using VHDL (VHSIC Hardware Description Language). The RTL (Register-Transfer Level) code used to program the FPGA, involving both the measurement procedures and the modules running the communication protocols, was entirely designed in the frame of this thesis activity. In order to fulfil the requirements in terms of speed, digital IO pins and number of available logic elements, an FPGA of the Altera Cyclone II family (EP2C35F672C6) has been employed. The FPGA, featuring the highest speed grade allowed for a Cyclone II device, is mounted on a development board (Altera DE2 Board), providing a number of additional components that add various functionalities to the measurement setup. The system clock, featuring a frequency of 50 MHz, is based on a ceramic oscillator, integrated in the DE2 Board. The oscillation frequency, managing the measurement procedures, is compliant with the speed requirement of the internal structures of the ASAP110LF chip. The DE2 Board contains also an array of programmable switches, which have been used to customize the measurement steps, an array of programmable LEDs, used for debug purposes, and two arrays of pins (2×40 pins totally), representing the physical interface between the FPGA and the other components of the measurement system.
The computer is used to program the FPGA, to control the measurement procedures, to collect the measurement data, and to communicate with all the external components, which are not part of the core measurement setup. The FPGA programming and the acquisition of the measurement data are

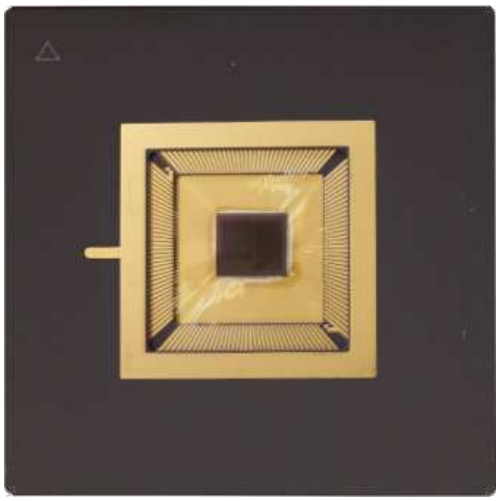


Figure 3.2: CPGA144 package accommodating a bonded ASAP110LF chip.

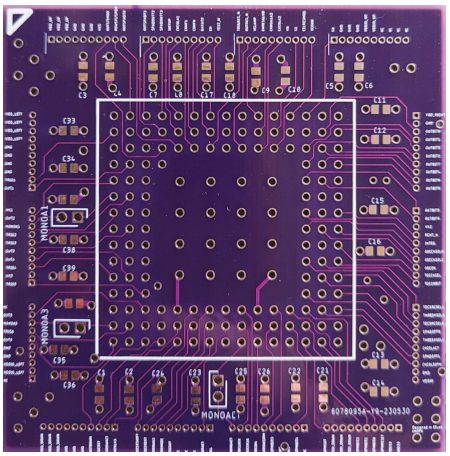


Figure 3.3: Photograph of the carrier board used for the characterization of the ASAP110LF chip.

performed on two different physical channels. During a measurement, the resulting data are sent from the FPGA to the computer by means of a hardware USART-USB interface. The latter, after receiving packets of data in the USART protocol from the FPGA, injects them into the USB port, which is directly connected to the computer. The communication signals, providing instructions to the FPGA, are transmitted in the opposite direction, through the same hardware interface. Once acquired, measurement data are stored in a cloud database, before being processed offline in the Matlab environment.
For the sake of characterization, some samples of the ASAP110LF chip have been packaged in a CPGA144 (Ceramic Pin Grid Array with 144 pins), with removable lid. A photograph of the above mentioned package, containing a bonded chip, is shown in Fig. 3.2.
Two custom boards, namely mother and carrier boards, were designed with the purpose of connecting the chip under test to the measurement system. During the characterization, the carrier board, containing the chip package, was mounted on top of the mother board, through a dedicated Preci-dip custom socket. The dual board approach provides greater flexibility in terms of modularity compared to a single board one, without significantly compromising the compactness and robustness of the overall system.
The top view of the carrier board used for the characterization of the structures inside the ASAP110LF chip is depicted in Fig. 3.3. The board, with a dimension of $7 \times 7\ cm^2$, is a PCB (Printed Circuit Board) with 4 layers, including some SMD capacitors, used to filter out high frequency components from

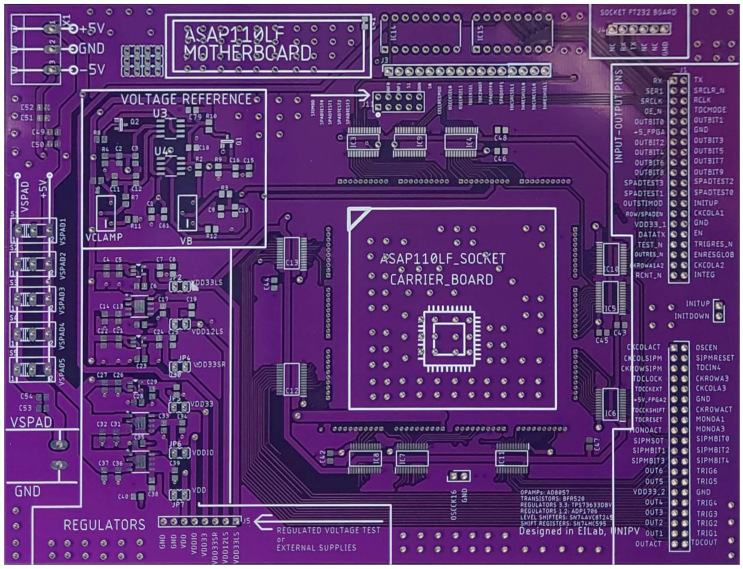


Figure 3.4: Photograph of the mother board used for the characterization of the ASAP110LF chip.

power supplies and reference voltages, and an array of 144 interconnecting plated through holes. The carrier board can accommodate either a CPGA144 compatible socket or directly the chip package. In the second case, the overall parasitic capacitance of the interconnections between chip and boards can be strongly reduced, even though a carrier board for each device under test is required. The three monostable signals generated by the ASAP110LF chip (MONOA1, MONOA3 and MONOACT) can be accessed directly on the carrier board through dedicated test points, located close to the chip package. By applying a jumper bridge shorting the two joint ends of a test point, the monostable signals can be routed towards the mother board.
The top view of the mother board is shown in Fig. 3.4. The mother board ($14 \times 18\ cm^2$) is in charge of regulating the voltages provided by the power supply, generating the voltage references required for chip operation and integrating various components acting as interfaces between the ASAP110LF chip and the measurement setup. The different sections making up the mother board can be subdivided as in the following list:

- Power connectors and switches, providing the power supply to the board and the SPAD sensors. The circuitry power supply ($5\ V$, $-5\ V$, GND) and the SPAD bias voltage (V_{SPAD}) can be provided to the board

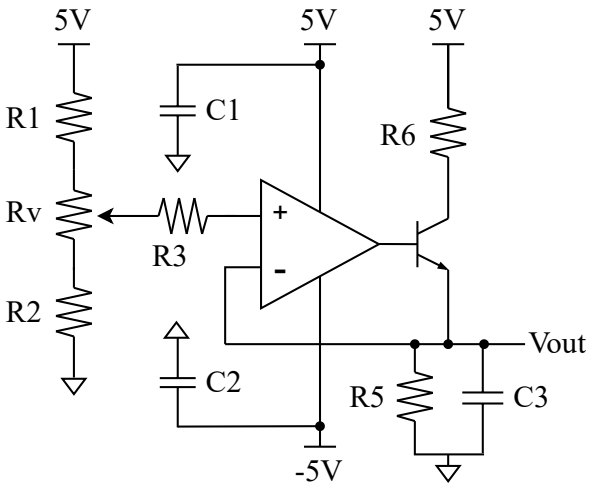


Figure 3.5: Circuit used for the generation of the VCLAMP and VB voltage references.

through these connectors, located on the left side of the board. Five switches, one per each VKx pad of the chip, are used to select the SPAD bias voltage. Depending on the switch position, the relevant SPADs can be either connected to 5 V or to the bias voltage available at the V_{SPAD} connector.

- Circuits for the generation of the VCLAMP and VB voltage references. The scheme used for the two circuits is shown in Fig. 3.5. The output voltage ($Vout$) is determined by the tunable resistive voltage divider, through the virtual short circuit at the OPAMP input terminals. The feedback provided by the n-p-n bipolar transistor ensures a stable output voltage, since the BJT emitter is directly connected to $Vout$, while the base can assume whatever voltage is needed to make the emitter voltage equal to the one generated by the resistor divider. This approach guarantees the generation, with a low output impedance, of a stable and tunable voltage reference, as required by the application.

- Framed connectors (2 arrays of 20×2 pins) that are used to communicate with the FPGA through a wired connection.

- A 16-bits shift register, storing the configuration bits provided to the ASAP110LF chip. Since the 16 configuration bits should remain constant throughout the entire measurement, the use of a shift register,

storing these signals, while properly driving them to the chip, allows to reduce the overall number of physical connections between the FPGA and the board. Two ICs, each implementing an 8-bit shift register, are cascaded to attain the required number of memory elements. Before the measurement starts, in the configuration phase, the FPGA drives the shift register control signals, so as to load the 16 configuration bits into the flip-flops of the ICs. Once the output drivers of the shift register are enabled, the 16 bits are provided to the device under test. If the two ICs are not mounted on the board, the configuration bits can be applied externally, through an array of 16 connectors. Alternatively, these connectors can be used as test point for the output signals generated by the shift register.

- Level shifters, used to adapt the voltage range of the FPGA signals ($0 - 3.3\ V$) to the one supported by the ASAP110LF chip ($0 - 1.2\ V$). The level shifting operation is performed for signals propagating from the FPGA to the chip under test, as well as in the opposite direction.

- Regulators, which generate the supply voltage used in the ASAP110LF chip and in multiple structures integrated within the board. Starting from the 5 V coming from the power supply, three 1.2 V regulators generate the $VDD_{CHIP-CORE}$, $VDD_{CHIP-IO}$, $VDD_{LOW\ level\ shifters}$ voltages. Analogously, the same number of 3.3 V regulators are used to provide the $VDD_{CHIP-33}$, $VDD_{shift\ register}$, $VDD_{HIGH\ level\ shifters}$ voltages. An array of 6 pins was added as a test point for the regulated voltages. The outputs of the six regulators can be selectively disconnected from the target circuits. In such a condition, the supply voltage can be directly provided to the structures in the board through the aforementioned 6-pin array.

- A socket for the accommodation of the carrier board, made of Preci-dip connectors arranged in a square shape.

3.2 Characterization results

In this section, the results from the characterization of the ASAP110LF chip will be presented. The aim of the measurement campaign, started in the frame of this thesis work, is to provide a thorough characterization of the structures integrated in the ASAP110LF chip. The main SPAD parameters, as well as the performance of the readout circuits described in chapter 2, will

be investigated, in view of future implementations targeting the development of more complex sensors. The study of sensor parameters like dark noise, breakdown voltage and detection efficiency is essential for the development of effective detection systems which may beneficially leverage SPAD features. By means of a measurement campaign stressing the sensors in different working conditions, various aspects of DCR can be analyzed, such as susceptibility to crosstalk, temperature dependence and random telegraph signal-like fluctuations in time. Understanding how SPAD parameters vary with operating conditions and identifying the main sensor weaknesses may represent the starting point of future research aiming at pushing the boundaries of CMOS SPAD array technologies.

When possible, the results from the characterization of the ASAP110LF chip will be compared to the results from SPADs fabricated in a 150 nm CMOS technology, developed in the framework of the APIX2/ASAP project. The sensors used in the comparison are integrated within the APIX2LF chip (see Section 1.4), representing a test chip consisting of a number of SPAD arrays, having different sensor sizes and readout circuits [5][34]. Although the APIX2LF chip is a two tier detection system, made up of two layers of overlapping SPADs, some samples implementing only a single layer of sensors were bonded onto dedicated packages and were available for test. The DCR characterization performed on single and dual layer samples of the APIX2LF chip had a pivotal role in demonstrating the beneficial impact that a coincidence-based structure has on the dark noise of the overall detection system. The characterization of the sensors in the ASAP110LF chip can be useful to understand whether the extremely low values of DCR, attained with the APIX2LF chip at room temperature, can be achieved also in more scaled technologies. In addition, the comparison between the noise performance featured by the two detection systems may provide significant results highlighting the effectiveness of a CIS technology in the production of SPAD-based sensors. It can be worth specifying that the ASAP110LF chip is currently available only as a single layer device, even though the entire design was carried out in view of a two tier assembly. However, an estimation of the DCR, potentially featured by a dual layer detection system based on the ASAP110LF chip, will be presented in the following.

The last part of this section will be devoted to the characterization of the new SiPM architecture developed in the framework of this Ph.D. program. Different measurements were performed to provide a first characterization of the readout electronics described in chapter 2. The input-output characteristic will be presented as a preliminary benchmark to evaluate the effectiveness of

the novel architecture, as implemented in the prototype. The measurements taken from this structure will serve as a foundation for further improvements to the current design.

3.2.1 Measurement methods

The characterization of the SPADs in the ASAP110LF chip was performed mainly in terms of detection efficiency, DCR and breakdown voltage. The availability of a relatively large number of sensors, with different dimensions and readout approach, allowed the collection of a statistically meaningful set of data. Different measurement conditions, potentially affecting SPAD performance, were applied to the DUTs. Unless specified, all the measurements described in the following were performed at $26^\circ C$, with an uncertainty of $\pm 0.5^\circ C$. Except for the measurements aiming at studying the detection efficiency of the sensors, the DUTs were kept under dark conditions by means of a ceramic removable lid, completely covering the die surface. In addition, the entire measurement setup, except for the monitoring components, was arranged inside a dark box, so as to completely isolate the sensors from the external light.
The measurement campaign involved four different chips, which will be referred to as C3, C4, C5 and C11 in the reminder of this manuscript. All the chips under measurement, except for C11, were bonded on independent packages before the measurement campaign started. During all the measurements, the supply voltages were provided to the chips through the voltage regulators mounted on the mother board. The VCLAMP and VB voltage references, representing respectively the gate voltage of the clamping transistor and the quenching voltage (see Section 2.3), were fixed respectively at 1.6 V and 0.8 V. Unless otherwise specified, the monostable duration was set to 2 ns, while, if allowed by the chip circuitry, parallel reading was always enabled to speed up the measurement procedures.

3.2.2 Breakdown voltage

The operational region of a photodiode in an avalanche detector is significantly influenced by the breakdown voltage (V_{BD}). As long as the bias voltage is maintained below the breakdown voltage, the detector operates linearly, with the output current being directly proportional to the incident light. Conversely, if the diode is biased with a voltage exceeding the breakdown voltage, a single photon or charged particle, interacting with the active volume, can induce an avalanche in the device. Therefore, an accurate characterization

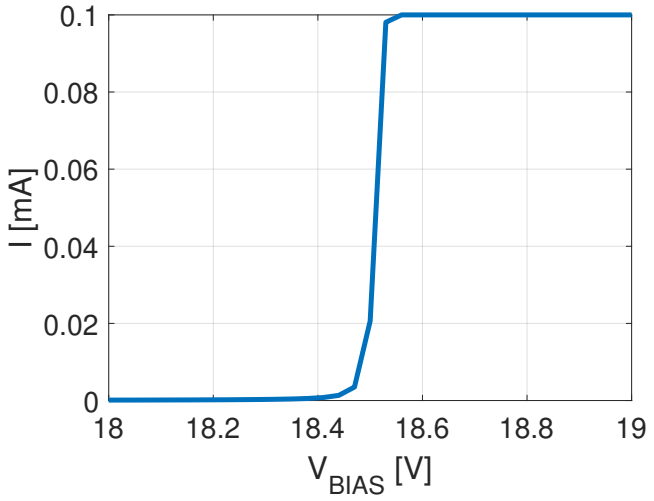


Figure 3.6: I-V curve measured on a single SPAD (SPAD1) in C11 ($V_{BIAS} = V_K - V_A$).

of the breakdown voltage is crucial to properly bias the device in the desired operating region.
Fig. 3.6 shows the I-V curve featured by the sensor marked as "SPAD1", belonging to the group of single sensors located within the padring of C11. The V_{BIAS}, represented on the x axis, is taken as the difference between the cathode and the anode voltage ($V_K - V_A$). I-V measurements were performed by using a micro-probing station, while data were acquired through a semiconductor parameter analyser (Agilent Technologies E5270B) equipped with precision medium-power SMU modules (HP E5281B). The measurement was managed in different steps, each of them involving the increment of the SPAD bias voltage by 0.02 V. The results shown in figure were obtained under a faint illumination condition, so as to guarantee the avalanche generation at all the bias voltages above the breakdown voltage. As shown in figure, before the SPAD bias voltage approaches 18.5 V, the sensor is biased in the linear gain region. In this operating condition, single carriers generated within the space charge region do not have a significant effect on the output current. For bias voltages around 18.5 V, the current swiftly rises, since single photons absorbed in the active volume can achieve a non-negligible probability of triggering a self-sustaining avalanche. As a result, 18.5 V can be considered as the breakdown voltage (V_{BD}) of the structure under measurement. In order to prevent any potential damage to the device under test, a compliance of 0.1 mA was set on the measured current, as evident at voltages exceeding the breakdown

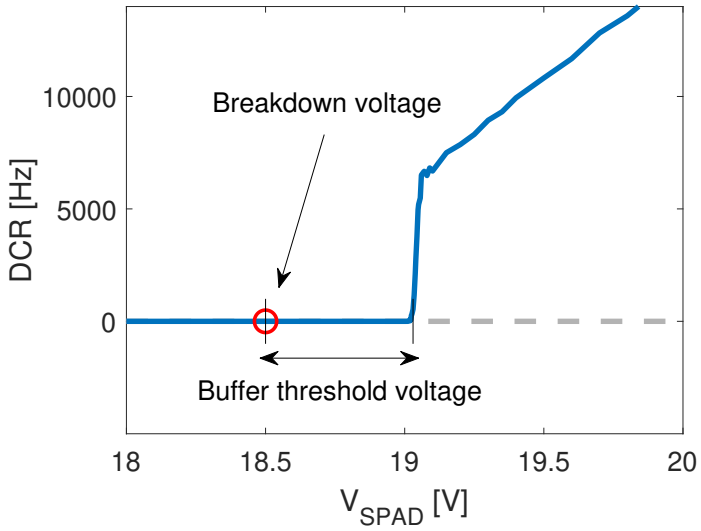


Figure 3.7: Example of breakdown voltage extraction.

voltage.
The breakdown voltage of the sensors integrated in the array structures was obtained with dedicated measurements. Since the anode terminals of the detectors are not reachable, due to their co-integration with the readout circuits, the direct acquisition of the sensor current, aiming at retrieving the I-V curve, is not a viable method. As shown in Fig. 3.7, the breakdown voltage was determined pixel by pixel, by subtracting the threshold voltage of the buffer integrated in the frontend channel (with reference to the circuit shown in Fig. 2.9) from the voltage at which the DCR vs V_{SPAD} curve abruptly departs from zero [135]. Monte Carlo simulations indicated that the buffer threshold voltage was distributed around the nominal value of 490 mV, with a standard deviation around 4 mV, negligibly affecting the breakdown voltage extraction. In a voltage range of about 40 mV around the swift DCR escalation, measurements were performed with a resolution of 2 mV. More details about the DCR dependence on the V_{SPAD} will be provided later on in this chapter.
In Fig. 3.8, the distributions of the breakdown voltages featured by A1 in C3 and A1, A2 and AAQ in C5 are shown. The average value featured by each distribution and the relevant standard deviation are also indicated in the diagrams. On the Y axis, the percentage of pixels, expressed in relation to the total number of pixels in the array, is reported, so as to allow a convenient comparison between distributions from arrays with a significantly different number of pixels. As shown in the figure, similar values for the breakdown voltage were

obtained between arrays belonging to different chips and between different arrays in the same chip. V_{BD} values ranging between 18.48 V and 18.56 V were found throughout all the arrays. These results are compliant with the I-V curve depicted in Fig. 3.6, featuring a breakdown voltage located in the aforementioned range. The average value of the breakdown voltage is in accordance with the results from the characterization of SPADs in the APIX2LF chip [5], which integrates sensors based on a similar p+/nwell junction. However, the standard deviations found for the devices in the ASAP110LF chip are significantly smaller than the ones featured by the SPADs fabricated in 150 nm CMOS technology.
The V_{BD} heat maps featured by the two A1 arrays in C3 and C5 are shown in Fig. 3.9. In both the arrays, the breakdown voltage looks higher for pixels located in the bottom right corner of the structures. The apparent V_{BD} gradient

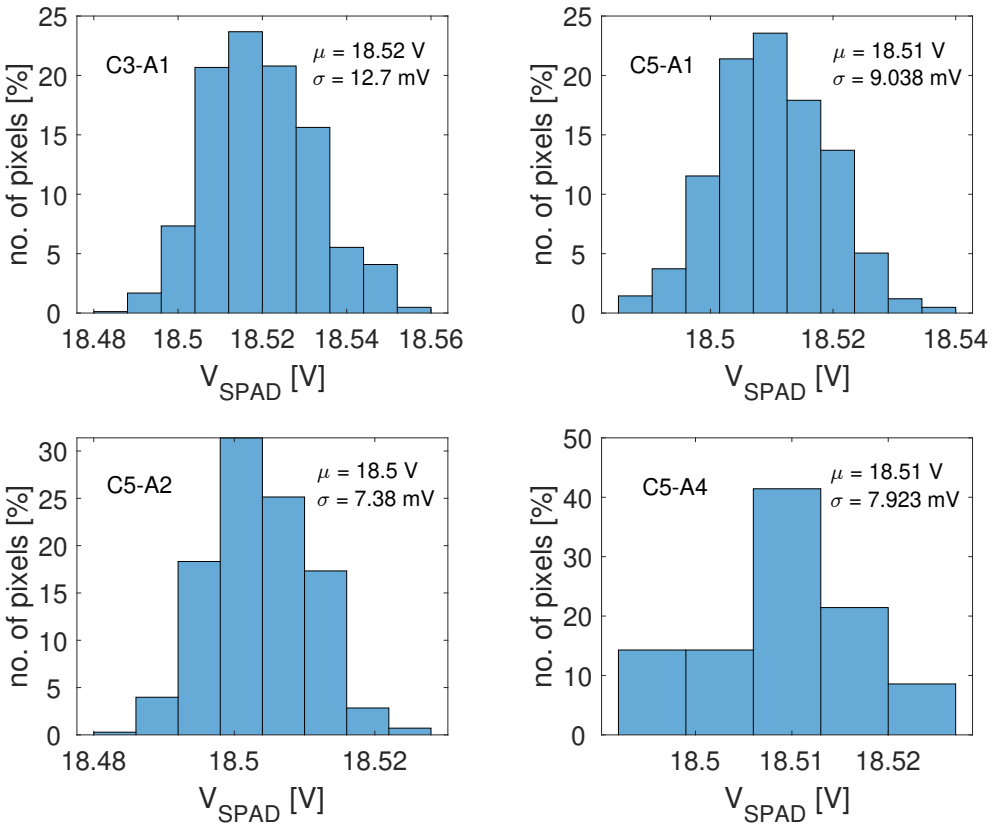


Figure 3.8: Breakdown voltage distributions for different arrays contained in C3 and C5.

