EPFL

Thèse n° 8720

High-Performance CMOS SPAD-Based Sensors for Time-of-Flight PET Applications

Présentée le 7 mars 2022

Faculté des sciences et techniques de l'ingénieur
Laboratoire d'architecture quantique
Programme doctoral en microsystèmes et microélectronique

pour l'obtention du grade de Docteur ès Sciences

par

Francesco GRAMUGLIA

Acceptée sur proposition du jury

Prof. A. P. Burg, président du jury
Prof. E. Charbon, Dr C. Bruschini, directeurs de thèse
Prof. F. Zappa, rapporteur
Prof. D. Schaart, rapporteur
Dr P.-A. Besse, rapporteur

École polytechnique fédérale de Lausanne

2022

To my family...

Acknowledgements

First of all, I would like to thank my supervisor Prof. Edoardo Charbon and my co-supervisor Dr. Claudio Bruschini for giving me the unique opportunity of participating in the AQUA lab and performing the research work discussed within this thesis. Thanks to them, I had the opportunity to tackle many research challenges and to pursue my goals, while having access to abundant resources and being surrounded by a very knowledgeable, supportive, and thought-provoking environment. The tough but worth high standard that they put on my path, pushed me to significantly improve my skills and achieve excellent results.

In addition, I would like to particularly thank Dr. Myung-Jae Lee, Dr. Esteban A. Venialgo, Dr. Scott Lindner, and Dr. Emanuele Ripiccini for all the support given during my work and also have been also friends outside the work place. Without their support many achievement of the thesis would have not been possible. Another particular thanks goes to Pouyan Keshavarzian, Ekin Kizilkan, and Ming-Lo Wu to have been unique *brother in arm* during a large part of this work.

Next, I want to thank all the other people that directly or indirectly helped me to succeed in this journey. In particular I thank Andrada A. Muntean that participated to the design of the 3D-stacked MD-SiPM presented in the thesis, and to Brigitte Khan to have patiently helped me with countless bureaucratic problems. Also, I want to thank the other members of AQUA Lab: Dr. Ivan Michel Antolovic, Dr. Samuel Burri, Dr. Feng Liu, Dr. Michael Alan Wayne, Ermanno Bernasconi, Andrei Ardelean, Jad Benserhir, Baris Can Efe, Simone Frasca, Utku Karaca, Tommaso Milanese, Alban Olivier E Morelle, Dr. Andrea Ruffino, Halil Kerim Yildirim, Bedirhan Ilik, Jiuxuan Zhao, Paul Mos, Chang Liu, Chufan Zhou, and Fulvio Martinelli, Dr. Arin Ulku, Dr. Preethi Padmanabhan. In addition, also I want to thank Prof. Lorenzo Paolozzi, Dr. Nicola Luzardi, Fabio Garzetti, Dr. Pieter Vlugter, Dr. Saood Nazir, Ruben Ricca, Gozden Torun, Enrico Casamenti, Dr. Nicolas Descharmes, and Dr. Davide Di Febbo.

Moreover, I would like to thank Dr. Antonino Caizzone to have been a very close friend and to have given several advice and suggestions throughout my entire thesis. I would also like to thank all the Globafoundries team, leaded by Dr. Elgin Queck for the support during the project that allowed the implementation of SPADs in 55 nm BCD technology, and Makoto Motoyoshi for the support during the development of the 3D device presented in the thesis. Further, I want to thank the UTOFPET consortium composed by Prof. Nicola Belcari, Dr. Giancarlo Sportelli, Maria G. Bisogni, Niccolò Camarlinghi, Dr. Pietro Carra, Dr. Esther Ciarrocchi, Dr.

Matteo Morrocchi, Giovanni Franchi; Dr. Karel Deprez, Charlotte Thyssen, Prof. Roel Van Holen, Dr. Mariele Stockhoff, Ewout Vansteenkiste, and Prof. Stefaan Vandenberghe.
Finally and most importantly, I want to thank my family, Lara, and all my closest friends Andrea, Claudio, Clara, Federico, Giuseppe, and Jessica to have supported me and allowed me to arrive to this point.

Lausanne, 13 Jan 2022 F. G.

Abstract

Detecting light, photon by photon, has been possible since the 1930s, with the invention of the photomultiplier tube (PMT). However, it is only since the 1970s, that solid-state single-photon detectors have emerged and only since 2003 that a new technology, known as single-photon avalanche diode (SPAD) has become CMOS-compatible.