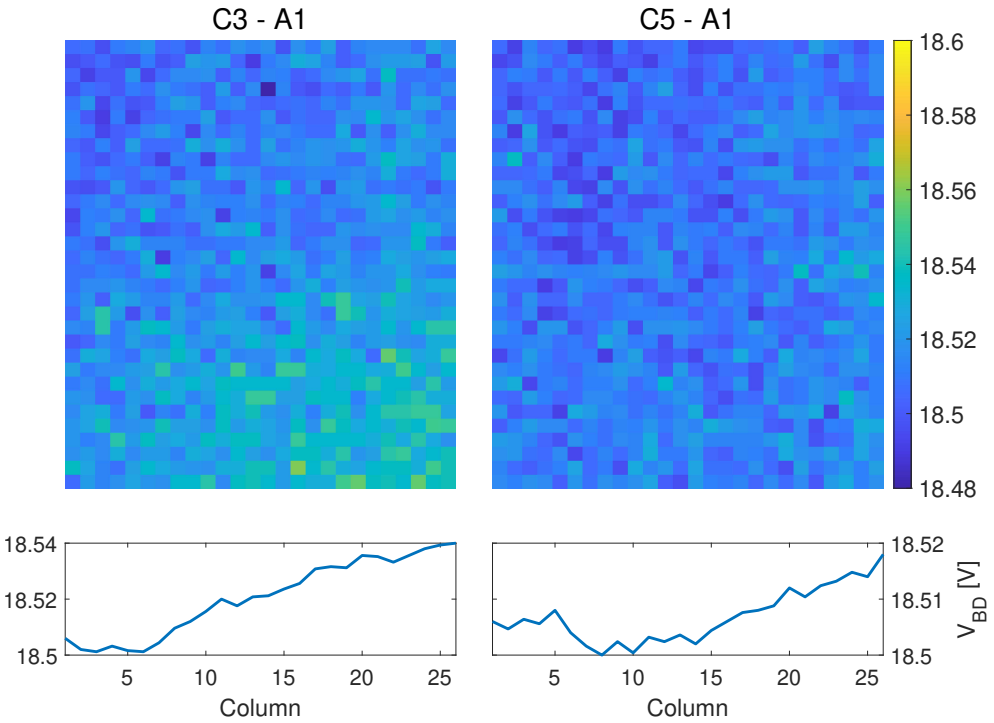


Figure 3.9: Heat map of the breakdown voltage for the A1 array of C3 and C5 chips. The diagrams showing the V_{BD} gradient along the main diagonal of the two arrays are included.

observed on these arrays is described in the two bottom diagrams of Fig. 3.9, representing the breakdown voltage of pixels along the main diagonal of the two arrays. This trend, contributing to increasing the standard deviation of the V_{BD} distributions, may be due to a non-uniformity in the SPAD bias voltage, which may differ throughout the array. Since the cathode voltage, which is shared between all the SPADs in the array, is provided through a single metal connection located in the top left corner of the structure, the voltage drop experienced by SPADs far from the VK1 pad cause non-homogeneous bias conditions among the SPADs of the array. Sensors located in the bottom right corner may be supplied with a bias voltage lower than the nominal one, thus resulting in a higher equivalent breakdown voltage. The existence of an external factor creating, at the array level, a gradient on the breakdown voltage, possibly identified as a voltage drop across the cathode voltage line, is demonstrated by the fact that the standard deviation decreases down to a minimum (7.3 mV) in subsets of adjacent pixels randomly chosen inside the

structure. However, the number of SPADs featuring a breakdown voltage significantly departing from the mean value is relatively small, as demonstrated by the histograms in Fig. 3.8. A gradient in the breakdown voltage was observed only in A1, which is the largest structure in the chip. All the other arrays, designed with smaller dimensions, feature an improved V_{BD} uniformity.

3.2.3 Photon detection

The PDP of the SPADs in the ASAP110LF chip was measured by illuminating a single device, located in A2 array of the chip C5, with monochromatic light at different wavelengths. To the measurement setup described in section 3.1, a broad-spectrum light source coupled with monochromatic optical filters was added to perform the characterization. The light intensity was measured using a calibrated power meter. The measured PDP spectrum, obtained for $V_{EX} = 1\ V$, is shown in Fig 3.10. The maximum of the detection probability (21%) was found around $\lambda = 450\ nm$, in accordance with the results obtained for similar sensors in different technologies [41]. For CMOS SPADs, the PDP is marginally dependent on the employed technology, while being strictly related to the SPAD design [136]. As compared to recent SPADs specifically targeting high values of detection efficiency [137], a relatively low PDP can be attained with the use of standard p+/nwell junctions, as the de-

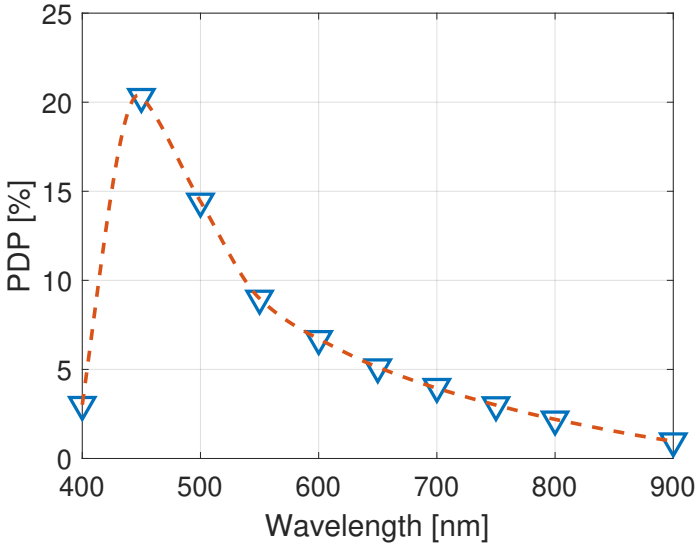
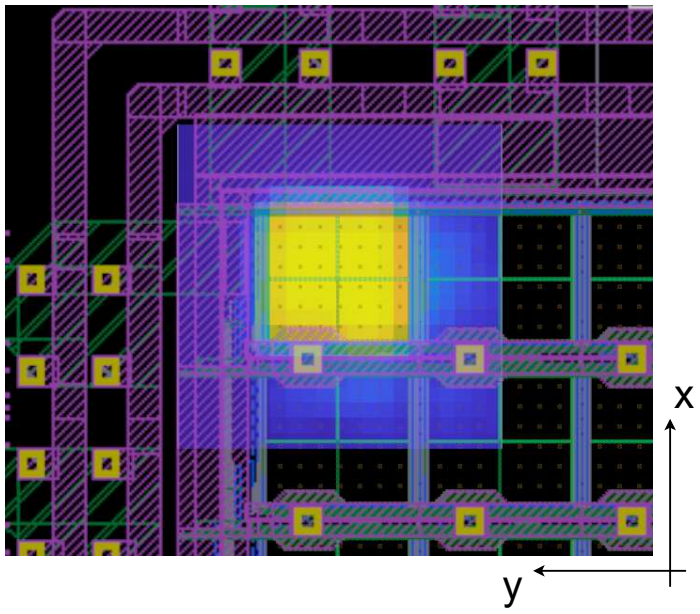
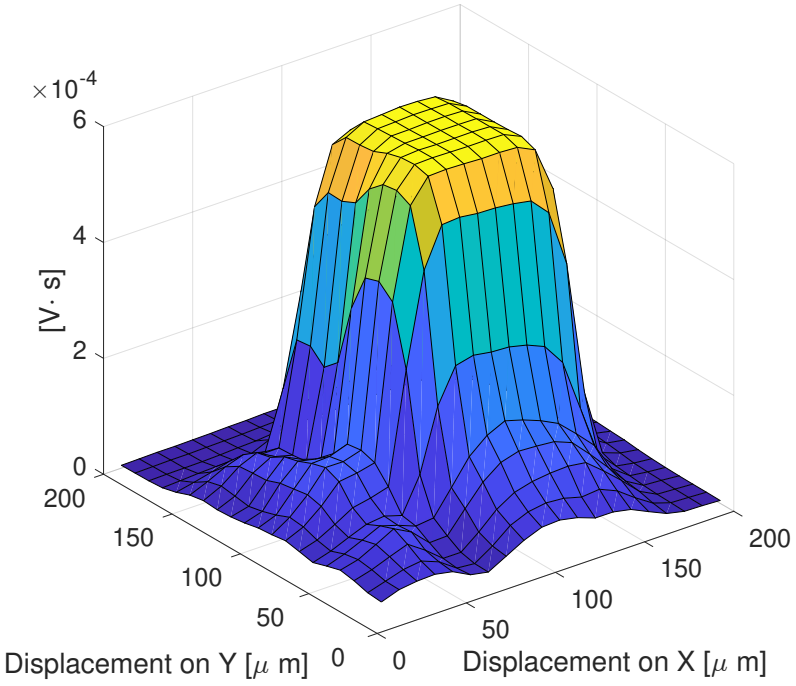


Figure 3.10: Photon detection probability (PDP) spectrum for a SPAD located in A2 of C5 ($V_{EX} = 1\ V$).

pleted layer width is shrunk by the high doping concentrations used to build up the sensing region. In [138], a maximum PDP of 29%, around $\lambda = 440\ nm$, was obtained with $V_{EX} = 1\ V$ by using n+/high-voltage p-well (N+/HVPW) sensors fabricated in the same technology as the one of this work. The PDP enhancement, as compared to the results shown in Fig 3.10, can be ascribed to the higher doping levels, used in the aforementioned work, leading to a higher breakdown probability. A compensation technique, acting on the doping profile of the HVPW layer, is also adopted to reduce the dark noise originating from the use of a high doping concentration. Similar results (PDP=32% at $\lambda = 450\ nm$, with $V_{EX} = 1\ V$) were obtained in [127], where the noise performance of CMOS SPADs, integrating PW/DNW junctions, are investigated as a function of the STI surrounding the sensor structure. Measurements at different excess voltages, potentially enabling further meaningful comparisons with PDP results presented in the literature, need to be performed, so as to provide more information about the carrier generation mechanisms affecting the avalanche triggering probability [41].
Photon detection experiments were conducted on the (1,1) pixel in A1 of C5, to qualitatively examine how various elements surrounding the active region of a SPAD may influence its detection efficiency. The measurement was performed on a pixel located at the boundary of the array, so as to appreciate the effect of high level metal layers, running on two sides of the pixel, on the detection capabilities of the sensor. A semiconductor laser ($\lambda = 1060\ nm$) was used, with a pulse frequency of 10 kHz and a 50% duty cycle. The surface of the pixel, and part of the surrounding area, was scanned, with a step of 10 μm, by means of a 3-axis micro positioning system, made of a mainframe (Newport ESP300) and three precision linear motors (Newport MFA-CC). A total area of $200 \times 200\ \mu m^2$ was covered, including the pixel area, a portion of top level metal tracks, and part of the three adjacent pixels. The laser was operated at a power level preventing the saturation of the SPAD response when pointing outside the active area of the sensor. During the measurement, which was performed at $V_{EX} = 1\ V$, only the front-end circuit of the pixel under investigation was enabled. The detection signal was recorded through the MONO_A1 output pad, which was connected to a digital oscilloscope (R&S RTO2014) in charge of collecting and processing the data. The monostable of the front-end circuit was set in transparent mode for all the measurement duration. The detection pulses, produced by the SPAD, were integrated over a time window of 1 ms. The measurement was repeated more than 5000 times per each laser position, thus providing a result consisting of the average value computed over the total number of acquired samples. In this way, a matrix



(a)



(b)

Figure 3.11: Photon detection qualitative measurement: (a) heat map superimposed onto the chip surface, (b) 3D diagram of the measurement data shown in (a).

of 400 points (20×20), each related to a different location around, or within, the pixel area, was built. The sum of the pulse areas was considered in place of the total number of pulses, so as to get rid of the paralyzable behaviour of the front-end circuit. Indeed, at high detection rates, the SPAD may be never reset back to the idle state, thus producing a long pulse, instead of a train of discrete pulses, as a response to a dense flux of impinging photons. In this case, counting the individual pulses generated by the sensor may produce a misleading result. If the pulse area is considered, the output data have the dimension of $V \cdot s$. Since it is not possible to distinguish between a photo-generated avalanche and a dark event, noise pulses were also recorded during the measurement. However, the effect that dark noise had on the detection characterization was negligible, since contributing only with a few $nV \cdot s$.
The measurement results are shown in Fig. 3.11. Fig. 3.11a consists of a two-dimensional heat map, which represents the measurement data, overlapped with the top view of the pixel under consideration. Fig. 3.11b shows the same data in a 3D diagram. As expected, the heat map points corresponding to the active area feature the highest values of pulse energy (in Fig. 3.11a, the active area can be easily identified by means of the two green perpendicular lines dividing the SPAD active surface in four quadrants), since the photons produced by the laser can be directly absorbed within the active volume of the sensor. In most of the portions of the active area that are directly exposed to the laser beam, the detection system reaches the saturation level ($600\ \mu V \cdot s = 1.2\ V \times 1\ ms/2$, where the division by 2 is due to the 50% duty cycle of the laser pulse). In this case, a rectangular waveform, with the same frequency of the pulsed signal controlling the laser, was generated at the front-end output terminal. Due to the high rate of photons absorbed in the active volume, the SPAD bias voltage cannot be completely reset to the initial condition, thus preventing the anode voltage from going below the threshold of the subsequent digital buffer. A lower sensitivity can be observed at the border of the active area, where edge effects reducing the PDP may occur [139]. It can be noted that, in proximity of the sensor corners, the number of absorbed photons is significantly lower than in the side regions, since the corners of the active area are cut with a 45° angle (not shown in figure). Moving away from the SPAD active area, in the direction of the other pixels of the array, the number of recorded pulses was found to decrease, since, in this case, photons could reach the active area of the enabled pixel only by being scattered by the surrounding structures. The high level metal tracks, represented in Fig. 3.11a as purple stripes, effectively shield the sensor active area against photons, as demonstrated by the abrupt change of colors that is featured in

the top and left sides of the two-dimensional heat map. Moreover, a portion of metal track, providing the power supply to all the pixels in the row, partially overlaps with the active area of the SPAD, thus leading to a substantial loss of sensitivity that can be observed in the bottom-left quadrant of the sensor. Even though pixels on the border are often affected by edge effects, the results obtained from the measurement discussed in this section are fairly in agreement with the pixel geometry, which is easily retrievable by inspection of the 20×20 points detection map.

3.2.4 Dark count rate

The dark count rate is a key parameter for SPAD detectors. Devices fabricated in CMOS technologies may exhibit significantly high DCR values as compared to devices fabricated in custom technologies. High levels of dark noise may jeopardize the sensor capabilities as radiation detector, thus making such devices completely unsuitable for applications that require single photon resolution. As a matter of a fact, DCR characterization is of paramount importance to determine the effectiveness of a CMOS SPAD detector.

3.2.4.1 DCR measurement procedures

The noise performance of most of the array structures in the ASAP110LF chip was investigated. For each array, the DCR was evaluated at the sensor level. Individual noise measurements were performed pixel by pixel, thus allowing the collection of relatively large arrays of data. When the DCR of a single SPAD was being measured, all the other SPADs in the array were prevented from generating noise pulses by disabling the relevant front-end circuits. During the characterization of pixels in A1, the DCR measurements were carried out by exploiting the parallel reading allowed by the relevant row selection register. In this case, four pixels (one per group) at a time were concurrently enabled, thus significantly speeding up the measurement procedure.
Depending on the array, different procedures were used to measure the DCR of the relevant SPADs. For the A2 array, where a 10 bit counter is integrated in each individual pixel, the DCR measurement procedure consisted of a preliminary phase performing pixel enabling, and a second part defining the measurement integration window, through the assertion of the INTEG signal. At the end of the measurement time, the binary word representing the DCR result was available at the output pads.
For pixels integrating the 1-bit memory, the DCR of each pixel was measured by enabling the quenching circuit and the readout electronics of the relevant

SPAD for a fixed time frame, referred to as Δt in the following. The output pad, connected to the enabled pixel, was observed to collect a potential dark pulse produced by the sensor. If a dark noise event occurred within Δt, a digital one was recorded. Then, the memory was reset, so as to allow the collection of a new noise event. The procedure was repeated N times, so as to cover a large time interval during which the SPAD was left free to produce dark pulses. This time interval, given by $N \cdot \Delta t$, is referred to as total integration time (TIT), and was set in a range of values between 100 ms and 300 s, depending on the SPAD under measurement. At the end of the procedure, an FPGA register, updated after each time frame Δt, was used to represent the DCR value of the currently enabled pixel. The individual DCR of SPADs integrated within the ASIPM array was measured in a similar way, by setting the SiPM threshold to 1 and collecting data from the SIPMSOT pad.
The choice of the time frame Δt and of the number of repetitions N is of paramount importance for the measurement accuracy, since it determines the maximum and the minimum DCR, DCR_{MAX} and DCR_{MIN}, that can be detected. In particular

$$DCR_{MAX} = \frac{N}{N\Delta t} = \frac{1}{\Delta t}, \tag{3.1}$$

$$DCR_{MIN} = \frac{1}{N\Delta t}. \tag{3.2}$$

After an evaluation of the DCR range featured by the pixels in the ASAP110LF chip, a time frame of 10 μs was chosen for most of the measurements shown in the following. Unless otherwise specified, $N = 10^4$, thus resulting in a TIT of 100 ms. With these parameters, the maximum rate that can be measured is 100 kHz, while the minimum one is 10 Hz.
By applying the measurement method discussed so far, a systematic error in the measured DCR must be taken into account. The discretization of the TIT into equally sized time frames of duration Δt may return a DCR value which is smaller than the actual one. Since the in-pixel memory can record a single noise event within each time frame, additional pulses after the first one, occurring in the same interval Δt, will remain undetected. Increasing the duration of the time frame leads to a higher number of lost pulses, thereby increasing the discrepancy between the measured DCR and the actual dark noise value. In a DCR measurement performed through the discretization of the TIT, the number of time frames producing no dark noise events (assuming that the probability density function of the dark counts follows a Poisson statistics) can be expressed as

$$n_0 = N \times P(0, \Delta t) = (n_1 + n_0)e^{-DCR_t \Delta t}, \tag{3.3}$$

where DCR_t is the actual (true) DCR featured by the sensor, n_1 is the number of time frames with at least a dark noise event, $N = n_1 + n_0$ indicates the total number of time frames within the TIT and $P(0, \Delta t)$ represents the probability of having no dark events in the time interval Δt, according to the Poisson statistics expressed in (1.14). The measured DCR can be written as

$$DCR_m = \frac{n_1}{N}\frac{1}{\Delta t}, \tag{3.4}$$

leading to

$$\frac{n_1}{n_0} = \frac{DCR_m \Delta t}{1 - DCR_m \Delta t}. \tag{3.5}$$

Starting from (3.3), by applying (3.5), the following expression can be obtained

$$DCR_t = \frac{1}{\Delta t}\, ln\left(\frac{1}{1 - DCR_m \Delta t}\right). \tag{3.6}$$

This relationship, expressing the actual DCR as a function of DCR_m and Δt, can be used as a correction factor to suppress the measurement error introduced with the discretization of the TIT (or pile-up error). Fig. 3.12 shows the DCR error as a function of the measured noise in the case of three

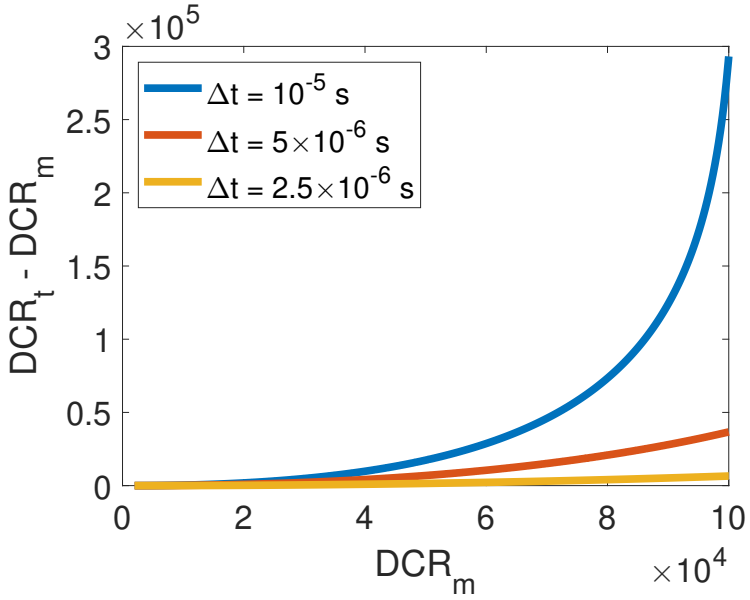
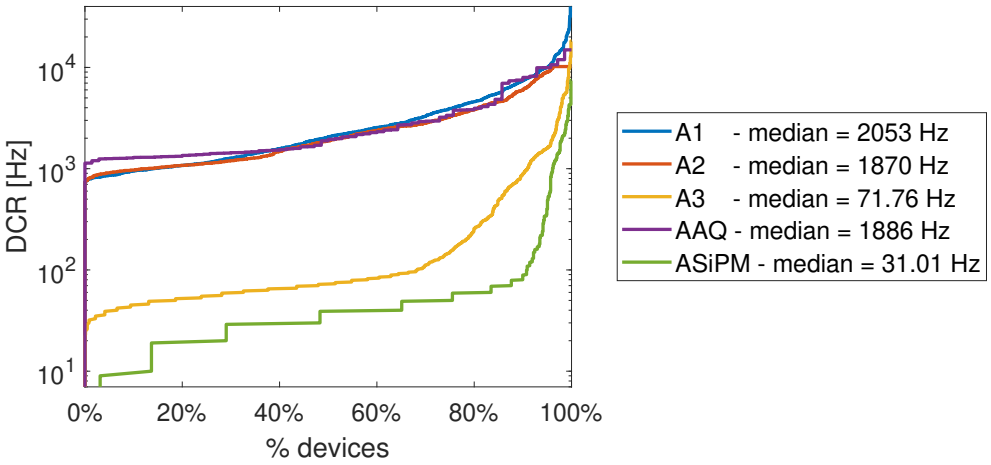


Figure 3.12: DCR error as a function of the measured DCR in the case of three different time frames.

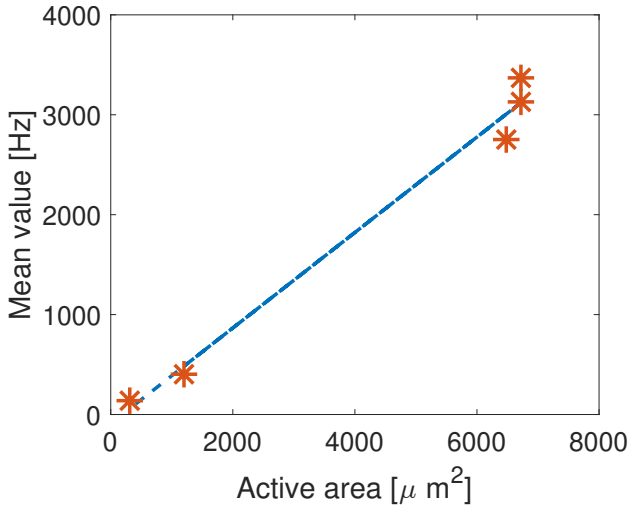
different time frames. In the diagram, a measured DCR not exceeding 10^5 Hz is assumed. It can be noticed that the discrepancy between the actual DCR and the measured one increases more than linearly as ever higher values of dark noise are measured. However, the measurement error is strongly reduced if small time frames are used, since the probability of losing dark pulses in a single time frame decreases. For the measurements shown in the following, a time frame equal to 10^{-5} s was chosen, since very few pixels were found to feature a DCR exceeding 30 kHz in the explored range of excess voltage.

3.2.4.2 DCR measurements

Fig. 3.13a shows the DCR cumulative distribution curves representing the noise performance of SPADs from different array structures in the ASAP110LF chip. All the curves were obtained at the excess voltage of 1 V, by means of the measurement procedures described in section 3.2.4.1. In the case of pixels from the AAQ array, equipped with an active quenching network, a hold-off time of 30 ns was selected. In order to perform individual DCR measurements on the subpixels of ASIPM, all the SPADs in the structure were considered as independent devices. Therefore, measurements were performed by enabling the front-end circuitry of a single subpixel at a time. The median value of each curve, taken as a meaningful parameter describing the noise performance of a homogeneous group of pixels, is shown in the figure. The cumulative distributions relevant to A1, A2 and AAQ cover less than two orders of magnitude, with a very small number of pixels exceeding 10 kHz (about 4%). ASIPM and A3 contain cells featuring heterogeneous levels of dark noise, ranging from 10 Hz to 6 kHz. This phenomenon is due to the Poissonian distribution of defects inside the depletion region of the sensors, that affects the DCR uniformity of pixels with relatively small active areas. However, the abrupt increase shown by the ASIPM distribution at high percentage values indicates that only a small number of SPADs is affected by a DCR above 100 Hz. The staircase-like shape of the ASIPM curve is due to the relatively low DCR featured by most of the subpixels, which returned dark noise levels close to the resolution of the measurement setup (10 Hz, with $\Delta t = 10^{-5}$ s and $TIT = 100$ ms). Moreover, it can be noticed that a small percentage of SPADs in ASIPM (about 3%) did not produce any dark pulse during the TIT, meaning that the noise values are below the minimum detectable DCR.
The DCR mean values relevant to the different array structures, as a function of the SPAD active area, are shown in Fig. 3.13b. The linear fit of the mean values is also included in the figure. As already demonstrated in [5], the dark noise was found to scale roughly linearly with the active area of the



(a)



(b)

Figure 3.13: Noise performance in different array structures of the ASAP110LF chip: (a) DCR cumulative distribution curves, (b) linear fit of the mean values as a function of the SPAD active area.

devices. As apparent from Fig. 3.13a, no significant differences were observed between curves relevant to arrays integrating SPADs with similar active areas. Conversely, the DCR obtained from A3 and ASIPM arrays, featuring a significantly smaller active area, is well below the DCR of arrays A1, A2 and AAQ.
Measurements at different SPAD cathode voltages were carried out, so as to evaluate the dependence of the DCR on the excess voltage. The $DCR-V_{SPAD}$ characteristics for five pixels of A1 in C3 are shown in Fig. 3.14. The number of dark counts significantly increases as the excess voltage exceeds the threshold voltage of the frontend buffer reading the SPAD anode voltage. As already discussed in section 3.2.2, the $DCR - V_{SPAD}$ curve of each pixel was used to extract the breakdown voltage of the relevant sensor. For bias voltages above 19.05 V, the DCR increases linearly with the SPAD voltage.
Fig. 3.15 shows the $DCR - V_{SPAD}$ characteristics featured by all the pixels contained in the same array as in Fig. 3.14. The 832 curves are separated in four subsets, according to the group division typical of A1.
Fig. 3.16 shows the heat maps of the DCR as a function of the SPAD bias voltage. At 19.04 V, an apparent gradient of the breakdown voltage, represented as a low-DCR region located in the bottom right corner of the array, can be observed. As already discussed, this effect can be ascribed to a non-uniform distribution of the SPAD bias voltage, which is connected to the structure

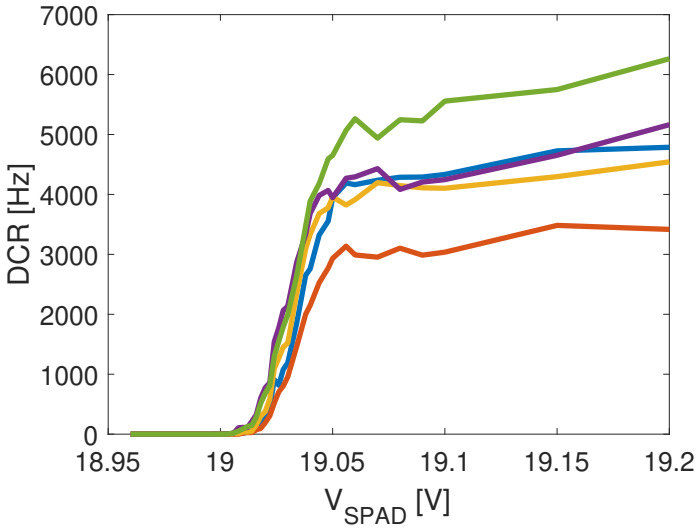


Figure 3.14: DCR curves as a function of the SPAD bias voltage for five pixels in A1 (chip C3).

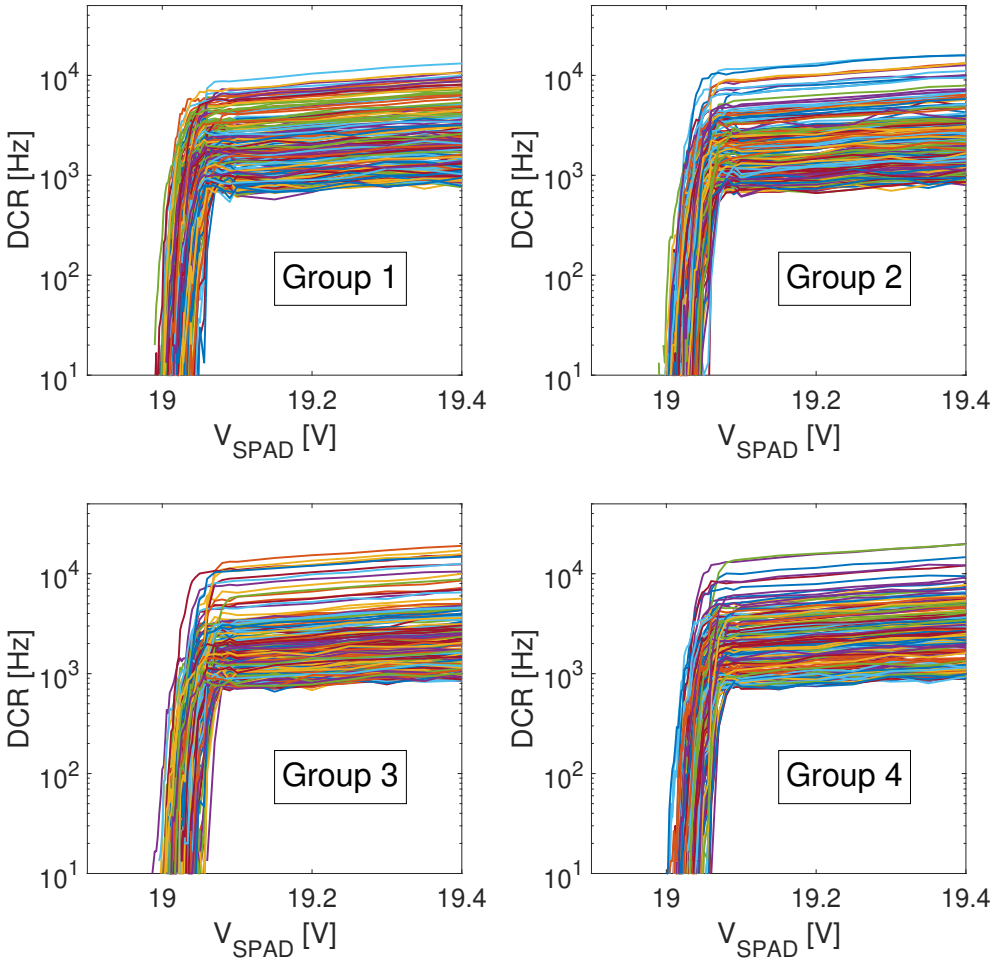


Figure 3.15: DCR curves as a function of the SPAD bias voltage for all the pixels in A1 (chip C3).

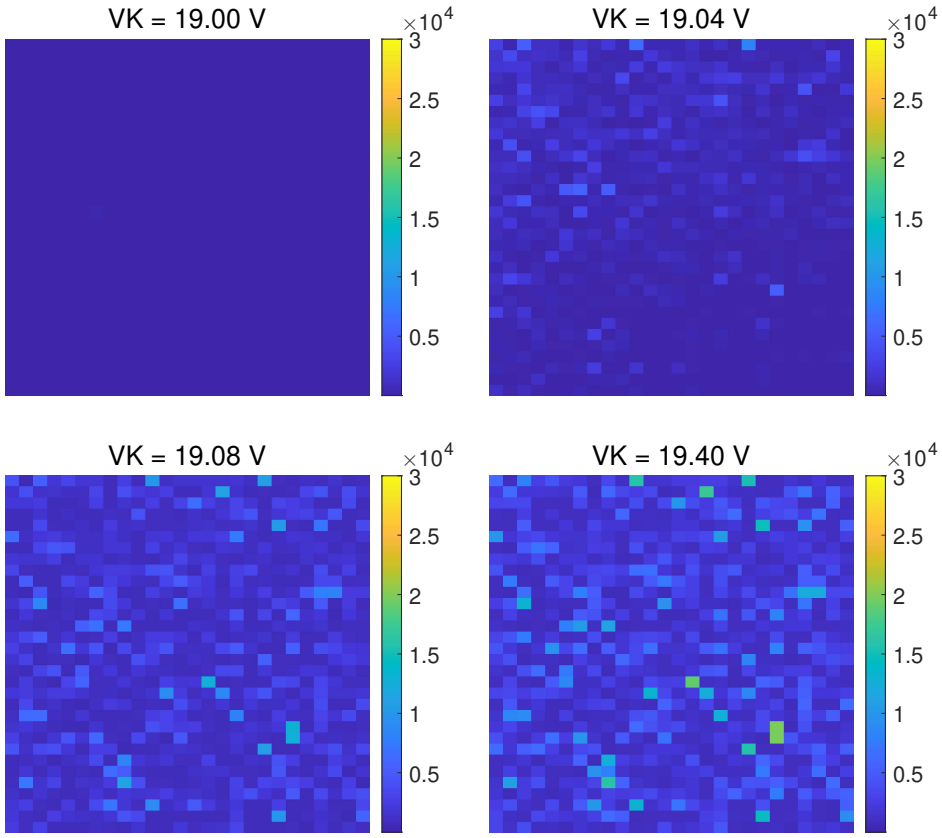
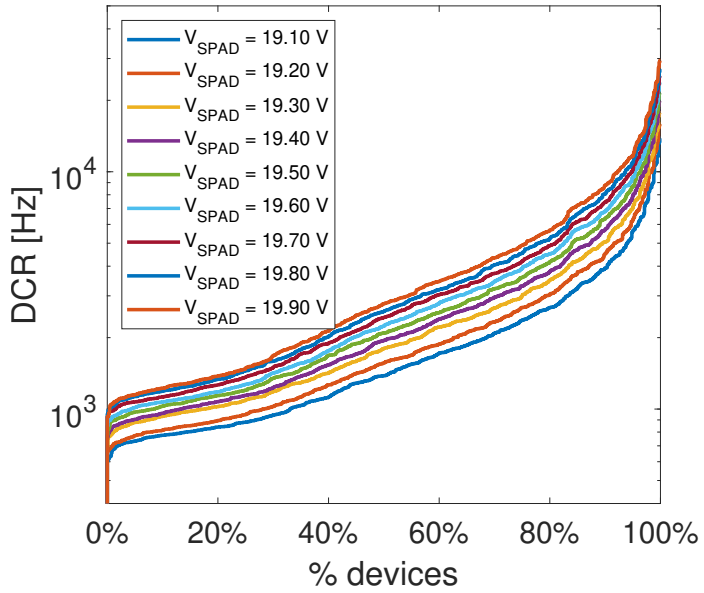


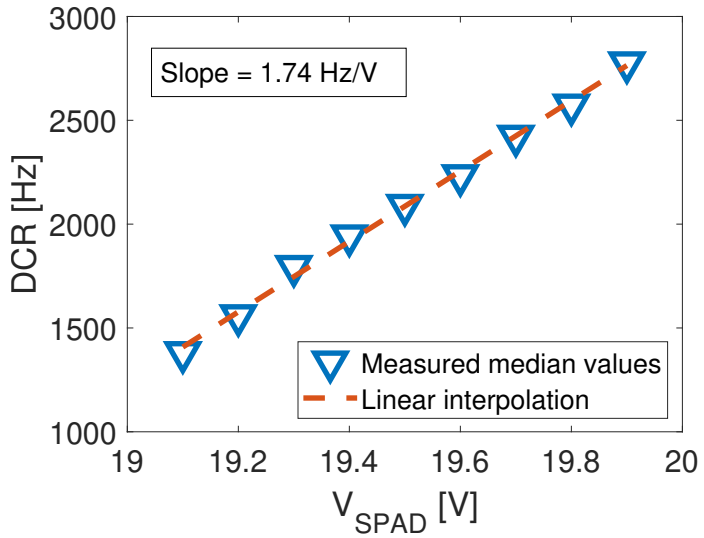
Figure 3.16: Heat maps representing the DCR as a function of the SPAD cathode voltage (A1 in C3).

through a high level metal track close to the top left corner of the array. At cathode voltages sufficiently above the breakdown one, the effects of the SPAD bias voltage drop are no longer discernible, since all the SPADs are biased well above the high slope region of the DCR-VSPAD characteristic. In addition, small clusters of particularly noisy pixels were found to be equally distributed throughout the array.

The DCR cumulative distributions of pixels in A1 (chip C3) at different SPAD bias voltages are depicted in Fig. 3.17a. Since the measurements were performed with excess voltages well above the buffer threshold level, all the pixels produced a number of dark pulses significantly exceeding the minimum detectable DCR. The linear dependence of the DCR on the V_{SPAD} is shown in Fig. 3.17b, where the median values extracted from curves in Fig. 3.17a are



(a)



(b)

Figure 3.17: DCR as a function of the SPAD bias voltage for SPADs in array A1 of chip C3: (a) cumulative distributions, (b) linear fit of the median values.

plotted against the SPAD bias voltage.
Preliminary measurements of the DCR as a function of the temperature were performed by using a DY110C ACS climatic chamber. Fig. 3.18 shows the cumulative distribution curves of DCR at different temperatures for SPADs in array A1 of chip C3. The temperature was varied in a range between $-40°C$ and $60°C$ through an automatic script, developed in the Labview environment and used to control the climatic chamber operation. The characterization was performed at the excess voltage of 1 V, by taking into account the linear dependence of the breakdown voltage on the temperature. Indeed, as the temperature increases, the carrier mean free path decreases, thus requiring a larger electric field to sustain the avalanche. A temperature coefficient of 15 $mV/°C$ was assumed [7][140], in accordance with an N-well doping level of a few 10^{16} cm^{-3} [141], which can be considered reasonable for the SPAD junctions integrated within the ASAP110LF chip. As expected, the number of dark pulses exponentially increases with the temperature. This effect is due to the enhancement of the thermal energy being available to the SPAD, promoting the thermal generation of an increasing number of free carriers within the space charge region. Fig. 3.19 shows the exponential fit computed over the median values reported in Fig. 3.18. At low temperatures, band to band tunneling is likely the primary mechanism contributing to dark noise [7][122], as, below $-20°C$, the DCR was found to significantly depart from the expo-

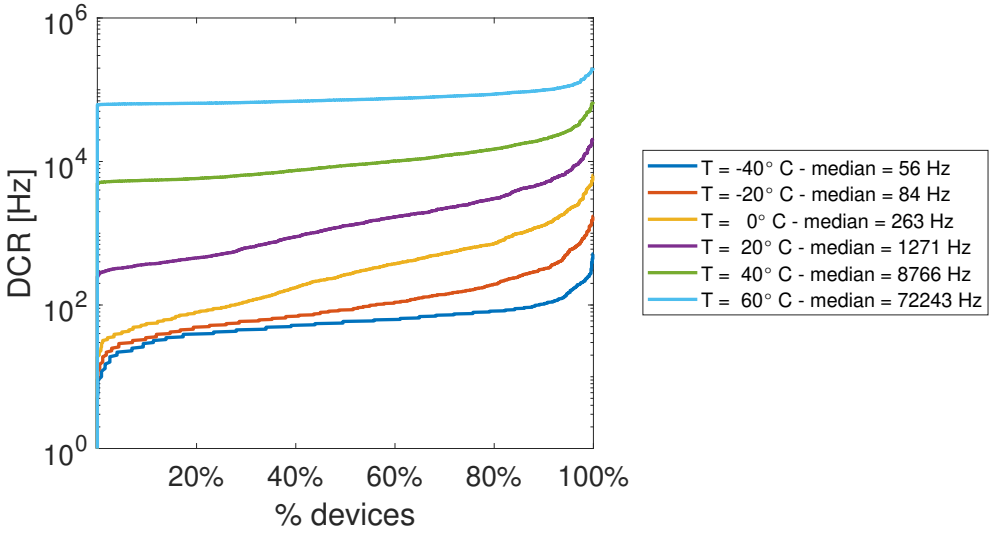


Figure 3.18: DCR cumulative distribution curves at different temperatures ($V_{EX} = 1\ V$).