The SPAD technology is scalable, low-cost, and it is natively digital, thereby enabling a variety of sensor architectures optimized for applications, such as bioimaging, high energy physics, and medical physics. Thanks to the adaptability of SPADs to many different CMOS technologies, it is possible to achieve systems of different complexity, operating conditions, and, ultimately, cost. The versatility of this device, makes it ideal in very different applications, where it is emerging as the sensor of choice.

This thesis focuses on improving CMOS SPAD performance for time-of-flight PET, while other applications, such as FLIM and LiDAR could also benefit from the technology developed here. We have designed SPAD-based sensors in two different silicon standard technology nodes and we have fully analyzed their performance. We have also explored the optimization of light extraction in inorganic scintillator-based detector modules and direct particle detection and we have studied two applications beyond PET.

The thesis is organized as follows. After a comprehensive description of SPAD technology, we focus on new SPADs structures in standard deep-submicron technology nodes, 180 nm CMOS and 55 nm BCD. The results achieved with these devices show how SPADs can achieve an unprecedented timing performance of 7.5 ps FWHM at room temperature with a sensitivity peak of 55% thanks to a fully integrated front-end readout system that is shown to play an essential role here.

We show that state-of-the-art performance is possible using advanced standard technology nodes, where outstanding timing and sensitivity performance were achieved by systematically engineering the electric field in multiplication and drift regions to achieve optimized carrier generation and transport upon detection.

We have demonstrated the first large-scale 3D-stacked frontside-illuminated (FSI) multi-digital silicon photomultiplier (MD-SiPM), whereas we discuss the sensor structure and its electronic components aimed at modularity and scalability, so as to ultimately achieve large-format SPAD sensors.

To demonstrate specific aspects of the intended applications, *in primis*, TOF-PET, we look at coupling between photodetectors and inorganic scintillators, where gamma radiations, originated from electron-positron annihilations, are detected. Using photonic crystals structures, nanoimprinted on the scintillator output surface, we have demonstrated a 40% improvement in light extraction and 25% better energy resolution in BGO crystals.

Finally, we explore the use of SPADs for direct time-of-flight coincidence measurements of minimum ionizing particles. We achieved an unprecedented 9.4 ps coincidence uncertainty in this novel and very thought-provoking SPAD application, corresponding to a 6 ps Gaussian sigma timing resolution for a single detector.

In summary, this thesis demonstrates new paradigms of SPAD performance, together with a novel 3D integration of FSI SPADs, and novel application solutions. In the future, we thus expect that these achievements will pave the way for further performance improvements and extensive scientific explorations.

Key words: 3D stacking, afterpulsing, avalanche diode, BCD, CMOS, coincidence, dead time, deep sub-micrometer, energy resolution, FSI, inorganic scintillator, light extraction, low noise, multi-digital SiPM, particle detector, PDP, PET, PhC, single-photon, SPAD, sub-65 micrometer, time-of-flight, timing jitter, substrate isolated.

Sommario

La rilevazione della luce, fotone per fotone, è stata resa possibile a partire dagli anni trenta grazie all'invenzione del tubo fotomoltiplicatore (PMT). Comunque, è solamente dagli anni settanta che i rilevatori a stato solido sono cominciati ad emergere e solo dal 2003 che questi dispositivi, conosciuti come diodi di rilevazione a singolo fotone (SPAD) sono diventati compatibili con tecnologie CMOS.

La tecnologia SPAD è scalabile, a basso costo e intrinsecamente digitale. Ciò permette l'implementazione di sensori con architetture ottimizzate per applicazioni come il bioimaging e fisica medica e delle alte energie. Grazie alla flessibilità degli SPAD ad adattarsi a differenti tecnologie CMOS, è possibile ottenere sistemi di diversa complessita, condizioni di funzionamento e costo. La versatilità di questo dispositivo fa sì che sia la scelta ideale per applicazioni molto diverse, diventandone il dispositivo di riferimento.