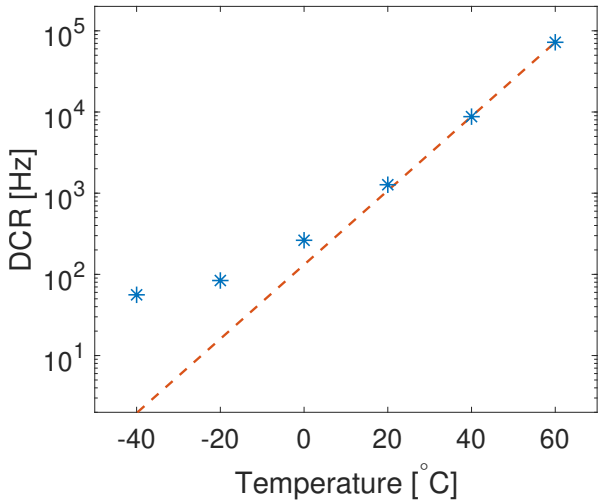
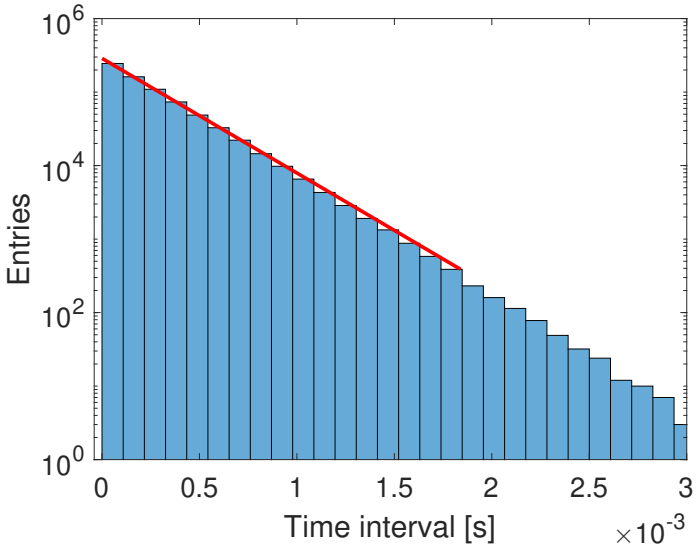
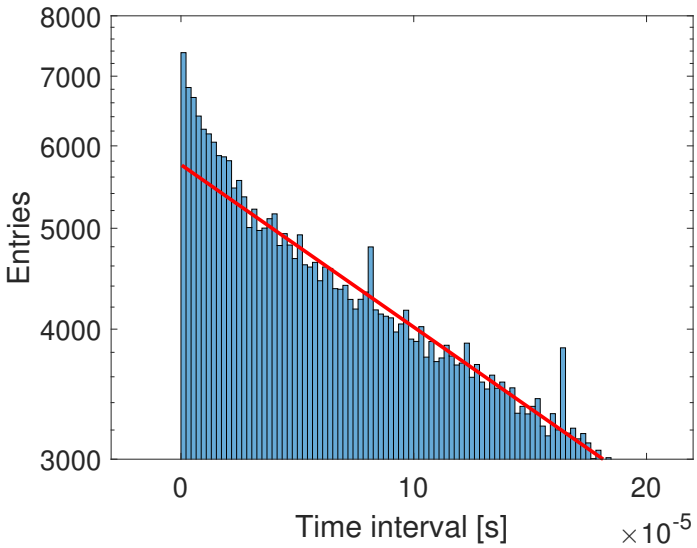


Figure 3.19: DCR median values for the cumulative curves shown in Fig. 3.18 as a function of the temperature.

nential fit. By keeping the temperature lower than $0°C$, DCR median values smaller than 250 Hz (37 kHz/mm^2) can be attained. At temperatures above $40°C$, the DCR median values significantly exceeds a few tens of kHz.
Further measurements need to be performed to accurately extract the dependence of the breakdown voltage on the temperature, as well as to obtain useful insights about the dominant noise mechanisms taking place in the devices of the ASAP110LF chip across the entire temperature range.
The DCR of pixels in A1 was also measured by using the TDC that is integrated within the ASAP110LF chip. Several time intervals, occurring between adjacent dark pulses in a single SPAD, were recorded through the on board time to digital converter. An external clock with an oscillation frequency of 25 MHz, generated by the FPGA, was used as clock reference for the counting operation of the TDC, which was internally connected to the monostable port of A1. The TDC was set in 20-bit mode, so as to allow the collection of time intervals larger than 50 μs. Before the actual DCR measurements, a characterization procedure was performed to assess the working condition of the TDC. In this case, pulses were generated, through the TEST_B signal, providing at the TDC input time intervals with duration equal to an integer multiple of the reference clock period. From the characterization process, stimulating all the possible output binary configurations of the TDC, no errors were found. Fig. 3.20 shows the distribution of the time intervals, occurring between dark events, featured by a pixel in A1. The SPAD under measurement exhibited



(a)



(b)

Figure 3.20: Distribution of the time intervals between dark noise events for pixel (1,1) with exponential fitting: (a) complete time distribution, (b) close up view of the distribution around the peak. The two distributions, obtained from the same set of data, are represented with a different number of bins.

a DCR, measured according to the time frame method described in section 3.2.4.1, of 3480 Hz (before the application of the correction formula). The time interval distribution, obtained after the acquisition of more than 7×10^5 entries, was found to follow an exponential behaviour, consistent with the Poissonian nature of the random generation of dark pulses in SPADs. Some entries, located around the peak of the distribution (Fig. 3.20b), depart from the exponential fit, thus potentially indicating the presence of afterpulsing phenomena, where secondary avalanches were generated in the depleted region, upon the release of secondary carriers. Afterpulsing phenomena decrease the distance between two noise events, thus producing a higher number of time intervals than one would expect in the case of Poissonian behavior in the first part of the distribution. In the pixel under measurement, afterpulsing involved carrier release times ranging between 40 ns, which is the minimum detectable time interval, and 30 μs. It is worth specifying that the measurement of time intervals between dark pulses, carried out by using the integrated TDC, cannot detect secondary noise events causing pulse merging, since in this case a single pulse is fed as input of the time to digital converter.
From the distribution in Fig. 3.20a, the DCR of the pixel under measurement can be extracted by considering the reciprocal of the average time interval, computed over all the entries of the distribution. The DCR obtained in this way, equal to 3590 Hz, is in fair agreement with the DCR value obtained from the discretization of the TIT, after the application of the correction formula (3550 Hz), thus demonstrating the validity of the two measurement procedures.

3.2.4.3 ASAP110LF vs APIX2LF

Fig. 3.21 shows the comparison, in terms of DCR, between the ASAP110LF chip and the APIX2LF chip. Since the SPADs integrated in the two detectors have different size, the DCR data shown in figure are normalized to the active area. The two sets of data, acquired with an excess voltage of 1.3 V, were obtained from the largest array structures integrated in the two chips (A1 of C3 for the ASAP110LF chip). The measurements relevant to the APIX2LF chip were obtained from a single layer sample. The array considered for the comparison is made of 2304 pixels (arranged in a 48×48 matrix), each of them equipped with a p+/nwell SPAD (active area of 70×52 μm^2), a front-end channel based on a passive quenching technique and a 1-bit memory. The cumulative distribution curve of the APIX2LF chip covers more than two orders of magnitude, thus indicating the presence of SPADs featuring significantly different values of DCR. In general, the DCR featured by SPADs in the

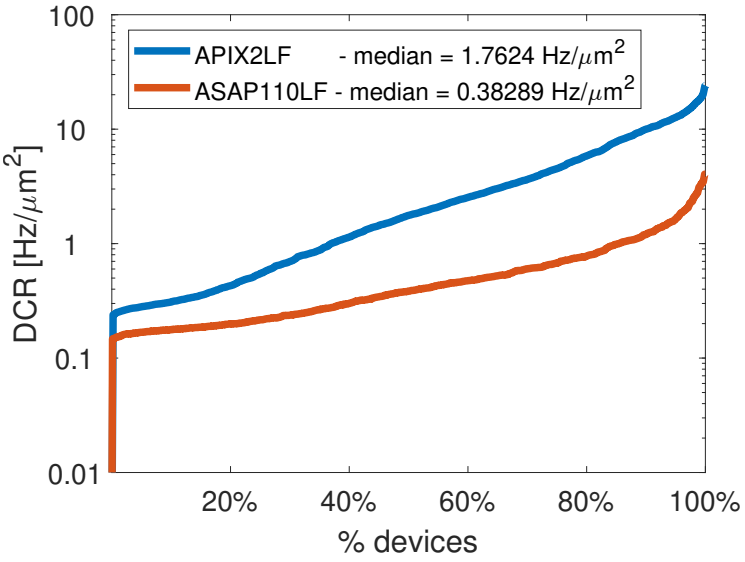


Figure 3.21: DCR cumulative distributions curves, normalized to the active area, relevant to SPADs in the ASAP110LF chip and the APIX2LF chip.

ASAP110LF chip is found to be significantly lower than the noise measured in the other chip. For the chip fabricated in the 150 nm technology node, DCR values exceeding 10^5 Hz were measured, while 27 kHz was the maximum noise value recorded for SPADs in the ASAP110LF chip. The performance improvement in terms of DCR, featured by the ASAP110LF chip, can be ascribed to the employed technology. The 110 nm CIS technology, specifically developed for the production of CMOS image sensors, may provide relatively low levels of noise, if compared to standard CMOS technologies, which may not represent the most convenient choice to achieve adequate performance in optoelectronic devices [138]. The main advantages of CIS technologies are concerned with the use of cleaner processes, obtained through the use of p-epitaxial substrates, and the enhancement of the detection efficiency, which may be attained by including antireflective films [142]. As a result, SPADs manufactured using these technologies can leverage the advantages provided by established CMOS processes, while exhibiting notable optical characteristics. As shown in Fig. 3.21 and 3.22, the DCR of the SPADs fabricated in 110 CIS technology, besides being lower than in the case of sensors in a standard CMOS technology, was also found to be more homogeneous throughout the array. As a matter of a fact, more than the 95% of the ASAP110LF cumulative distribution curve is enclosed in less than a decade. As already discussed, also the breakdown volt-

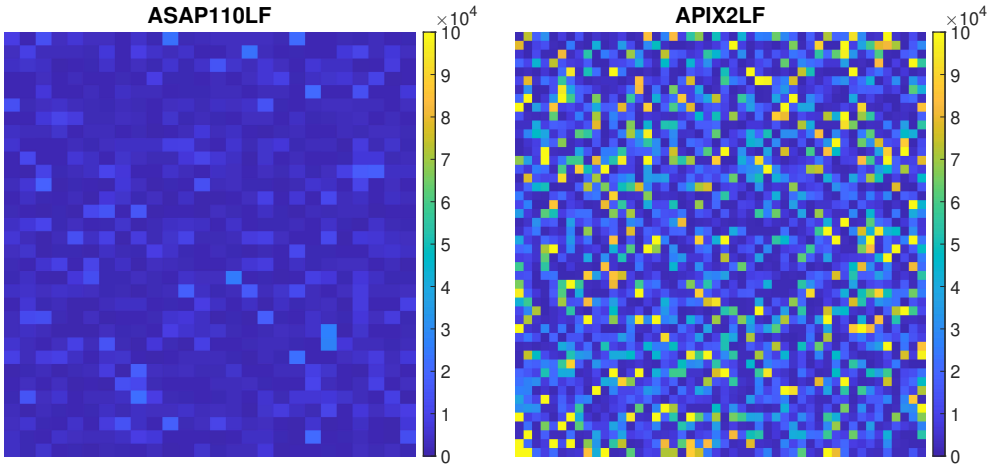


Figure 3.22: Heat maps representing the DCR featured by SPAD arrays in the ASAP110LF chip and in the APIX2LF chip.

age features better uniformity in SPAD arrays produced in CIS technology. Standard deviations from 7 to 12 mV were measured on the V_{BD} of arrays in the ASAP110LF chip (Fig. 3.8), while values ranging from 20 to 32 mV were recorded for SPADs fabricated in standard 150 nm CMOS technology [7].
Fig. 3.23 shows an estimation of the DCR which could be obtained if a dual layer chip was fabricated by vertically interconnecting two ASAP110LF chip layers. The cumulative distribution curves measured on two single layer chips (array A1 in C3 and C5) are also included in the figure. The two curves were obtained at a SPAD bias voltage of 19.8 V, thus corresponding, in both the arrays, to an excess voltage of 1.3 V. The same bias voltage was applied to SPADs in the two chips, as in the case they were vertically connected to form a dual layer detection system. The dual layer DCR estimation curves, represented for three different coincidence windows, were derived by applying (1.31) to pairs of SPADs potentially forming a dual layer pixel. The accuracy of the DCR model expressed in (1.31) was already demonstrated in [5] and [121], where actual DCR measurements on dual layer chips were found to be in good agreement with the estimated dark noise. To attain the dual layer solid curves shown in the figure, SPADs were matched according to their position in the array. A series of Monte Carlo runs, performing randomly pairing between SPADs from the two different chips, yielded a set of noise distribution curves featuring negligible difference among each other. In the figure, the cumulative distribution from DCR measurements performed on a dual layer APIX2LF chip is included, for the sake of comparison. The curve shown in the

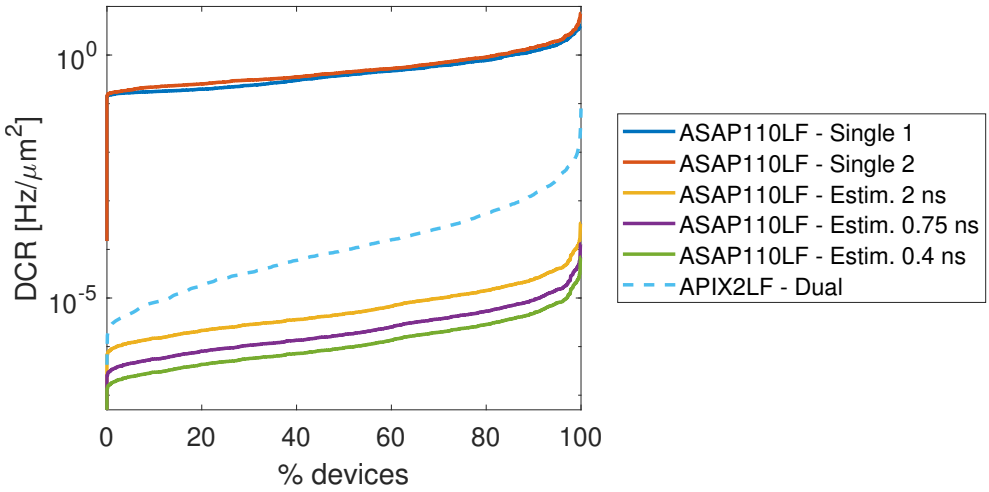


Figure 3.23: DCR cumulative distribution curves showing the estimated DCR featured by a dual layer ASAP110LF chip and the noise performance measured on a dual layer APIX2LF chip (the noise curves are normalized to the relevant active areas).

figure was obtained with a coincidence window of 2 ns and an excess voltage of 1.3 V. The DCR which can be attained with a dual layer sensor based on the ASAP110LF chip is remarkably lower than that featured by the APIX2LF chip. More than one order of magnitude can be detected between the dashed curve and the yellow solid one, which remains below 10^{-4} $Hz/\mu m^2$ until the 99% of the pixels is reached. Towards the right end of the plot, the two distributions differ by two decades.
In applications involving recent upgrades to the CMS tracker [143], considered as a valid reference to classify the noise performance of the coincidence-based detection system described in this work, the maximum noise hit rate allowed by the system is 1.6 $mHz/\mu m^2$. This value, set as noise requirement for the specific application, is two orders of magnitude higher than 1.13×10^{-5} $Hz/\mu m^2$, representing the DCR mean value estimated for a dual layer sample of the ASAP110LF chip (A1 array, $\delta t = 2$ ns).
In [7], the results from DCR measurements at different temperatures, performed on an APIX2 dual layer sample, are shown. As expected, the noise performance of dual layer sensors, in the explored range of temperatures ($-20°C$ to $50°C$), was found to adhere to the DCR model in (1.31).

3.2.4.4 RTS measurements

Defects within the bulk material, or at interfaces, pose significant challenges for solid-state sensors with small areas. The manufacturing processes, involving steps such as etching or implantation, often introduce defects and lattice impurities, which may degrade the sensor performance by increasing the dark current. The latter may fluctuate between multiple discrete levels, leading to the phenomenon known as Random Telegraph Signal (RTS) [144][145]. This mechanism can significantly impact applications requiring low and stable noise over long integration times, since calibration errors can be introduced. In addition, RTS tends to escalate in radiation environments, where both ionizing and non-ionizing radiation may be responsible for increasing the dark current and triggering fluctuations of the random telegraph signal kind, thereby undermining sensor performance [146][147]. Comprehending the underlying mechanisms is crucial for devising strategies to mitigate RTS effects. Despite RTS being observed since many years, the fundamental causes remain uncertain. Different factors may give rise to the RTS phenomenon, including bulk Shockley-Read-Hall (SRH) generation centers and defects at the oxide interface. The discrete levels of RTS centers represent stable configurations where carriers are generated at a particular rate. If a pixel contains a single center, the number of levels is equal to the number of visible configurations, while the amplitude discrepancy between levels indicates differences in the generation rates of the center. In the case of two-level RTS, the time intervals spent in the two states are exponentially distributed, and the switching between the two states is described by a Poisson process [148]. However, RTS events in SPADs may be the result of a superposition of different levels originating from multilevel sources, independent bi-level sources, or a combination of both [149]. If two-level RTSs are considered, the following parameters are mainly of interest:

- τ_{up}, representing the average time spent in the high level;
- τ_{down}, representing the average time spent in the low level;
- A_{RTS}, which is the amplitude of switching between the high and low levels.

The dark noise of all the pixels of A1 in C3 was monitored over a time span of approximately five days, inside a climatic chamber at $25°C$. The DCR was acquired in loop, for all the SPADs in the array, with a time resolution of 40 seconds. The measurement, carried out through the measurement procedure

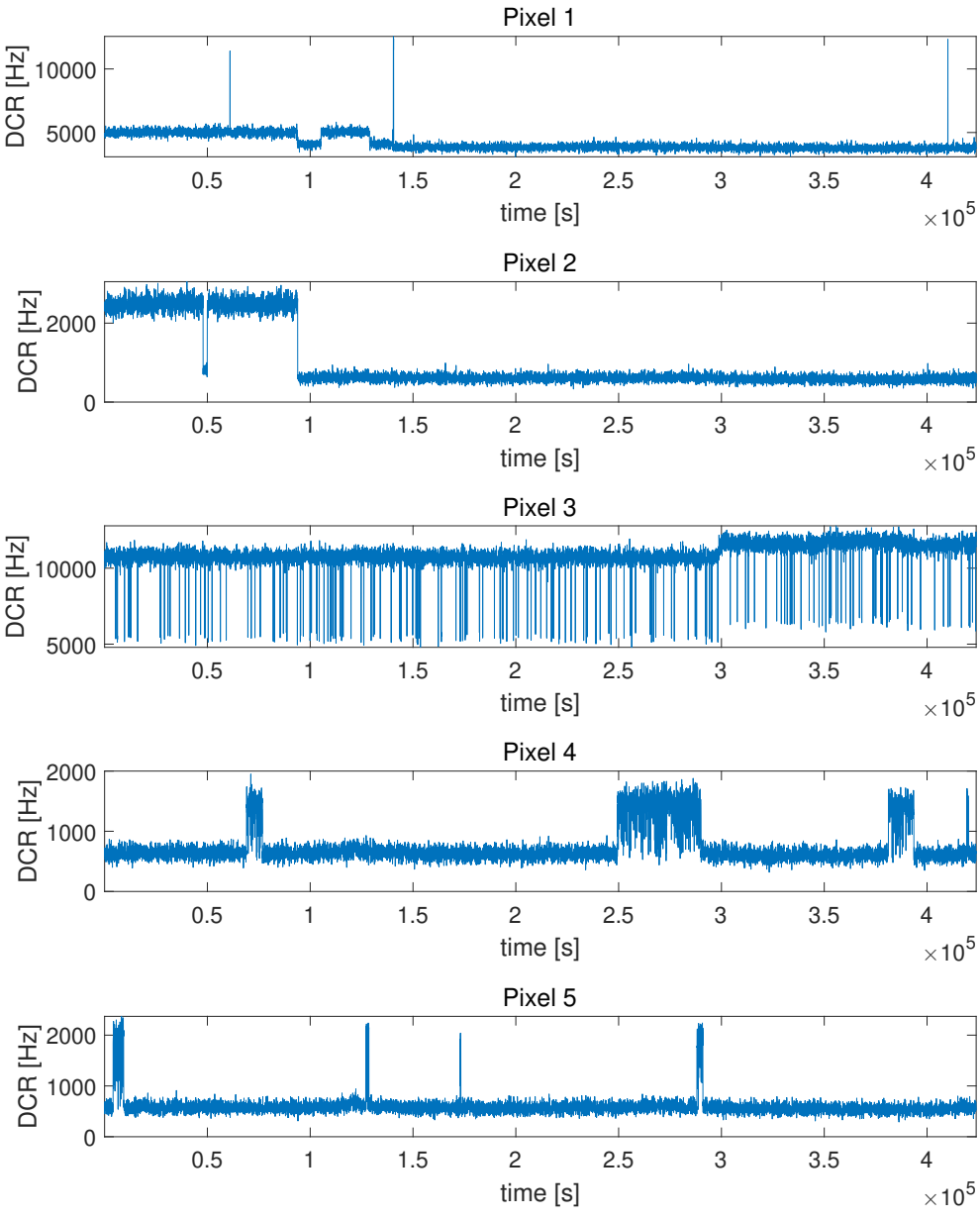


Figure 3.24: DCR vs time in pixels of A1 exhibiting RTS behavior.

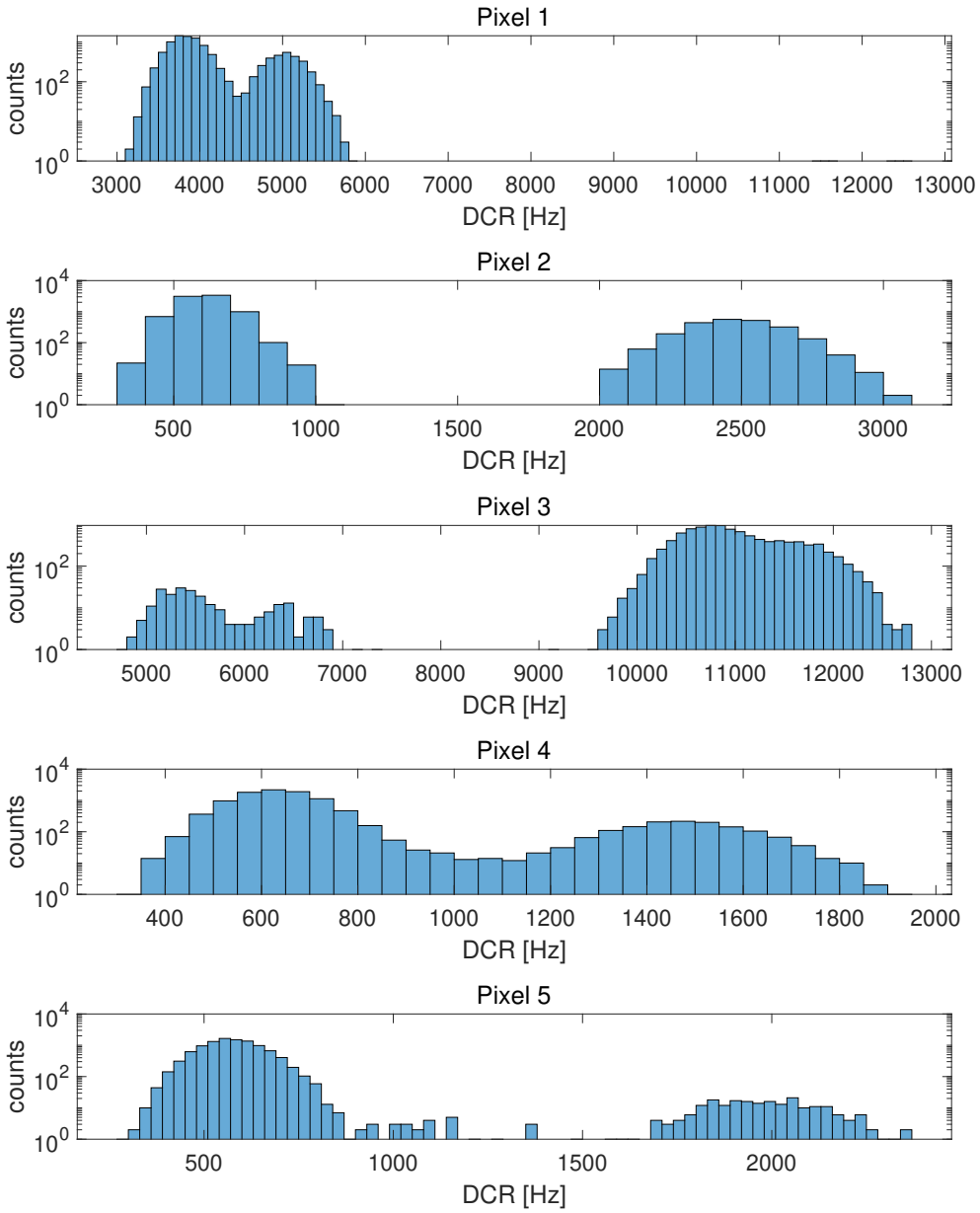


Figure 3.25: DCR distribution for pixels exhibiting RTS behavior.

discussed in 3.2.4.1, was performed at a fixed excess voltage of 1 V. A remarkable effort was essential to ascertain whether a SPAD exhibits RTS behavior and to establish the number of levels per pixel. Methodologies, including visual counting [150] and histogram analysis [151] were adopted to identify the RTS characteristics. An automated algorithm integrating the aforementioned techniques, along with threshold-based methods, statistical analysis of properties, and a step-shaped filter, is currently under development. As the sampling time was not sufficiently short to conduct a detailed characterization of RTS and its characteristic times, the results that will be shown in the following must be considered as a preliminary overview of the RTS features of the devices under test. Thorough measurements, targeting specific pixels and using a higher time resolution and a larger time span, need to be carried out to delve deeper into RTS features by means of more advanced analysis techniques.
Fig. 3.24 shows the DCR as a function of time for five pixels featuring RTS. In all the time diagrams, except for the ones relevant to pixel 1 and 3, two different levels of DCR can be observed. In the case of pixel 3, approximately at $t = 3 \times 10^5$ s, the entire DCR characteristics exhibits a slight upward shift, thus generating two additional noise levels. As a matter of a fact, four RTS levels can be detected in the DCR of the pixel under discussion, even though its RTS behaviour can be considered as piecewise bi-level, since the two pairs of RTS levels are clearly separated in time. In the diagram relevant to pixel 1, a third level, featuring a very small number of transitions and fast characteristic times, can be detected. Relatively fast transition times, in the order of a few minutes, can be observed for pixel 3, while the other pixels show longer transition times, in the order of several hours. Very likely, a larger time span may be required to perform an accurate analysis of the time constants associated with the levels of pixels 1,2,4 and 5, since a small number of transitions was collected during the acquisition time window. In Fig. 3.25, the DCR distributions for the pixels shown in Fig. 3.24 are depicted. The Y-axis of the diagrams is logarithmically scaled to help distinguish the different peaks of the distributions. In the case of a two-level signal, each level is represented by a Gaussian distribution, with the peaks corresponding to the mean values of the high and low levels of the RTS fluctuation. The distance between the two peaks denotes the RTS amplitude. For pixels 1 and 4 the two main distributions are not well separated, indicating that the two RTS levels share some DCR values. The distribution relevant to pixel 3 confirms that more than two levels can be detected in the DCR of this SPAD.
In the case of a bi-level RTS pixel, the extraction of the time constant associated with each level is straightforward, as it corresponds to the average time

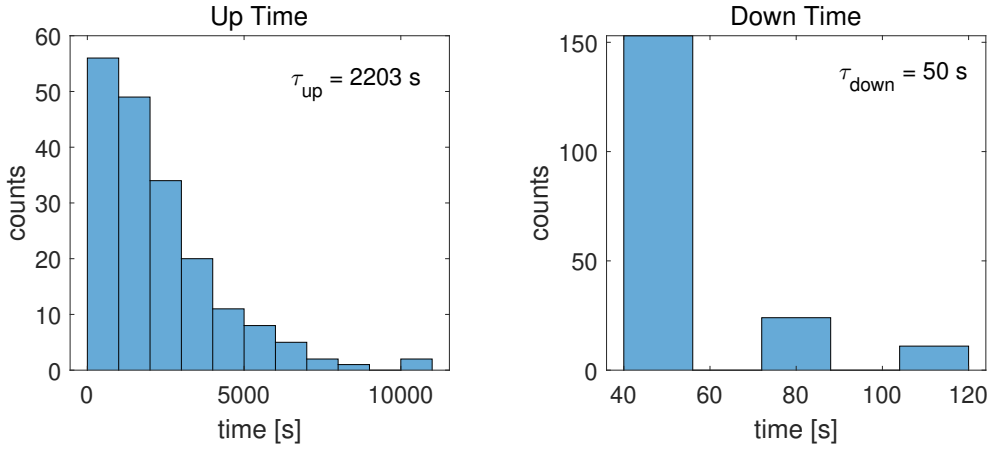


Figure 3.26: Up and down time distributions for a pixel exhibiting RTS behavior.

spent in the specific level. Fig. 3.26 shows the distribution of time intervals spent in the lower level (down time) and in the upper level (up time) for the sensor indicated as pixel 3 in Fig. 3.24 and 3.25. Despite not showing a pure bi-level RTS characteristic, pixel 3 was selected for this analysis, due the high number of transitions shown in the acquisition time window. However, only DCR entries before $t = 3 \times 10^5$ s were considered, so as to deal with a bi-level RTS behaviour. In the diagrams, the time constants associated to the two levels are also reported. Because of the coarse time resolution employed for the RTS measurement, only three distinct intervals, each a multiple of the minimum measurable interval (40 s), were recorded for the down time. The diagrams illustrate that the time intervals follow an exponential distribution, as expected for a Poissonian process.
The RTS phenomenon may, in part, originate from the shallow trench isolations, which are integrated in the structures under characterization to ensure isolation between adjacent sensors. Defects at the STI-silicon interface may inject free carriers from their surface to the depletion region, thus increasing the probability of observing RTS behaviour [152]. This detrimental effect may be also enhanced in SPADs with reduced active area [7].
The RTS phenomenon is expected to be strongly dependent on temperature. The latter plays a crucial role in the performance of a semiconductor detector since it affects the mechanism of carrier generation inside the device. A temperature rise may lead to a higher probability of DCR switching and an increased RTS amplitude, as already demonstrated in [7]. Conversely, at lower

temperatures, the DCR switching frequency may decrease, and RTS phenomena could even disappear for certain pixels. Further research, involving measurements at different temperatures and SPAD bias voltages, is underway to investigate these phenomena more thoroughly and gain a deeper understanding of their behavior.
From these preliminary measurements, only five pixels, representing the 0.6% of the total number of cells, were found to be affected by RTS phenomena in DCR. The number of pixels showing RTS behaviour is significantly reduced if compared to the APIX2LF chip [7][153], where a non negligible fraction of SPADs (around 10%, with slight differences from one chip to the other) feature a DCR fluctuating between two, three and four levels. The use of a CIS technology, allowing the implementation of SPAD junctions with remarkably lower defect density as compared to standard technologies, has proven to be effective in providing sensors with better noise performance. However, future measurements on pixels of the ASAP110LF chip, monitoring the DCR of the entire array within a larger time window, may reveal other pixels featuring RTS levels with larger characteristic times.

3.2.4.5 Crosstalk measurements

Crosstalk measurements were performed on the ASAP110LF chip to study how the generation of dark pulses in a single pixel may affect the noise performance of the neighbouring ones. The crosstalk depends on factors like the DCR of individual sensors [43], the substrate thickness [42], the technology process and the layout of the array. Reducing crosstalk is of paramount importance to preserve the resolution of a digital detection system, as well as to improve the noise performance of the entire structure. During a crosstalk measurement, the DCR of each pixel was measured, while all the other pixels in the array were kept enabled. Therefore, all the sensors in the array were left free to produce dark pulses, according to their own dark event generation rate.
Fig. 3.27 shows the cumulative distribution curve from a crosstalk measurement performed on pixels of A1 in C3. For the sake of comparison, the cumulative distribution curve obtained from an individual DCR measurement (the curve with $V_{SPAD} = 19.9\ V$, shown in Fig. 3.17a), performed by means of the measurement procedure described in section 3.2.4.1, is included in figure. The distribution taking into account the crosstalk effects partially overlaps with the blue curve. The dark count rate of approximately 60% of the SPADs in the array was found to be systematically lower during the crosstalk measurements rather than in individual DCR measurements. This effect may be

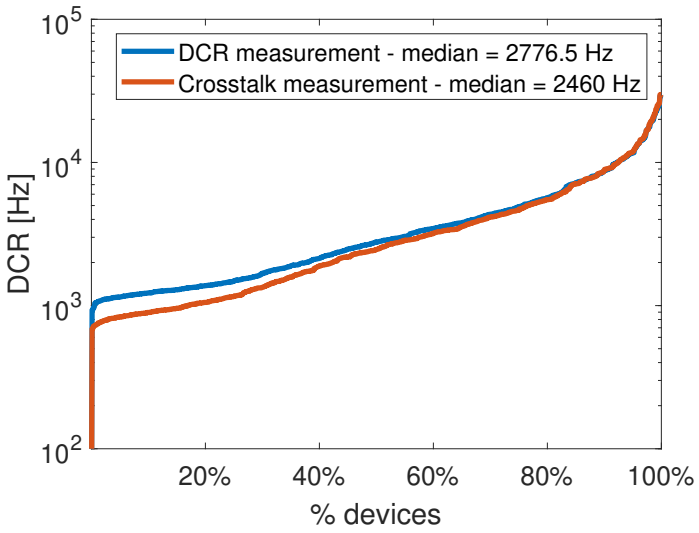


Figure 3.27: DCR cumulative distribution curves relevant to a crosstalk measurement performed in the ASAP110LF chip (A1 in C3) and an individual DCR measurement from the same chip.

ascribed to a non negligible voltage drop on the SPAD bias voltage line, taking place during the crosstalk characterization. In this kind of measurements, all the sensors in the array were concurrently enabled to produce dark pulses, thus demanding for a larger amount of current to flow through the V_{SPAD} voltage grid, as compared to the measurements where one single pixel is enabled at a time. As a consequence, the apparent voltage breakdown gradient, observed in Fig 3.9, may have been further amplified and the SPADs located in the bottom right corner of the chip may be biased with a supply voltage which was non-negligibly lower than the nominal one. From the cumulative distributions in the figure, it can be inferred that the ASAP110LF chip turned out to be completely screamer-free, as the crosstalk curve never exceeds, in a significant way, the blue one. In addition, at high percentage of pixels, the two curves overlap, thus indicating that the DCR of noisy pixels is not affected by crosstalk effects. The minimal susceptibility to crosstalk, exhibited by SPADs in 110 nm CIS technology, may be ascribed to the lack of particularly noisy pixels affecting the DCR of the entire array. This result allows for improved noise performance, while preserving the overall detection efficiency of the system, since no screaming pixels need to be selectively switched off.
For the sake of comparison, the crosstalk performance of the APIX2LF chip is shown in Fig. 3.28. For SPADs fabricated in 150 nm CMOS technology, a sig-

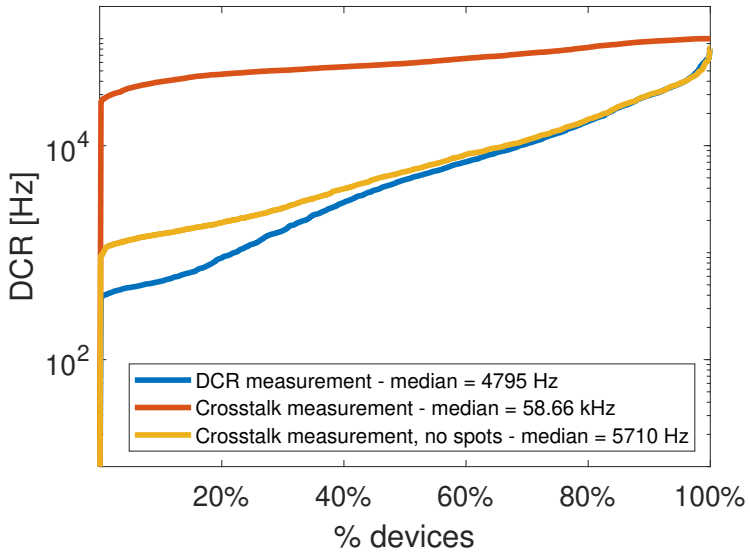


Figure 3.28: DCR cumulative distribution curves describing the crosstalk performance of the APIX2LF chip.

nificant noise degradation can be observed due to the effect of crosstalk. The DCR obtained from crosstalk measurements (red curve) was found to be remarkably higher than the one acquired through individual DCR measurements (blue curve). The third curve presented in figure ("no spots") resulted from a crosstalk measurement that was performed after disabling a group of notably noisy pixels. The exact location of the excluded cells, indicated as screamers, was obtained by inspection from the original crosstalk measurement. During the measurement procedure leading to the yellow curve, nine screamer pixels out of 2304 were disabled, thus preventing them from producing dark pulses, while the DCR of the other pixels in the array was being evaluated. The small difference between the individual DCR distribution and the "no spots" curve suggests that most of the crosstalk contribution was provided by the identified screamer pixels, significantly affecting the DCR of all the other SPADs in the array.

3.2.5 Digital SiPM measurements

The first prototype of digital SiPM, based on a parallel counter architecture, was characterized to test the effectiveness of the readout circuit shown in Fig. 2.24. The results shown in this section have to be considered as preliminary. A complete measurement campaign, investigating the performance of

all the SiPM structures in the array, needs to be planned to assess the advantage of using a parallel counter in the readout of a digital SiPM. The characterization, which was performed on a single SiPM in C4, was carried out by selectively enabling the front-end of a specified number of subpixels. All the SPADs in the SiPM under test were prevented from producing dark pulses by setting their cathode voltage to 5 V, which is well below the breakdown voltage. The SPAD firing was emulated through the assertion of the TEST_B signal. Upon each negative pulse of the TEST_B signal, all the enabled frontend circuits produce a pulse, with a fixed duration, propagating to the input terminals of the parallel counter. Since the TEST_B signal is provided globally to all the subpixels in the SiPM, the pulses processed by the parallel counter are simultaneously generated by the enabled front-end circuits. The entire procedure was repeated for all the 2^{16} combinations of the enabled pixels, thus covering all the potential bit configurations which can be fed to the 16-bit input parallel counter. During the measurements leading to the readout circuit characterization, a SiPM threshold of 1 was used. In order to accomplish the results shown in this section, an FPGA program properly driving the TEST_B signal was developed. Before providing the test pulse, setting specific inputs of the parallel counter, the subpixel enabling procedure is performed. The latter is managed by a 16-bit word, stored in an FPGA register, that is incremented before each pulse of the TEST_B signal. A one-to-one correspondence between the 16 bits of the FPGA word and the 16 subpixels of the SiPM under measurement was determined in the FPGA

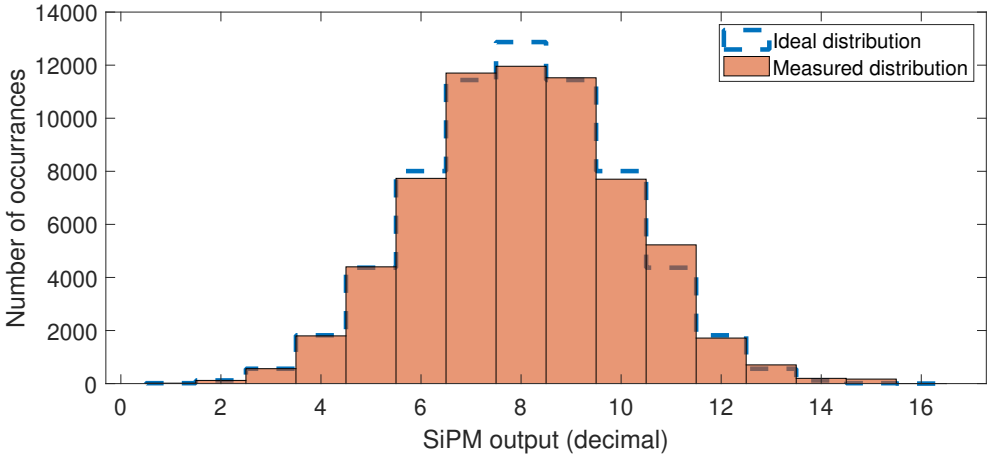


Figure 3.29: Distribution of the binary configurations read at the output of a SiPM in C4, during the characterization of the digital readout circuit.

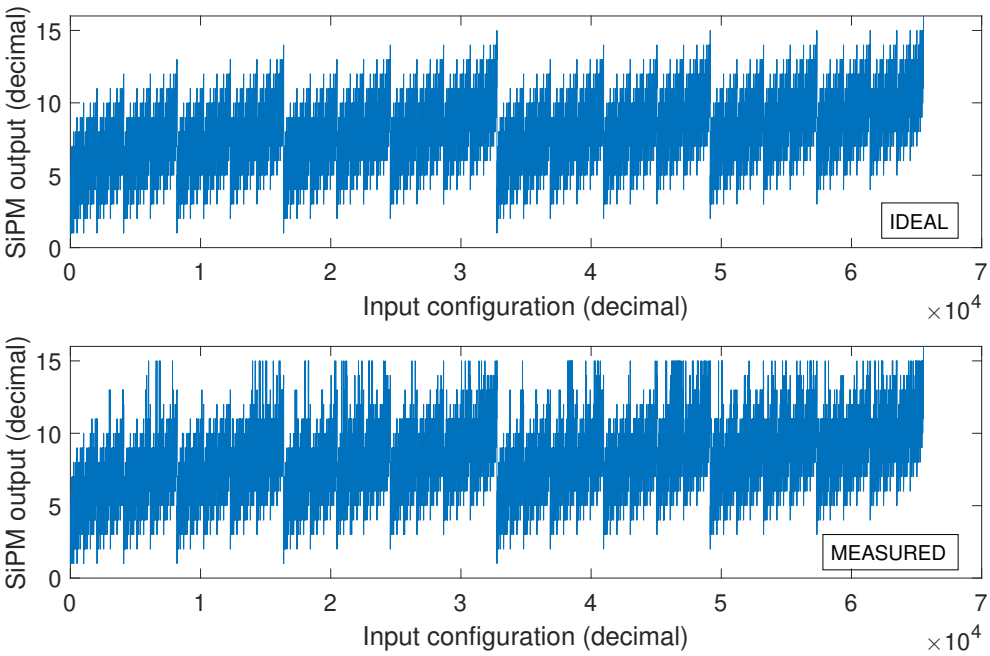


Figure 3.30: Input-output characteristics of an ideal SiPM (top) and of a SiPM contained in C4 (bottom).

algorithm. The value of each bit (1 or 0) in the 16-bit word represents the enable status of the corresponding subpixel. All the different combinations of enabled subpixels can be tested by making the 16-bit FPGA word vary from 1 to 65535.
The distribution of the binary configurations read at the output of the SiPM, during the characterization procedure, is shown in Fig. 3.29. In the same diagram, the outline of the histogram that should be obtained in the case of a fault-free SiPM is represented with a dashed line. Fig. 3.30 shows the comparison between the ideal input-output characteristics and the measured one. In these diagram, the x-axis represents the 16-bit binary word (in decimal) determining the different combinations of concurrently enabled pixels. Each digit of the 16-bits identifies a specific subpixel in the 4×4 array. The number of enabled pixels is equal to the number of bits at 1 contained into the binary word. Assuming no faults in the front-end circuits of the subpixels, the 16-bits binary words, represented in the x-axis, can be considered also as the binary configurations provided at the input of the parallel counter, during