Questa tesi si concentra sul miglioramento delle prestazioni degli SPAD implementati in tecnologie CMOS per tomografia ad emissione positronica a tempo di volo (TOF-PET). Allo stesso tempo, altre applicazioni come FLIM e LiDAR possono comunque beneficiare della tecnologia sviluppata in questo lavoro di tesi. Sensori basati su SPAD sono stati progettati in due diversi nodi tecnologici in silicio e le loro performance sono state analizzate in maniera completa. Inoltre, anche l'ottimizzazione dell'estrazione di luce da scintillatori inorganici e la rilevazione diretta di particelle sono state oggetto di studio come applicazioni oltre la PET.

La tesi è organizzata nel seguente modo. Dopo una completa descrizione della tecnologia SPAD, l'attenzione viene spostata sulle nuove strutture SPAD realizzate in due nodi tecnologici standard submicrometrici: 180 nm CMOS e 55 nm BCD. I risultati ottenuti con questi dispositivi mostrano come gli SPAD possano raggiungere le prestazioni temporali senza precedenti di 7.5 ps FWHM a temperatura ambiente e una sensitività di picco del 55%. Tutto questo grazie anche all'uso di una elettronica di front-end integrata che assume un ruolo essenziale nel miglioramento delle prestazioni.

Nella tesi viene anche dimostrato come sia possibile ottenere prestazioni pari allo stato dell'arte utilizzando nodi tecnologici standard molto avanzati, in cui ottime prestazioni temporali e

di sensitività sono state ottenute tramite un'ingegnerizzazione sistematica del campo elettrico nelle zone di moltiplicazione e di deriva per l'ottimizzazione della generazione e del trasporto di portatori di carica al fine della rilevazione.

In questa tesi viene anche discussa la dimostrazione del primo sensore 3D implementato con approccio FSI e basato sull'architettura di un photomoltiplicatore in silicio digitale multicanale di grandi dimensioni. La struttura e tutti i suoi componenti sono oggetto di discussione in questa tesi. L'architettura proposta si prefigge come obbiettivo la modularità e scalabilità, avendo come fine l'implementazione di un sensore SPAD di grandi dimensioni.

Per dimostrare aspetti specifici dell'applicazione di riferimento, in primis TOF-PET, è stato effettuato uno studio sull'accoppiamento tra fotorilevatori e scintillatori inorganici. Questi ultimi corrispondono al luogo dove viene rilevata una radiazione gamma emessa dall'annichilazione di un elettrone con un positrone. Usando delle strutture di cristalli fotonici, fabbricati tramite tecniche di nanoimprinting sulla superficie di uscita dello scintillatore, è stato possibile dimostrare un guadagno in estrazione di luce del 40% e un corrispondente guadagno in risoluzione di energia del 25% utilizzando cristalli di BGO.

Infine, è stato esplorato l'uso di SPAD per la misura diretta di tempo di volo in coicidenza di particelle al minimo di ionizzazione (MIP). Questo studio ha prodotto dei risultati senza precedenti mostrando una incertezza sul tempo di coincidenza di 9.4 ps per questa nuova e molto interessante applicazione per gli SPAD. Questo risultato corrisponde ad una risoluzione temporale del singolo rilevatore di 6 ps (sigma Gaussiana).
Per concludere, in questa tesi sono stati dimostrati nuovi paradigmi per le prestazioni degli SPAD insieme a un sensore integrato in 3D e a nuove soluzioni applicative. Sicuramente questi risultati spianeranno la strada verso ulteriori miglioramenti in termini di prestazioni e verso più ampie esplorazioni scientifiche.

Parole chiave: 3D stacking, afterpulsing, basso rumore, BCD, CMOS, coincidenza, dead time, diodo a valanga, estrazione di luce, FSI, multi-digital SiPM, PDP, PET, PhC, precisione temporale, risoluzione in energia, rivelatori di particelle, scintillatori inorganici, singolo fotone, SPAD, sub-65 micrometri, sub-micrometrico, substrato isolato, tempo di volo.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Detectors for Single-Photon Applications

The quantization of light represents one of the most influential theories of modern physics. It demonstrates that light is quantized in individual packets of energy named photons [1]–[4]. The term *photon*, first introduced by G.N. Lewis in 1926 [5], is defined as an elementary excitation of a single mode of the quantized electromagnetic field. The concept of quantized radiation was initially introduced by Planck, followed by Einstein and Compton in 1900, 1905, and 1926 respectively. However, only after a few years, in 1927, a formal quantization of the electromagnetic field was proposed by Dirac [6], [7].
The energy of a single photon in the wavelength range comprised between near-ultraviolet and near-infrared is of around 10^{-19} J, or few electron volts (eV).
In general, the energy of a photon can be calculated as:

$$E_{ph} = h\nu = \frac{hc}{\lambda}, \tag{1.1}$$

where E_{ph} is the photon's energy, h is the Planck's constant, ν the photon mode frequency, c the speed of light in a vacuum, and λ the photon's wavelength.