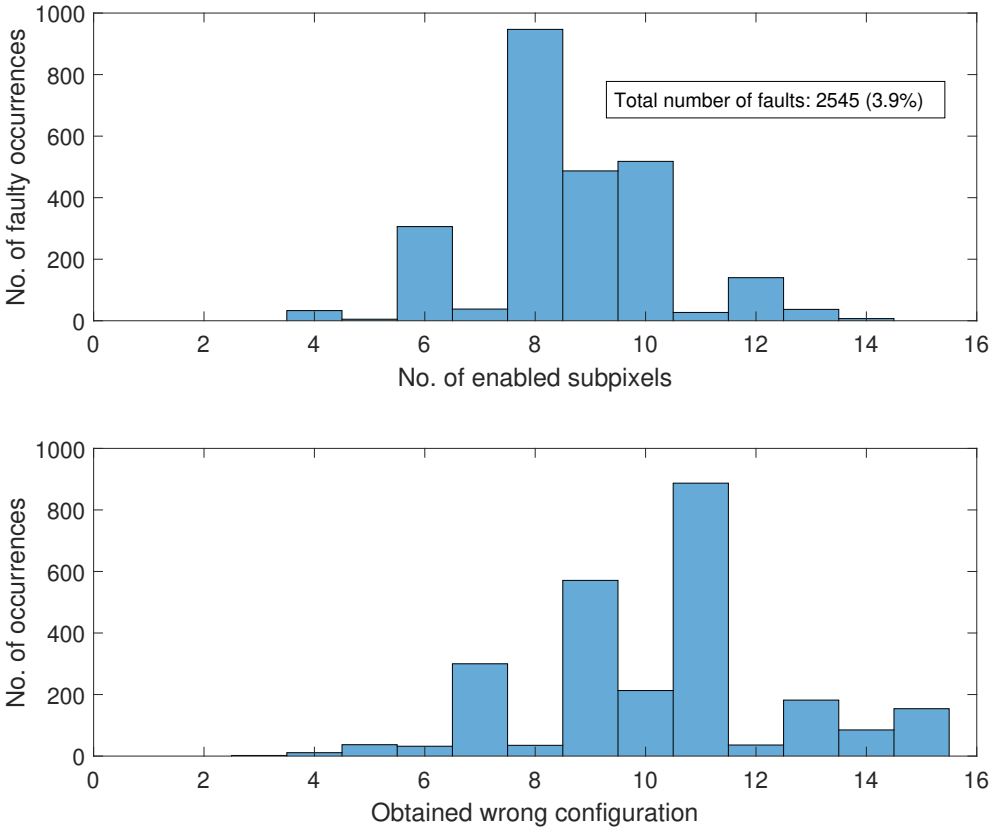


Figure 3.31: Number of wrong output configurations as a function of the number of enabled subpixels (top) and distribution of the wrong output configurations (bottom).

each step of the characterization. The non-zero difference between the two distributions in Fig. 3.29, as well as the non-regular spikes featured by the measured input-output characteristics of Fig. 3.30, indicates that wrong values were detected at the output of the SiPM. The number of wrong output configurations as a function of the number of enabled subpixels is shown in Fig. 3.31. The distribution of the wrong binary words read at the output of the SiPM is also included. From the characterization, 2545 input configurations, out of 2^{16} (3.9%), returned a wrong value. Repeated measurements revealed that the errors are systematic, thus indicating a design-related malfunctioning in the processing circuit. Due to statistical reasons, the highest number of errors occurs when 8 subpixels are concurrently enabled, as this value has the

greatest number of occurrences among the 2^{16} possible configurations. The distributions of the wrong output configurations, separated according to the number of concurrently enabled pixels, is shown in Fig. 3.32. The value in brackets, reported within each diagram in the figure, represents the percentage of errors that was obtained for each number of enabled subpixels.
The main source of error, producing wrong output values, may be identified in the generation of residual count glitches, which are not completely filtered out by the auto triggering mechanism. Depending on the amplitude and on the time duration of the spurious pulses, binary output words in excess of or below the expected one can be generated. Timing violations on the flip-flops, due to count glitches featuring a duration similar to the delay in Fig. 2.24, may prevent the memory elements from settling to the correct value. If the setup and hold times of the flip-flops are not met, the output of the memory blocks becomes unpredictable, thus potentially evolving into a wrong value after a metastability time interval having non-deterministic duration. Due to timing violations on the memory elements, output words above or below the expected one can be produced.
Spurious pulses, featuring a time duration larger than the delay of the auto-triggering network, can incorrectly set the output flip-flops. In this case, a logic 1 is stored in place of an actual 0, thus contributing with an additive effect to the generation of wrong binary words. By inspection of the diagrams in Fig. 3.32, it can be inferred that most of the wrong output words are in excess of the relevant expected values, since, in this case, both the effects from timing violations on the output flip-flops and the generation of large count glitches sum up. Moreover, most of the errors are concentrated within one or two units around the expected one, thus indicating bits 0 and 1 as the main source of glitches. In particular, in the case of 8 ('b01000) and 12 ('b01100) concurrently enabled subpixels, the contributions of count glitches, affecting the two least significant bits, lead to the highest percentage of error. As shown in the diagrams, all the even binary configurations have the maximum of the error distribution located in the subsequent odd binary number, thus indicating that the strongest glitch contribution, between the two LSBs, comes from bit 0. The detrimental effect of glitches on this bit can be noted also in the bottom diagram of Fig. 3.31, where the wrong binary words featuring the highest number of occurrences are odd numbers.
Glitches occurring on bits with binary weights exceeding 1 were effectively filtered out by the auto-triggering network. As shown in the diagrams in Fig. 3.32, the error distributions are sufficiently narrow around the expected binary word, with sporadic wrong values differing by 3 from the correct one.

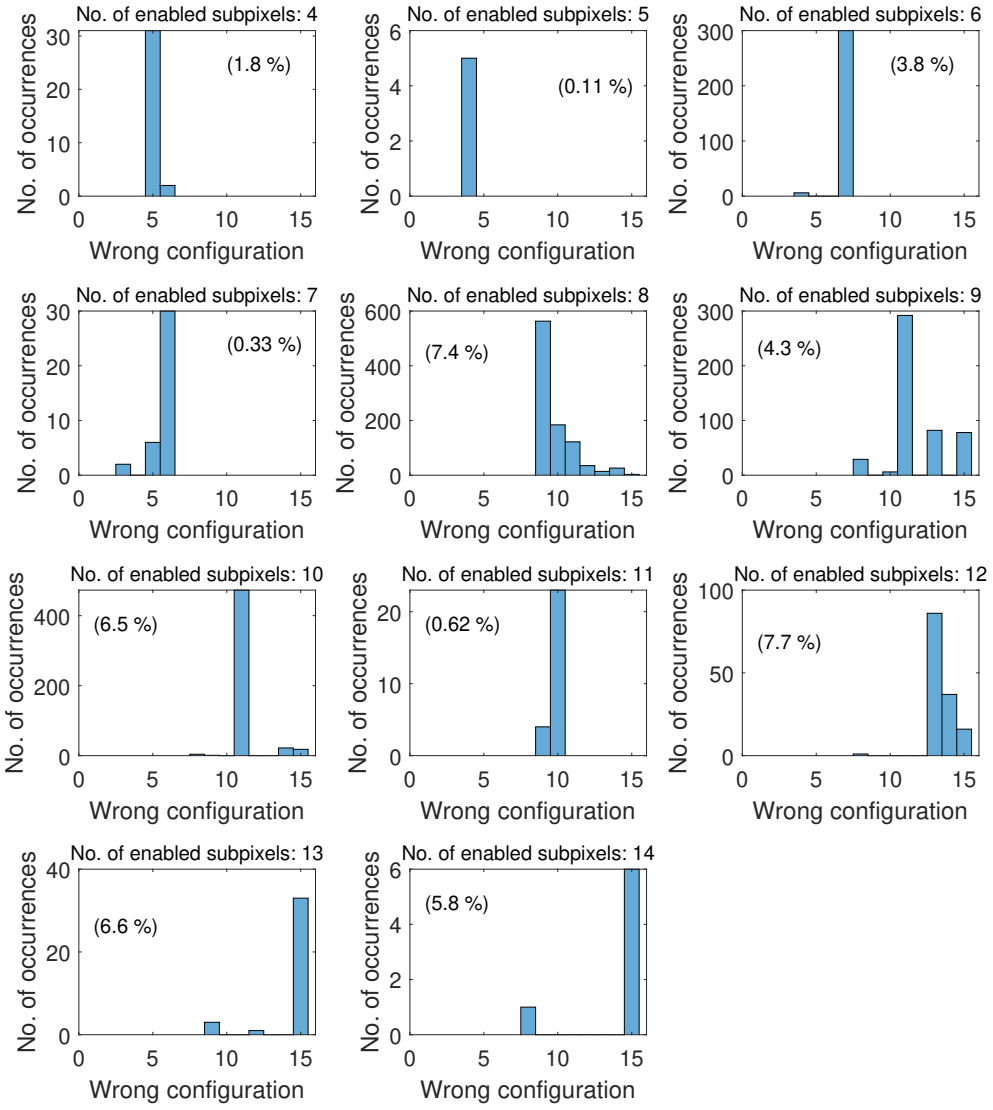


Figure 3.32: Distribution of the wrong output configurations for each number of enabled subpixels. Fault-free input configurations were not included.

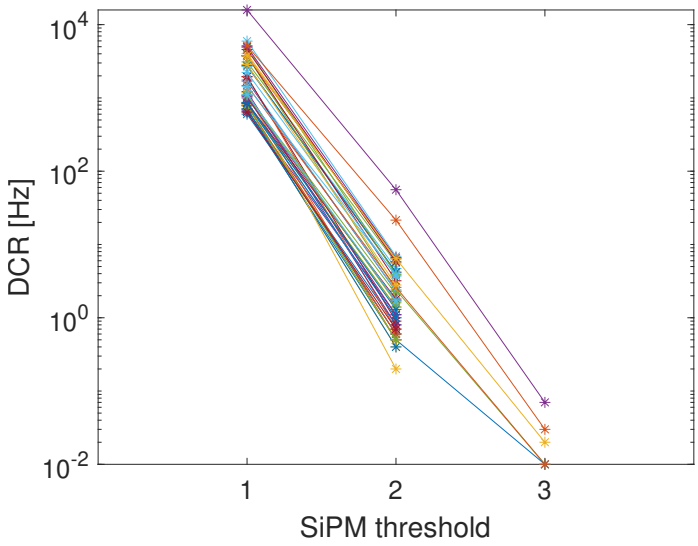


Figure 3.33: DCR of all the SiPM structures in C5, as a function of the threshold value. Data were collected through the SIPMSOT pad.

However, it is worth specifying that a relatively low number of spurious pulses was expected on high-weighted bits, as already described in Fig. 2.32.
Fig. 3.33 shows the results from a DCR measurement performed, at different threshold values, on all the pixels of the ASIPM array in C5. The measurement, carried out at $V_{SPAD} = 19.5\ V$, was performed pixel by pixel by enabling all the subpixels in a single SiPM. The SOT bit, revealing whether a number of SPADs higher than the selected threshold value was triggered, was read on the dedicated output pad. The DCR measurement procedure, already discussed in section 3.2.4.1, was used to extract the set of data shown in the figure. For SiPM threshold values exceeding 1, a TIT ranging between 10 s and 100 s was used to allow the collection of noise rates less than 10 Hz. Due to the extremely low probability of having four concurrent dark events in a time window of 2 ns (which is the duration of the monostable pulse generated by the front-end circuit of each subpixel), no SOT signal was detected, with the used TIT, after setting the SiPM threshold to 4. For the same reason, most of the SiPMs were non responsive while the threshold value was set to 3. As expected from the counting statistics of random events, the DCR of the SiPM was found to scale exponentially with the threshold value. In the structure under consideration, the coincidence DCR as a function of the SiPM

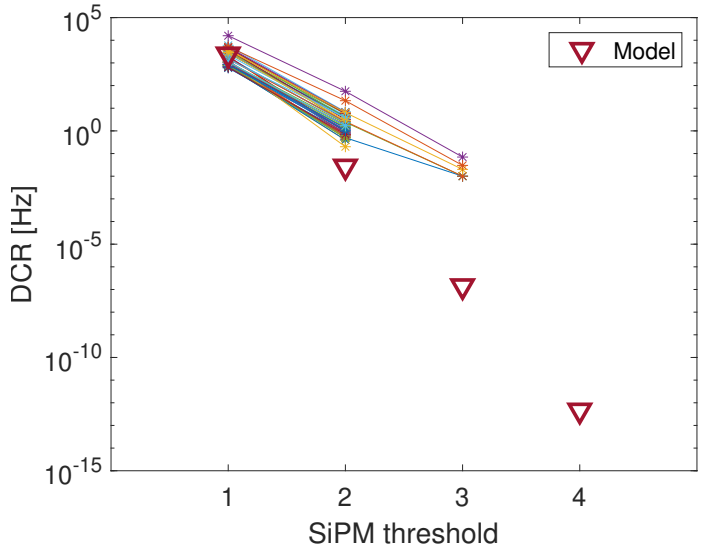


Figure 3.34: Comparison between the theoretical model, expressing the SiPM DCR as a function of the threshold value, and the experimental noise data, collected through the SIPMSOT pad.

threshold can be expressed through the following model:

$$DCR_{th} = \sum_{j=1}^{\binom{16}{n_{th}}} n_{th} \left(\prod_{k=1}^{n_{th}} DCR_k \right) \delta t^{n_{th}-1}, \tag{3.7}$$

where n_{th} is the selected threshold value, DCR_k is the DCR of the k^{th} subpixel and δt is the duration of the pulse generated by the front-end circuits. In Fig. 3.34, the results from the DCR measurements are compared with the theoretical model. An average subpixel DCR of 100 Hz and a δt of 2 ns were assumed for the development of the model in figure. It can be noted as, for $n_{th} = 1$, the measured DCR was found to perfectly adhere to the theoretical model, which reduces to the sum of 16 DCR values. For threshold values in excess to 1, the experimental data significantly differ from the model, which reports lower DCR levels. This discrepancy, occurring for threshold values larger than 1, may be due to the effect of low energy spurious pulses, that are successfully filtered out by the auto-triggering networks, but may affect the generation of the SOT bit. It is worth specifying that the combinational network producing the SOT bit is directly connected to the output bits of the parallel counter, thus potentially being the circuit mostly impacted by the oc-

currence of spurious pulses. Further measurements, studying the dependence of the SOT bit on the residual count glitches generated by the parallel counter, need to be carried out in a dedicated characterization procedure.

Conclusions

Summary

This dissertation presented the design of a test platform supporting the characterization of CMOS SPADs in a 110 nm CIS technology. A new test chip and a complete measurement setup were designed in the framework of the ASAP project, whose main purpose was the development of low-noise position-sensitive sensors, targeting applications to charged particle tracking.
The new chip, referred to as ASAP110LF chip, integrates a number of array structures differing in the readout approach, the pixel geometrical features and the sensor type. The arrangement of sensors in arrays of pixels makes it possible to study the performance of SPADs integrated in a complete detection system, as well as to provide a statistically meaningful set of measurement data. Each pixel of the arrays is equipped with a front-end circuit made of a passive or active quenching network and a readout chain performing pixel selection and signal gating. Some arrays include also memory elements to store the information relevant to single or multiple detection events.
The SPADs integrated in all the array structures consist of p+/nwell junctions, isolated from the substrate by means of a deep nwell. Single SPADs, making the cathode and anode terminals available on dedicated pads, were also included in the chip, so as to enable a direct extraction of the I-V curves. A time to digital converter (TDC), used to measure the time intervals between adjacent pulses for dark count rate (DCR) measurement and after-pulsing characterization, was added to the chip design.
Part of the chip is devoted to an array of digital Silicon Photomultipliers (dSiPMs), which were designed by using a novel architecture based on parallel counters. A readout network, made of a parallel counter and a memory circuit, is used to count the number of simultaneously firing subpixels and store the result in dedicated flip-flops. The combinational logic network implemented through a parallel counter works with a very low latency time (a few nanoseconds), thus enabling virtually real time photon counting.
An automatic measurement setup was designed to accomplish the characterization of the ASAP110LF chip. Two custom boards were developed to connect the chip to the measurement setup. The digital signals needed for the correct chip operation were generated through an FPGA, implementing all the algorithms required by the different measurements.
SPADs from four chip samples were characterized in terms of breakdown volt-

age, photon detection probability (PDP) and DCR. The mean value of the breakdown voltage, for sensors located in different arrays and in different chips, was found to be around 18.5 V, with $\sigma \leq 12\ mV$ for pixels in the same array. A maximum PDP of 21% at $\lambda = 450\ nm$ was found with an excess voltage of 1 V, according to similar SPAD structures found in the literature.
DCR measurements at different temperatures and SPAD cathode voltages were carried out on different arrays. At $25°C$, by using an excess voltage of 1 V, median values around 0.30 $Hz/\mu m^2$ were measured in the array structures of all the chips under characterization. At temperatures lower than $0°C$, the production of dark pulses may be ascribed to band to band tunneling, probably representing the primary mechanism contributing to carrier generation in the active region. As expected, the DCR was found to linearly scale with the SPAD cathode voltage.
The SPADs integrated in the ASAP110LF chip feature overall lower DCR than sensors produced in a commercial 150 nm CMOS technology, thus demonstrating the effectiveness of a CIS technology in producing sensors with improved noise performance.
The onboard TDC was used to measure the time intervals between adjacent dark pulses, occurring in a pixel of the A1 array. The distribution of the time intervals was found to follow an exponential behavior, consistent with the Poissonian nature of the random generation of dark pulses in SPADs.
The DCR of a small percentage of pixels (0.6%) in the A1 array was found to be affected by fluctuations of the RTS (random telegraph signal) type. The measurement was performed by continuously evaluating the DCR of all the pixels in the array over an acquisition window of 5 days, with a time resolution of 40 s and an excess voltage of 1 V.
Unlike SPADs fabricated in 150 nm CMOS technology, the DCR of sensors in the ASAP110LF chip, was found to be minimally affected by crosstalk effects. The robustness against crosstalk may be ascribed to the lack of pixels featuring significantly high noise values.
A preliminary input-output characteristic of a dSiPM prototype was extracted by triggering all the 65535 possible combinations of simultaneously firing subpixels. During the characterization, systematic errors (3.9%) were produced by the structure, probably due to residual count glitches, not successfully filtered out by the readout network. Further measurements are needed to get more insights about the effectiveness of the new architecture, and to assess the working condition of the parallel counter integrated in the SiPM structure.

Future perspectives

Most of the experimental data presented in this thesis work represent preliminary results from the structures integrated in the ASAP110LF chip. Further measurements exploiting the capabilities of the presented chip need to be carried out in a dedicated characterization campaign. Some of the future activities, that may significantly enhance the scientific contribution of this PhD thesis, are listed in the following:

- The DCR measurements must be extended to all the SPADs in the chip, and to all the available chip samples, by using different combinations of SPAD cathode voltages and temperatures. These measurements may provide useful insights about the carrier generation mechanisms taking place in the sensors under analysis, as well as identify a convenient operating condition representing the best trade off between detection efficiency (assumed to increase with the SPAD bias voltage) and dark pulse generation.

- PDP measurements at different excess voltages need to be performed to study the dependence of the avalanche triggering probability on the SPAD bias voltage and identify the PDP saturation level that may occur due to the maximization of the avalanche triggering probability.

- The crosstalk performance of the sensors integrated in the ASAP110LF chip need to be further investigated through crosstalk probability measurements. The latter, aiming at selectively studying the influence of a single pixel on the DCR of the neighbouring ones, can be managed by creating pairs of emitter-receiver pixels, in a subset of cells surrounding the SPAD under measurement.

- RTS measurements need to be repeated on a higher number of chip samples and for all the array structures. The RTS characterization should be performed, at different SPAD bias voltages, with higher time resolution and larger time span as compared to the ones used in this work. In addition, an AI-based algorithm, automatically detecting pixels affected by RTS in DCR, is currently under development.

- More samples, in different chips, of the dSiPM with parallel counter architecture need to be characterized, in order to quantitatively study the effectiveness of the presented architecture. Dedicated measurements should be performed on the structures of the last row in ASIPM, which implement memory elements within the sub-pixels and a reduced readout

circuit consisting of a parallel counter and a SOT logic network. By exploiting these measurements, a direct evaluation of the latency time provided by the structure can be performed.

- Further studies are necessary to investigate the feasibility of larger dSiPMs, consisting of more than 16 SPADs, read out by means of the parallel counter architecture. In addition, the presented structure should be provided with timestamp capabilities, as required by most of the applications exploiting Time Correlated Single Photon Counting.

- A characterization of the timing resolution featured by the SPADs integrated in the new chip needs to be performed at different SPAD bias voltages, temperatures and wavelengths. To achieve this purpose, pixels in the APXT array, minimally affecting the timing jitter of the output signal, or single sensors, located in the chip boundary, can be used. The SPAD timing resolution is a key parameter in biomedical applications (positron emission tomography, time-resolved spectroscopy, fluorescence decay analysis, Raman spectroscopy), thus representing an essential step for a complete sensor characterization.

- Measurements on dual layer samples need to be performed, so as to prove, also in a 110 nm CIS technology, the beneficial impact of a two-tier structure on the DCR of the detection system. The production of extremely low noise sensors based on CMOS SPADs operated at room temperature may have several implications in a number of fields like Nuclear Medicine and High Energy Physics experiments.

- In view of applications involving space-borne particle detection or charged particle tracking in High Energy Physics experiments, an irradiation campaign, aiming at investigating the radiation hardness of SPADs fabricated in 110 nm CIS technology, needs to be planned. The exposure of SPADs to ionizing and non-ionizing radiation may significantly affect the DCR of the target sensors, contributing with permanent or temporary damage to the lattice structure. During and after the irradiation procedure, the DCR of the SPADs under measurement can be monitored to investigate the dynamics of defect formation and short-term annealing.

Scientific contributions

In this section, a summary of the scientific contributions made by this PhD thesis is provided. Most of the work was carried out with the main purpose of

providing a test platform that can support the characterization of p+/nwell SPADs, fabricated in a 110 nm CIS technology. The properties of these sensors might be beneficially exploited in the field of charged particle tracking, representing the primary scientific field addressed by this work. In particular, CMOS SPADs can be considered as a valid alternative to hybrid pixels, that, with their relatively large amount of material, may not represent the most suitable choice in tracking applications at the next-generation linear colliders and B-factories. Moving into the field of calorimetry in High Energy Physics, especially as far as applications requiring high dynamic range are concerned, the research work was focused on the development of a digital SiPM featuring latency performance that can compete with the typical response of a fully analog structure. The final scope of this activity was to provide a SiPM architecture enabling real time photon counting in the digital domain.
The main scientific contributions of this thesis can be summarized as follows:

- Design and characterization of dual layer position-sensitive detection systems, based on CMOS SPADs, drastically reducing the impact of the sensor DCR through a coincidence readout logic, implemented at the pixel level.

- Design of a chip, consisting of several test structures, and an FPGA-based measurement setup supporting the characterization of CMOS SPADs, fabricated in a 110 nm CIS technology, under different test conditions and with various readout approaches.

- Collection of statistically meaningful sets of measurement data demonstrating the remarkable performance, in terms of photon detection and DCR, of CMOS SPADs fabricated in a 110 nm CIS technology.

- Demonstration of the effectiveness of a CIS technology in improving the noise performance of SPADs, through a direct comparison between the results from the sensors integrated in the ASAP110LF chip and SPADs fabricated in a 150 nm CMOS technology.

- Design and preliminary characterization of a novel digital SiPM architecture based on parallel counters, capable of returning the number of simultaneously firing SPADs with very low latency time (a few nanoseconds), thus potentially enabling real time photon counting with the use of a fully digital circuit.

Bibliography

[1] P. Seitz and A. J. P. Theuwisse, *Single-Photon Imaging*. Springer, 2011. ISBN: 978-3-642-18443-7, part of the book series: Springer Series in Optical Sciences (SSOS, volume 160).

[2] S. Cova, M. Ghioni, A. Lotito, I. Rech, and F. Zappa, "Evolution and prospects for single-photon avalanche diodes and quenching circuits," *Journal of Modern Optics*, vol. 51, no. 9-10, pp. 1267–1288, 2004.

[3] F. Acerbi and M. Perenzoni, "High sensitivity photodetector for photon-counting applications," in *Photon Counting - Fundamentals and Applications* (N. Britun, ed.), ch. 2, IntechOpen, 2018.

[4] G.-F. D. Betta, L. Pancheri, D. Stoppa, R. Henderson, and J. Richardson, "Avalanche photodiodes in submicron CMOS technologies for high-sensitivity imaging," in *Advances in Photodiodes* (G. F. D. Betta, ed.), ch. 11, Rijeka: IntechOpen, 2011.

[5] L. Ratti, P. Brogi, G. Collazuol, G.-F. Dalla Betta, P. S. Marrocchesi, L. Pancheri, A. Sulay, G. Torilla, and C. Vacchi, "Layered CMOS SPADs for low noise detection of charged particles," *Frontiers in Physics*, vol. 8, 2021.

[6] N. D'Ascenzo, P. S. Marrocchesi, C. S. Moon, F. Morsani, L. Ratti, V. Saveliev, A. S. Navarro, and Q. Xie, "Silicon avalanche pixel sensor for high precision tracking," *Journal of Instrumentation*, vol. 9, p. C03027, mar 2014.

[7] J. Minga, *Imaging Probe for Charged Particle Detection Based on SPAD Sensors*. PhD thesis, Università degli Studi di Pavia, 2022/2023.

[8] F. Villa, E. Conca, V. Sesta, N. Lusardi, F. Garzetti, A. Geraci, and F. Zappa, "SPADs and TDCs for photon-counting, timing and gated-imaging at 30 ps resolution and 60% efficiency," in *2018 IEEE Nuclear*

Science Symposium and Medical Imaging Conference (NSS/MIC), pp. 1–5, 2018.

[9] F. Gramuglia, M.-J. Lee, E. Venialgo, C. Bruschini, and E. Charbon, "Towards 10 ps SPTR and ultra-low DCR in SiPMs through the combination of microlenses and photonic crystals," in *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–3, 2017.

[10] I. Vornicu, J. M. Lopez-Martinez, F. N. Bandi, R. C. Galan, and A. Rodríguez-Vazquez, "Design of high-efficiency SPADs for LiDAR applications in 110 nm CIS technology," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 4776–4785, 2021.

[11] H.-H. Huang, C.-H. Liu, T.-Y. Huang, S.-D. Lin, and C.-Y. Lee, "Self-restoring and low-jitter circuits for high timing-resolution SPAD sensing applications," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2023.

[12] F. Gramuglia, M.-L. Wu, M.-J. Lee, C. Bruschini, and E. Charbon, "SPAD microcells with 12.1 ps SPTR for SiPMs in TOF-PET applications," in *2021 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–2, 2021.

[13] O. Kumagai, J. Ohmachi, M. Matsumura, S. Yagi, K. Tayu, K. Amagawa, T. Matsukawa, O. Ozawa, D. Hirono, Y. Shinozuka, R. Homma, K. Mahara, T. Ohyama, Y. Morita, S. Shimada, T. Ueno, A. Matsumoto, Y. Otake, T. Wakano, and T. Izawa, "A 189x600 back-illuminated stacked SPAD direct time-of-flight depth sensor for automotive LiDAR systems," in *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 64, pp. 110–112, 2021.

[14] D. P. Palubiak and M. J. Deen, "CMOS SPADs: Design issues and research challenges for detectors, circuits, and arrays," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 409–426, 2014.

[15] G. Lutz, *Semiconductor Radiation Detector*. Springer, 1999. ISBN: 978-3-540-71679-2.

[16] M. Hofbauer, K. Schneider-Hornstein, and H. Zimmermann, *Single-photon Detection for Data Communication and Quantum Systems*. 2053-2563, IOP Publishing, 2021.

[17] R. S. Muller, T. I. Kamins, and M. Chan, "pn junctions," in *Device Electronics for Integrated Circuits*, ch. 4, Wiley, 3 ed., 2002.

[18] A. Luque and S. Hegedus, *Handbook of Photovoltaic Science and Engineering*. Edited by A. Luque and S. Hegedus, 2003. ISBN: 978-0-471-49196-5.

[19] L. Rossi, P. Fischer, T. Rohe, and N. Wermes, *Pixel Detectors: From Fundamentals to Applications*. Springer, 2006. ISBN: 978-3-540-28332-4.

[20] J. Heinonen, A. Haarahiltunen, M. Serue, D. Kriukova, V. Vähänissi, T. P. Pasanen, H. Savin, and M. A. Juntunen, "Temperature dependency of responsivity and dark current of nearly ideal black silicon photodiodes," in *Optical Components and Materials XVIII* (S. Jiang and M. J. F. Digonnet, eds.), vol. 11682, p. 1168207, International Society for Optics and Photonics, SPIE, 2021.

[21] C. Wong, W. H. W. Hasan, and S. Isaak, "The design and characterization of breakdown mechanism on p+/n- well single photon avalanche diode (SPAD)," in *Journal of Advanced Research in Applied Mechanics*, vol. 13, pp. 12–23, 2015.

[22] B. F. Aull, A. H. Loomis, D. J. Young, R. M. Heinrichs, B. J. Felton, P. J. Daniels, and D. J. Landers, "Geiger-mode avalanche photodiodes for three-dimensional imaging," *Lincoln laboratory journal*, vol. 13, no. 2, pp. 335–349, 2002.

[23] S. Gundacker and A. Heering, "The silicon photomultiplier: fundamentals and applications of a modern solid-state photon detector," *Physics in Medicine and Biology*, vol. 65, p. 17TR01, aug 2020.

[24] W. Oldham, R. Samuelson, and P. Antognetti, "Triggering phenomena in avalanche diodes," *IEEE Transactions on Electron Devices*, vol. 19, no. 9, pp. 1056–1060, 1972.

[25] G. Chesi, *Advances in Quantum Nonlinear Optics A nonclassical journey from the optimization of Silicon Photomultipliers for Quantum Optics to quantum Second-Harmonic Generation*. PhD thesis, Università dell'Insubria Dipartimento di Scienza e Alta Tecnologia, 03 2022.

[26] A. Stewart, V. Saveliev, S. Bellis, D. Herbert, P. Hughes, and C. Jackson, "Performance of 1-mm^2 silicon photomultiplier," *Quantum Electronics, IEEE Journal of*, vol. 44, pp. 157 – 164, 03 2008.

[27] M. Ghioni, A. Gulinatti, I. Rech, F. Zappa, and S. Cova, "Progress in silicon single-photon avalanche diodes," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 13, pp. 852 – 862, 08 2007.

[28] G. Bonanno, M. Belluso, S. Billotta, P. Finocchiaro, and A. Pappalardo, "Geiger avalanche photodiodes (G-APDs) and their characterization," in *Photodiodes* (J.-W. Park, ed.), ch. 11, Rijeka: IntechOpen, 2011.

[29] E. Webster, J. Richardson, L. Grant, D. Renshaw, and R. Henderson, "A single-photon avalanche diode in 90 nm CMOS imaging technology with 44photon detection efficiency at 690 nm," *Electron Device Letters, IEEE*, vol. 33, pp. 694–696, 05 2012.

[30] A. Vilà, A. Arbat, E. Vilella, and A. Dieguez, "Geiger-mode avalanche photodiodes in standard CMOS technologies," in *Photodetectors* (S. Gateva, ed.), ch. 9, Rijeka: IntechOpen, 2012.

[31] H. Xu, L. Pancheri, G.-F. D. Betta, and D. Stoppa, "Design and characterization of a p+/n-well SPAD array in 150nm CMOS process," *Opt. Express*, vol. 25, pp. 12765–12778, May 2017.

[32] M. Zarghami, L. Gasparini, L. Parmesan, M. Moreno-García, A. Stefanov, B. Bessire, M. Unternahrer, and M. Perenzoni, "A 32 × 32-pixel CMOS imager for quantum optics with per-SPAD TDC, 19.48% fill-factor in a 44.64 μm pitch reaching 1-MHz observation rate," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 10, pp. 2819–2830, 2020.

[33] R. J. McIntyre, "Multiplication noise in uniform avalanche diodes," *IEEE Transactions on Electron Devices*, vol. ED-13, pp. 164–168, Jan 1966.

[34] M. Musacci, *Design and characterization of CMOS SPADs for charged particle detectors*. PhD thesis, Università degli Studi di Pavia, 2017/2018.

[35] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Appl. Opt.*, vol. 35, pp. 1956–1976, Apr 1996.

[36] L. Ratti, P. Brogi, G. Collazuol, G. . Dalla Betta, A. Ficorella, L. Lodola, P. S. Marrocchesi, S. Mattiazzo, F. Morsani, M. Musacci, L. Pancheri, and C. Vacchi, "Dark count rate degradation in CMOS SPADs exposed to X-rays and neutrons," *IEEE Transactions on Nuclear Science*, vol. 66, pp. 567–574, Feb 2019.

[37] E. Kamrani, F. Lesage, and M. Sawan, "Premature edge breakdown prevention techniques in CMOS APD fabrication," in *10th IEEE International NEWCAS Conference*, pp. 345–348, 06 2012.

[38] G. Bonanno, D. Marano, M. Belluso, S. Billotta, A. Grillo, S. Garozzo, G. Romeo, and M. C. Timpanaro, "Characterization measurements methodology and instrumental set-up optimization for new SiPM detectors - part I: electrical tests," *IEEE Sensors Journal*, vol. 14, pp. 3557–3566, Oct 2014.

[39] K. T. Lim, H. Kim, J. Kim, and G. Cho, "Effect of electric field on primary dark pulses in SPADs for advanced radiation detection applications," *Nuclear Engineering and Technology*, vol. 53, no. 2, pp. 618–625, 2021.

[40] Y. Xu, P. Xiang, and X. Xie, "Comprehensive understanding of dark count mechanisms of single-photon avalanche diodes fabricated in deep sub-micron CMOS technologies," *Solid-State Electronics*, vol. 129, pp. 168–174, 2017.

[41] L. Pancheri, D. Stoppa, and G.-F. Dalla Betta, "Characterization and modeling of breakdown probability in sub-micrometer CMOS SPADs," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 328–335, 2014.

[42] A. Ficorella, L. Pancheri, P. Brogi, G. Collazuol, G. D. Betta, P. S. Marrocchesi, F. Morsani, L. Ratti, and A. Savoy-Navarro, "Crosstalk characterization of a two-tier pixelated avalanche sensor for charged particle detection," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, pp. 1–8, March 2018.

[43] G. Torilla, J. Minga, P. Brogi, G. Collazuol, G.-F. Dalla Betta, P. Marrocchesi, L. Pancheri, L. Ratti, and C. Vacchi, "DCR and crosstalk characterization of a bi-layered 24 × 72 CMOS SPAD array for charged particle detection," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1046, p. 167693, 2023.

[44] A. Ficorella, L. Pancheri, G.-F. D. Betta, P. Brogi, G. Collazuol, P. S. Marrocchesi, F. Morsani, L. Ratti, and A. Savoy-Navarro, "Crosstalk mapping in CMOS SPAD arrays," in *2016 46th European Solid-State Device Research Conference (ESSDERC)*, pp. 101–104, 2016.

[45] M. Musacci, P. Brogi, G. Collazuol, G. D. Betta, A. Ficorella, L. Lodola, P. Marrocchesi, F. Morsani, L. Pancheri, L. Ratti, A. Savoy-Navarro, and C. Vacchi, "Geiger-mode avalanche pixels in 180 nm HV CMOS process for a dual-layer particle detector," in *2016 IEEE Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop (NSS/MIC/RTSD)*, pp. 1–5, 2016.

[46] F. Gramuglia, P. Keshavarzian, E. Kizilkan, C. Bruschini, S. S. Tan, M. Tng, E. Quek, M.-J. Lee, and E. Charbon, "Engineering breakdown probability profile for PDP and DCR optimization in a SPAD fabricated in a standard 55 nm BCD process," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 28, no. 2: Optical Detectors, pp. 1–10, 2022.

[47] M. Moreno-García, H. Xu, L. Gasparini, and M. Perenzoni, "Low-noise single photon avalanche diodes in a 110nm CIS technology," in *2018 48th European Solid-State Device Research Conference (ESSDERC)*, pp. 94–97, 2018.

[48] E. Charbon, "Single-photon imaging in complementary metal oxide semiconductor processes," *Philos Trans A Math Phys Eng Sci. 2014*, vol. 372, 02 2014.

[49] F. Gramuglia, M.-L. Wu, C. Bruschini, M.-J. Lee, and E. Charbon, "A low-noise CMOS SPAD pixel with 12.1 ps SPTR and 3 ns dead time," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 28, no. 2: Optical Detectors, pp. 1–9, 2022.

[50] F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, and I. Rech, "Fully integrated active quenching circuit driving custom-technology SPADs with 6.2 ns dead time," *IEEE Photonics Technology Letters*, vol. 31, no. 1, pp. 102–105, 2019.

[51] G. Hurkx, D. Klaassen, and M. Knuvers, "A new recombination model for device simulation including tunneling," *IEEE Transactions on Electron Devices*, vol. 39, no. 2, pp. 331–338, 1992.

[52] C. Wang, J. Wang, Z. Xu, J. Li, J. Zhao, C. Wu, Y. Wei, and Y. Zhu, "Optimization of dead time in SPAD-based photon-counting communication system with afterpulsing," in *2019 18th International Conference on Optical Communications and Networks (ICOCN)*, pp. 1–3, 2019.

[53] P. Xiang and Y. Xu, "Improved photon detection efficiency of single photon avalanche diodes with buried layer structure," in *2016 13th IEEE*

International Conference on Solid-State and Integrated Circuit Technology (ICSICT), pp. 691–693, 2016.

[54] M. Gersbach, C. Niclass, E. Charbon, J. Richardson, R. Henderson, and L. Grant, "A single photon detector implemented in a 130nm CMOS imaging process," in *ESSDERC 2008 - 38th European Solid-State Device Research Conference*, pp. 270–273, 2008.

[55] M. H. U. Habib, F. Quaiyum, S. K. Islam, and N. McFarlane, "Optimization of perimeter gated SPADs in a standard CMOS process," in *SENSORS, 2014 IEEE*, pp. 1668–1671, 2014.

[56] M.-J. Lee, A. R. Ximenes, P. Padmanabhan, T.-J. Wang, K.-C. Huang, Y. Yamashita, D.-N. Yaung, and E. Charbon, "High-performance back-illuminated three-dimensional stacked single-photon avalanche diode implemented in 45-nm CMOS technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, pp. 1–9, 2018.

[57] B. Nouri, M. Dandin, and P. Abshire, "Large-area low-noise single-photon avalanche diodes in standard CMOS," in *SENSORS, 2012 IEEE*, pp. 1–5, 2012.

[58] F. Acerbi, G. Paternoster, A. Gola, N. Zorzi, and C. Piemonte, "Silicon photomultipliers and single-photon avalanche diodes with enhanced NIR detection efficiency at FBK," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 912, pp. 309–314, 2018. New Developments In Photodetection 2017.

[59] M. Ghioni, S. Cova, F. Zappa, and C. Samori, "Compact active quenching circuit for fast photon counting with avalanche photodiodes," *Review of Scientific Instruments*, vol. 67, no. 10, pp. 3440–3448, 1996.

[60] C. Niclass, A. Rochas, P.-A. Besse, R. Popovic, and E. Charbon, "A 4μs integration time imager based on CMOS single photon avalanche diode technology," *Sensors and Actuators A: Physical*, vol. 130-131, pp. 273 – 281, 2006. Selected Papers from TRANSDUCERS '05.

[61] L. Pancheri and D. Stoppa, "Low-noise CMOS single-photon avalanche diodes with 32 ns dead time," in *ESSDERC 2007 - 37th European Solid State Device Research Conference*, pp. 362–365, Sep. 2007.

[62] C. Niclass, *Single-photon image sensors in CMOS: picosecond resolution for three-dimensional imaging*. PhD thesis, EPFL, no.4161 (2008), DOI:10.5075/epfl-thesis-4161, 2008.

[63] J. Hu, X. Xin, X. Li, J. H. Zhao, B. VanMil, K.-K. Lew, R. Myers-Ward, E. Jr, and Gaskill, "4H-SiC visible-blind single-photon avalanche diode for ultraviolet detection at 280 and 350 nm," *Electron Devices, IEEE Transactions on*, vol. 55, pp. 1977 – 1983, 09 2008.

[64] A. Gallivanoni, I. Rech, and M. Ghioni, "Progress in quenching circuits for single photon avalanche diodes," *IEEE Transactions on nuclear science*, vol. 57, no. 6, pp. 3815–3826, 2010.

[65] C. Niclass, A. Rochas, P. Besse, and E. Charbon, "Toward a 3-D camera based on single photon avalanche diodes," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 10, pp. 796–802, July 2004.

[66] E. Charbon, "Single-photon imaging in CMOS," in *Detectors and Imaging Devices: Infrared, Focal Plane, Single Photon* (E. L. Dereniak, J. P. Hartke, P. D. LeVan, R. E. Longshore, A. K. Sood, and M. Razeghi, eds.), vol. 7780, pp. 319 – 333, International Society for Optics and Photonics, SPIE, 2010.

[67] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, "A 128 × 128 single-photon image sensor with column-level 10-bit time-to-digital converter array," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 2977–2989, Dec 2008.

[68] R. K. Henderson, N. Johnston, F. Mattioli Della Rocca, H. Chen, D. Day-Uei Li, G. Hungerford, R. Hirsch, D. Mcloskey, P. Yip, and D. J. S. Birch, "A 192 × 128 time correlated SPAD image sensor in 40-nm CMOS technology," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 7, pp. 1907–1916, 2019.

[69] R. K. Henderson, N. Johnston, S. W. Hutchings, I. Gyongy, T. A. Abbas, N. Dutton, M. Tyler, S. Chan, and J. Leach, "A 256 × 256 40nm/90nm CMOS 3D-stacked 120db dynamic-range reconfigurable time-resolved SPAD imager," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 106–108, 2019.

[70] L. Parmesan, N. Dutton, N. Calder, N. Krstaji, A. Holmes, L. Grant, and R. Henderson, "A 256 × 256 SPAD array with in-pixel time to amplitude

conversion for fluorescence lifetime amplitude conversion for fluorescence lifetime imaging microscopy," in *2015 International Image Sensor Workshop*, 06 2015.

[71] A. C. Ulku, C. Bruschini, I. M. Antolovic, Y. Kuo, R. Ankri, S. Weiss, X. Michalet, and E. Charbon, "A 512 × 512 SPAD image sensor with integrated gating for widefield FLIM," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1–12, 2019.

[72] A. R. Ximenes, P. Padmanabhan, M.-J. Lee, Y. Yamashita, D. N. Yaung, and E. Charbon, "A 256 × 256 45/65nm 3D-stacked SPAD-based direct TOF image sensor for LiDAR applications with optical polar modulation for up to 18.6db interference suppression," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, pp. 96–98, 2018.

[73] D. Renker and E. Lorenz, "Advances in solid state photon detectors," *Journal of Instrumentation*, vol. 4, p. P04004, apr 2009.

[74] F. Corsi, C. Marzocca, A. Perrotta, A. Dragone, M. Foresta, A. Del Guerra, S. Marcatili, G. Llosa, G. Collazzuol, G. F. D. Betta, N. Dinu, C. Piemonte, G. U. Pignatel, and G. Levi, "Electrical characterization of silicon photo-multiplier detectors for optimal front-end design," in *2006 IEEE Nuclear Science Symposium Conference Record*, vol. 2, pp. 1276–1280, 2006.

[75] J.-F. Pratte, F. Nolet, S. Parent, F. Vachon, N. Roy, T. Rossignol, K. Deslandes, H. Dautet, R. Fontaine, and S. A. Charlebois, "3D photon-to-digital converter for radiation instrumentation: Motivation and future works," *Sensors*, vol. 21, no. 2, 2021.

[76] D. R. Schaart, E. Charbon, T. Frach, and V. Schulz, "Advances in digital SiPMs and their application in biomedical imaging," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 809, pp. 31–52, 2016. Advances in detectors and applications for medicine.

[77] Y. Haemisch, T. Frach, C. Degenhardt, and A. Thon, "Fully digital arrays of silicon photomultipliers (dSiPM). A scalable alternative to vacuum photomultiplier tubes (PMT)," *Physics Procedia*, vol. 37, pp. 1546–1560, 2012. Proceedings of the 2nd International Conference on Technology and Instrumentation in Particle Physics (TIPP 2011).

[78] R. Radojcic, *More-than-Moore 2.5D and 3D SiP integration.* Springer Cham, 02 2017.