An important quantum phenomenon theorized in the same period by Einstein concerns the interaction between electromagnetic radiation (such as light) and a material. This type of effect, known as the photoelectric effect, consists of the emission of electrons from or within a material when it absorbs electromagnetic radiations [1], [8], [9].

Electrons bound to atoms occur in specific electronic configurations. The highest energy configuration (or energy band) normally occupied by electrons for a given material is known as the valence band, and the number of these electrons determines the material's electrical conductivity. In particular, semiconductors generally have their valence bands filled, but very little energy is required to excite an electron from the valence band to the next allowed energy band—known as the conduction band. Thus, in this higher energy level, an electron is relatively free, and it can cause a current flow. The electrons excited to the conduction band by the absorption of electromagnetic radiation are defined as photoelectrons.
For example, if we consider silicon (Si), the energy needed to excite an electron to the conduction band (energy gap) is about 1.12 eV.

During the years, it has been understood that it is possible to detect light through the charge generated during the optical photons' interaction with matter. Consequently, there has been an ongoing effort to develop devices that can detect always increasingly lower light level. This trend, mainly driven by the semiconductor device community, led to the introduction of various devices capable of providing a measurable electrical signal from a single photoelectron. The need to detect an extremely low light level, up to single photons, grew out of interest in

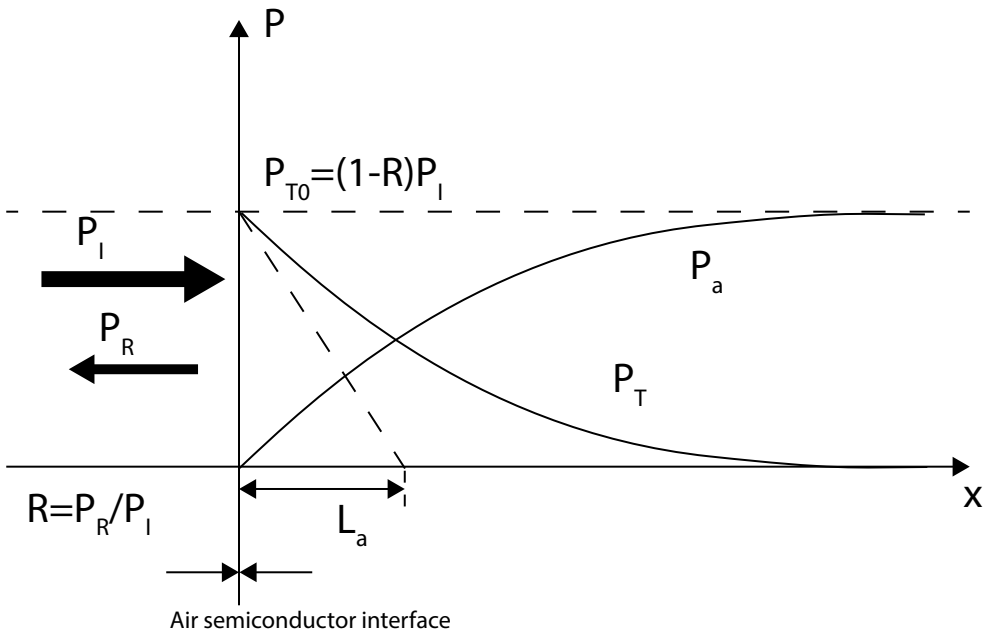


Figure 1.1: Absorption of photons in materials. For a certain incident power (P_I) on an air-material interface, part of this power is transmitted (P_{T0}) and part is reflected back (P_R). We define the reflectance (R) as the ratio between the reflected power and the incident power. Inside the material, the light is transmitted following an exponential decay and we define the optical absorption length (L_a) as the decay constant of this exponential trend. This trend also defines the power absorbed in the material (P_a).

exploring a broad spectrum of emerging applications, such as fluorescence microscopy or radiation detection through scintillators.