[79] B.-L. Berube, V. Rheaume, A. Corbeil Therrien, s. Parent, L. Maurais, A. Boisvert, G. Carini, S. Charlebois, R. Fontaine, and J.-F. Pratte, "Development of a single photon avalanche diode (SPAD) array in high voltage CMOS 0.8 μm dedicated to a 3D integrated circuit (3DIC)," in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1835–1839, 10 2012.

[80] T. Frach, G. Prescher, C. Degenhardt, and B. Zwaans, "The digital silicon photomultiplier. System architecture and performance evaluation," in *2010 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, pp. 1722–1727, 2010.

[81] T. Frach, G. Prescher, C. Degenhardt, R. de Gruyter, A. Schmitz, and R. Ballizany, "The digital silicon photomultiplier. Principle of operation and intrinsic detector performance," in *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, pp. 1959–1965, 2009.

[82] E. Charbon, C. E. Bruschini, C. Veerappan, L. H. C. Braga, N. Massari, M. Perenzoni, L. Gasparini, D. Stoppa, R. J. Walker, A. T. Erdogan, R. K. Henderson, S. East, L. Grant, B. Játékos, F. Ujhelyi, G. Erdei, E. Lorincz, L. André, L. Maingault, V. Reboud, L. Verger, E. G. d'aillon, P. Major, Z. Papp, and G. Nerneth, "SPADnet: A fully digital, networked approach to MRI compatible PET systems based on deep-submicron CMOS technology," *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC)*, pp. 1–5, 2013.

[83] L. H. C. Braga, L. Gasparini, L. Grant, R. K. Henderson, N. Massari, M. Perenzoni, D. Stoppa, and R. Walker, "A fully digital 8×16 SiPM array for PET applications with per-pixel TDCs and real-time energy output," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 301–314, 2014.

[84] E. Garutti, "EndoTOFPET-US a novel multimodal tool for endoscopy and positron emission tomography," in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, pp. 2096–2101, 2012.

[85] N. Aubry, E. Auffray, F. B. Mimoun, N. Brillouet, R. Bugalho, E. Charbon, O. Charles, D. Cortinovis, P. Courday, A. Cserkaszky, C. Damon,

K. Doroud, J. M. Fischer, G. Fornaro, J. M. Fourmigue, B. Frisch, B. Fürst, J. Gardiazabal, K. Gadow, E. Garutti, C. Gaston, A. Gil-Ortiz, E. Guedj, T. Harion, P. Jarron, J. Kabadanian, T. Lasser, R. Laugier, P. Lecoq, D. Lombardo, S. Mandai, E. Mas, T. Meyer, O. Mundler, N. Navab, C. Ortigao, M. Paganoni, D. Perrodin, M. Pizzichemi, J. O. Prior, T. Reichl, M. Reinecke, M. Rolo, H. C. Schultz-Coulon, M. Schwaiger, W. Shen, A. Silenzi, J. C. Silva, R. Silva, I. S. Schweiger, R. Stamen, J. Traub, J. Varela, V. Veckalns, V. Vidal, J. Vishwas, T. Wendler, C. Xu, S. Ziegler, and M. Zvolsky, "EndoTOFPET-US: a novel multimodal tool for endoscopy and positron emission tomography," *Journal of Instrumentation*, vol. 8, p. C04002, apr 2013.

[86] S. Mandai and E. Charbon, "A 4x4x416 digital SiPM array with 192 TDCs for multiple high-resolution timestamp acquisition," *Journal of Instrumentation*, vol. 8, p. P05024, may 2013.

[87] P. Fischer, T. Armbruster, R. Blanco, M. Ritzert, I. Sacco, and S. Weyers, "A dense SPAD array with full frame readout and fast cluster position reconstruction," in *Proceedings of the IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Seattle, WA, USA*, pp. 8–15, 2014.

[88] P. Fischer, R. Blanco, I. Sacco, M. Ritzert, and S. Weyers, "SPAD array chips with full frame readout for crystal characterization," in *EJNMMI Phys 2 (Suppl 1), A3 (2015).*, 2015.

[89] F. Nolet, W. Lemaire, F. Dubois, N. Roy, S. Carrier, A. Samson, S. A. Charlebois, R. Fontaine, and J.-F. Pratte, "A 256 pixelated SPAD readout ASIC with in-pixel TDC and embedded digital signal processing for uniformity and skew correction," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 949, p. 162891, 2020.

[90] F. Nolet, F. Dubois, N. Roy, S. Parent, W. Lemaire, A. Massie-Godon, S. A. Charlebois, R. Fontaine, and J.-F. Pratte, "Digital SiPM channel integrated in CMOS 65 nm with 17.5 ps FWHM single photon timing resolution," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 912, pp. 29–32, 2018. New Developments In Photodetection 2017.

[91] N. Roy, F. Nolet, F. Dubois, M.-O. Mercier, R. Fontaine, and J.-F. Pratte, "Low power and small area, 6.9 ps RMS time-to-digital converter for 3D digital SiPM," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 1, no. 6, pp. 486–494, 2017.

[92] L. Braga, L. Gasparini, L. Grant, R. Henderson, N. Massari, M. Perenzoni, D. Stoppa, and R. Walker, "A fully digital 8 × 16 SiPM array for PET applications with per-pixel TDCs and real-time energy output," *Solid-State Circuits, IEEE Journal of*, vol. 49, pp. 301–314, 01 2014.

[93] M. N. WERNICK and J. N. AARSVOLD, "Chapter 2 - introduction to emission tomography," in *Emission Tomography* (Miles N. Wernick and John N. Aarsvold, ed.), pp. 11–23, San Diego: Academic Press, 2004.

[94] A. Del Guerra, N. Belcari, M. G. Bisogni, G. Llosá, S. Marcatili, and S. Moehrs, "Advances in position-sensitive photodetectors for PET applications," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 604, no. 1, pp. 319–322, 2009. PSD8.

[95] F. Gramuglia, A. Muntean, E. Venialgo, M.-J. Lee, S. Lindner, M. Motoyoshi, A. Ardelean, C. Bruschini, and E. Charbon, "CMOS 3D-stacked FSI multi-channel digital SiPM for time-of-flight PET applications," in *2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–3, 2020.

[96] P. Lecoq, C. Morel, J. O. Prior, D. Visvikis, S. Gundacker, E. Auffray, P. Krizan, R. M. Turtos, D. Thers, E. Charbon, J. Varela, C. de La Taille, A. Rivetti, D. Breton, J.-F. Pratte, J. Nuyts, S. Surti, S. Vandenberghe, P. Marsden, K. Parodi, J. M. Benlloch, and M. Benoit, "Roadmap toward the 10 ps time-of-flight PET challenge," *Physics in Medicine and Biology*, vol. 65, p. 21RM01, oct 2020.

[97] B. Bai, "Celesteion. time-of-flight technology," tech. rep., Canon Medical Systems USA, Inc., 2018.

[98] M. Perenzoni, D. Perenzoni, and D. Stoppa, "A 64 × 64-pixels digital silicon photomultiplier direct TOF sensor with 100-Mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6 km for spacecraft navigation and landing," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 151–160, 2017.

[99] D. Li, J. Hu, R. Ma, X. Wang, Y. Liu, and Z. Zhu, "SPAD-based LiDAR with real-time accuracy calibration and laser power regulation," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 70, no. 2, pp. 431–435, 2023.

[100] K. Morimoto, A. Ardelean, M.-L. Wu, A. C. Ulku, I. M. Antolovic, C. Bruschini, and E. Charbon, "Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications," *Optica*, vol. 7, pp. 346–354, Apr 2020.

[101] A. Gupta, A. Ingle, and M. Gupta, "Asynchronous single-photon 3D imaging," in *Proc. ICCV*, October 2019.

[102] E. Charbon, C. Bruschini, and M.-J. Lee, "3D-stacked CMOS SPAD image sensors: Technology and applications," in *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 1–4, 2018.

[103] C. Chang, D. Sud, and M. Mycek, "Fluorescence lifetime imaging microscopy," in *Digital Microscopy, 3rd Edition*, vol. 81 of *Methods in Cell Biology*, pp. 495–524, Academic Press, 2007.

[104] D. Li, Z. Wu, Y. Xu, and T. Zhao, "A novel three observation-windows measurement scheme for SPAD fluorescence lifetime imaging detector," in *2018 China Semiconductor Technology International Conference (CSTIC)*, pp. 1–3, 2018.

[105] D. E. Schwartz, E. Charbon, and K. L. Shepard, "A single-photon avalanche diode array for fluorescence lifetime imaging microscopy," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 11, pp. 2546–2557, 2008.

[106] G.-J. Bakker, V. Andresen, R. M. Hoffman, and P. Friedl, "Chapter five - fluorescence lifetime microscopy of tumor cell invasion, drug delivery, and cytotoxicity," in *Imaging and Spectroscopic Analysis of Living Cells* (P. M. conn, ed.), vol. 504 of *Methods in Enzymology*, pp. 109–125, Academic Press, 2012.

[107] M. Benetti, M. Popleteeva, G.-F. Dalla Betta, L. Pancheri, and D. Stoppa, "Characterization of a CMOS SPAD sensor designed for fluorescence lifetime spectroscopy," in *2011 7th Conference on Ph.D. Research in Microelectronics and Electronics*, pp. 185–188, 2011.

[108] H. C. Gerritsen, M. A. H. Asselbergs, A. V. Agronskaia, and W. G. J. H. M. Van Sark, "Fluorescence lifetime imaging in scanning microscopes: acquisition speed, photon economy and lifetime resolution," *Journal of Microscopy*, vol. 206, no. 3, pp. 218–224, 2002.

[109] S. Moreno, V. Moro, and A. Dieguez, "A 72-bin in-pixel mixed-signal TDC for SPAD-based fluorescence lifetime measurements," in *2022 37th Conference on Design of Circuits and Integrated Circuits (DCIS)*, pp. 01–05, 2022.

[110] Y. Maruyama and E. Charbon, "An all-digital, time-gated 128x128 SPAD array for on-chip, filter-less fluorescence detection," in *2011 16th International Solid-State Sensors, Actuators and Microsystems Conference*, pp. 1180–1183, 2011.

[111] N. Massari, A. Tontini, L. Parmesan, M. Perenzoni, M. Gruijc, I. Verbauwhede, T. Strohm, D. Oshinubi, I. Herrmann, and A. Brenneis, "A monolithic SPAD-based random number generator for cryptographic application," in *ESSCIRC 2022- IEEE 48th European Solid State Circuits Conference (ESSCIRC)*, pp. 73–76, 2022.

[112] N. Massari, L. Gasparini, A. Tomasi, A. Meneghetti, H. Xu, D. Perenzoni, G. Morgari, and D. Stoppa, "A 16x16 pixels SPAD-based 128mb/s quantum random number generator with -74db light rejection ratio and -6.7ppm/degree Celcius bias sensitivity on temperature," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 292–293, 2016.

[113] F.-X. Wang, C. Wang, W. Chen, S. Wang, F.-S. Lv, D.-Y. He, Z.-Q. Yin, H.-W. Li, G.-C. Guo, and Z.-F. Han, "Robust quantum random number generator based on avalanche photodiodes," *Journal of Lightwave Technology*, vol. 33, pp. 3319–3326, aug 2015.

[114] F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, I. Rech, and R. Osellame, "Recent advances and future perspectives of single-photon avalanche diodes for quantum photonics applications," *Advanced Quantum Technologies*, vol. 4, no. 2, p. 2000102, 2021.

[115] G. Massaro, P. Mos, S. Vasiukov, F. D. Lena, F. Scattarella, F. V. Pepe, A. Ulku, D. Giannella, E. Charbon, C. Bruschini, and M. D'Angelo, "Quantum imaging at 10 volumetric images per second," 2022.

[116] X. Jiang, M. Itzler, K. O'Donnell, M. Entwistle, M. Owens, K. Slomkowski, and S. Rangwala, "InP-based single-photon detectors and geiger-mode APD arrays for quantum communications applications," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 21, no. 3, pp. 5–16, 2015.

[117] J. Minga, P. Brogi, G. Collazuol, G.-F. Dalla Betta, P. S. Marrocchesi, F. Morsani, L. Pancheri, L. Ratti, G. Torilla, and C. Vacchi, "A wireless, battery-powered probe based on a dual-tier CMOS SPAD array for charged particle sensing," *Electronics*, vol. 12, no. 11, 2023.

[118] A. F. Bulling and I. Underwood, "Accelerated electron detection using single photon avalanche diodes," in *2018 IEEE SENSORS*, pp. 1–4, 2018.

[119] E. Manuzzato, L. Gasparini, M. Perenzoni, Y. Zou, L. Parmesan, G. Battistoni, M. D. Simoni, Y. Dong, M. Fischetti, E. Gioscio, I. Mattei, R. Mirabelli, V. Patera, A. Sarti, A. Schiavi, A. Sciubba, S. M. Valle, G. Traini, and M. Marafini, "A 16x8 digital-SiPM array with distributed trigger generator for low SNR particle tracking," in *ESSCIRC 2019 - IEEE 45th European Solid State Circuits Conference (ESSCIRC)*, pp. 75–78, 2019.

[120] E. Vilella, O. Alonso, and A. Dieguez, "3D integration of geiger-mode avalanche photodiodes aimed to very high fill-factor pixels for future linear colliders," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 731, pp. 103–108, 2013. PIXEL 2012.

[121] L. Pancheri, A. Ficorella, P. Brogi, G. Collazuol, G.-F. Dalla Betta, P. Marrocchesi, F. Morsani, L. Ratti, A. Savoy-Navarro, and A. Sulaj, "First demonstration of a two-tier pixelated avalanche sensor for charged particle detection," *IEEE Journal of the Electron Devices Society*, vol. PP, pp. 1–1, 08 2017.

[122] L. Pancheri, P. Brogi, G. Collazuol, G.-F. D. Betta, A. Ficorella, P. Marrocchesi, F. Morsani, L. Ratti, and A. Savoy-Navarro, "First prototypes of two-tier avalanche pixel sensors for particle detection," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 845, pp. 143 – 146, 2017. Proceedings of the Vienna Conference on Instrumentation 2016.

[123] W. Riegler and P. Windischhofer, "Time resolution and efficiency of SPADs and SiPMs for photons and charged particles," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1003, p. 165265, 2021.

[124] G. Torilla, S. Giroletti, P. Brogi, G. Collazuol, G.-F. D. Betta, P. Marrocchesi, J. Minga, F. Morsani, L. Pancheri, L. Ratti, and C. Vacchi, "Digital SiPMs in a 110 nm CMOS technology with fast parallel counter architecture," in *2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–5, 2022.

[125] L. Ratti, P. Brogi, G. Collazuol, G.-F. D. Betta, A. Ficorella, P. S. Marrocchesi, F. Morsani, L. Pancheri, G. Torilla, and C. Vacchi, "DCR performance in neutron-irradiated CMOS SPADs from 150- to 180-nm technologies," *IEEE Transactions on Nuclear Science*, vol. 67, no. 7, pp. 1293–1301, 2020.

[126] L. Ratti, P. Brogi, G. Collazuol, G.-F. Dalla Betta, A. Ficorella, P. S. Marrocchesi, L. Pancheri, and C. Vacchi, "Dark count rate distribution in neutron-irradiated CMOS SPADs," *IEEE Transactions on Electron Devices*, vol. 66, no. 12, pp. 5230–5237, 2019.

[127] W.-Y. Ha, E. Park, B. Park, Y. Chae, W.-Y. Choi, and M.-J. Lee, "Noise optimization of single-photon avalanche diodes fabricated in 110 nm CMOS image sensor technology," *Opt. Express*, vol. 30, pp. 14958–14965, Apr 2022.

[128] B. Steindl, W. Gaberl, R. Enne, S. Schidl, K. Schneider-Hornstein, and H. Zimmermann, "Linear mode avalanche photodiode with 1-GHz bandwidth fabricated in 0.35- μ m CMOS," *IEEE Photonics Technology Letters*, vol. 26, no. 15, pp. 1511–1514, 2014.

[129] L. Dadda, "On parallel digital multipliers," *Alta Frequenza*, vol. 45, pp. 574–580, 1976.

[130] E. Swartzlander, "Parallel counters," *IEEE Transactions on Computers*, vol. C-22, no. 11, pp. 1021–1024, 1973.

[131] E. Jones R. F., Swartzlander, "Parallel counter implementation," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 7, pp. 223–232, 1994.

[132] S. Giroletti, L. Ratti, and C. Vacchi, "Design algorithm for N-bit input parallel counters in application to dSiPM readout," in *2022 17th Conference on Ph.D Research in Microelectronics and Electronics (PRIME)*, pp. 37–40, 2022.

[133] F. Gramuglia, A. Muntean, C. A. Fenoglio, E. Venialgo, M.-J. Lee, S. Lindner, M. Motoyoshi, A. Ardelean, C. Bruschini, and E. Charbon, "Architecture and characterization of a cmos 3d-stacked fsi multi-channel digital sipm for time-of-flight pet applications," in *2021 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–2, 2021.

[134] C. Foster and F. Stockton, "Counting responders in an associative memory," *IEEE Transactions on Computers*, vol. C-20, no. 12, pp. 1580–1583, 1971.

[135] L. Pancheri, G.-F. Dalla Betta, L. H. Campos Braga, H. Xu, and D. Stoppa, "A single-photon avalanche diode test chip in 150nm CMOS technology," in *2014 International Conference on Microelectronic Test Structures (ICMTS)*, pp. 161–164, 2014.

[136] D. Bronzi, F. Villa, S. Tisa, A. Tosi, and F. Zappa, "SPAD figures of merit for photon-counting, photon-timing, and imaging applications: A review," *IEEE Sensors Journal*, vol. 16, no. 1, pp. 3–12, 2016.

[137] C. Veerappan and E. Charbon, "CMOS SPAD based on photo-carrier diffusion achieving PDP >40% from 440 to 580 nm at 4 V excess bias," *IEEE Photonics Technology Letters*, vol. 27, no. 23, pp. 2445–2448, 2015.

[138] M.-J. Lee, U. Karaca, E. Kizilkan, C. Bruschini, and E. Charbon, "A 73% peak PDP single-photon avalanche diode implemented in 110 nm CIS technology with doping compensation," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. PP, pp. 1–10, 01 2023.

[139] X. Qian, W. Jiang, and M. J. Deen, "Enhanced photon detection probability model for single-photon avalanche diodes in TCAD with machine learning," in *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–6, 2022.

[140] I. Vornicu, J. M. López-Martínez, F. N. Bandi, R. C. Galán, and A. Rodríguez-Vázquez, "Design of high-efficiency SPADs for LiDAR applications in 110nm CIS technology," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 4776–4785, 2021.

[141] C. Chang, S. Chiu, and L. Hsu, "Temperature dependence of breakdown voltage in silicon abrupt p-n junctions," *IEEE Transactions on Electron Devices*, vol. 18, no. 6, pp. 391–393, 1971.

[142] M. Moreno-García, H. Xu, L. Gasparini, and M. Perenzoni, "Low-noise single photon avalanche diodes in a 110nm CIS technology," in *2018 48th European Solid-State Device Research Conference (ESSDERC)*, pp. 94–97, 2018.

[143] E. Monteil, *Front-end electronics in 65 nm CMOS technology for the HL-LHC upgrades*. PhD thesis, Università degli Studi di Torino, 2016.

[144] M. A. Karami, L. Carrara, C. Niclass, M. Fishburn, and E. Charbon, "RTS noise characterization in single-photon avalanche diodes," *IEEE Electron Device Letters*, vol. 31, no. 7, pp. 692–694, 2010.

[145] V. Goiffon, G. R. Hopkinson, P. Magnan, F. Bernard, G. Rolland, and O. Saint-Pe, "Multilevel RTS in proton irradiated CMOS image sensors manufactured in a deep submicron technology," *IEEE Transactions on Nuclear Science*, vol. 56, no. 4, pp. 2132–2141, 2009.

[146] L. Ratti, P. Brogi, G. Collazuol, G.-F. Dalla Betta, J. Delgado, P. Marrocchesi, J. Minga, F. Morsani, L. Pancheri, F. Pino, A. Selva, F. Stolzi, G. Torilla, and C. Vacchi, "Online dark count rate measurements in 150 nm CMOS SPADs exposed to low neutron fluxes," *IEEE Transactions on Nuclear Science*, pp. 1–1, 2024.

[147] F. Di Capua, M. Campajola, L. Campajola, C. Nappi, E. Sarnelli, L. Gasparini, and H. Xu, "Random telegraph signal in proton irradiated single-photon avalanche diodes," *IEEE Transactions on Nuclear Science*, vol. 65, no. 8, pp. 1654–1660, 2018.

[148] M. A. Karami, C. Niclass, and E. Charbon, "Random telegraph signal in single-photon avalanche diodes," 01 2009.

[149] C. Durnez, V. Goiffon, C. Virmontois, J.-M. Belloir, P. Magnan, and L. Rubaldo, "In-depth analysis on radiation induced multi-level dark current random telegraph signal in silicon solid state image sensors," *IEEE Transactions on Nuclear Science*, vol. 64, no. 1, pp. 19–26, 2017.

[150] I. Hopkins and G. Hopkinson, "Further measurements of random telegraph signals in proton irradiated CCDs," *IEEE Transactions on Nuclear Science*, vol. 42, no. 6, pp. 2074–2081, 1995.

[151] D. Smith, A. Holland, and I. Hutchinson, "Random telegraph signals in charge coupled devices," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 530, no. 3, pp. 521–535, 2004.

[152] D. Fiore, *Random Telegraph Signal in CMOS Single Photon Avalanche Diodes*. PhD thesis, Università della Calabria, 2019.

[153] L. Ratti, P. Brogi, G. Collazuol, G.-F. D. Betta, P. Marrocchesi, J. Minga, F. Morsani, L. Pancheri, G. Torilla, and C. Vacchi, "Cross-talk and RTS noise characterization of 1- and 2-tier CMOS SPADs in a 150 nm process," in *2021 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–4, 2021.