As mentioned before, an essential concept to be taken into account when detecting optical photons is their interaction with a material (Fig. 1.1). Optical photons passing through matter have a statistical probability of being transmitted that decays exponentially. The transmitted power (P_T) can be expressed by the following differential equation along the depth x of the material:

$$-dP_T = \alpha P_T dx = P_T \frac{dx}{L_a}. \tag{1.2}$$

Integrating it along the depth x, we obtain:

$$P_T = P_{T0} e^{-\alpha x} = P_{T0} e^{-x/L_a}. \tag{1.3}$$

From the previous result, we can derive the absorbed power in the material as:

$$P_a = P_{T0} - P_T = P_{T0}(1 - e^{-\alpha x}) = P_{T0}(1 - e^{-x/L_a}), \tag{1.4}$$

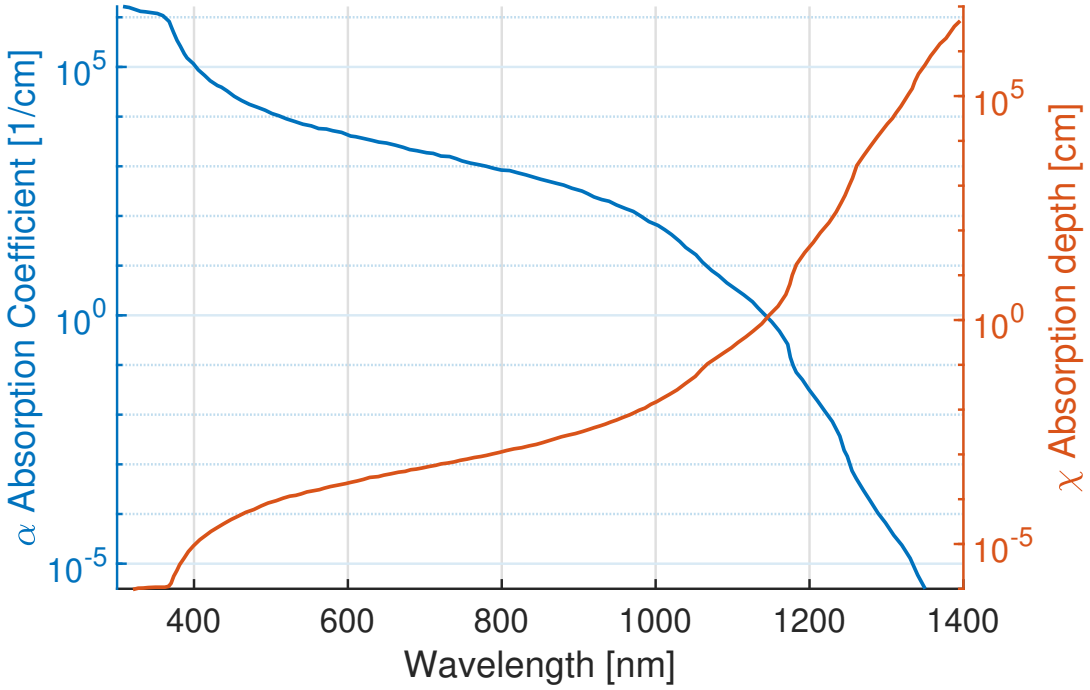


Figure 1.2: Absorption coefficient and absorption depth as a function of wavelength for silicon [12].

where P_{T0} is the power transmitted at the optical interface, P_T is the transmitted power as a function of the depth, P_a is the absorbed power as a function of the depth, α is the optical absorption coefficient, and L_a is the optical absorption length. It is also important to note that the optical absorption coefficient (and the absorption length) strongly depends on the wavelength. Fig. 1.2 and Fig. 1.3 show an example of the wavelength dependency of the absorption coefficient for several semiconductor materials.
Usually, single-photon detectors provide a sufficiently high gain and low noise, enabling the detection of a single photon [4], [10], [11]. Various strategies are used to increase the signal gain on single-photon detectors. In solid-state devices, the primary mechanism used for single-photon detection relies on converting optical photons into electron-hole pairs and using high voltage to induce a multiplication, or avalanche, process to convert a single charge into a macroscopic current pulse.

Not only solid state detectors, based on semiconductors materials are used for single-photon detection (or more in general for light detection). In the following paragraphs, two examples of presently available single-photon detectors are briefly presented.

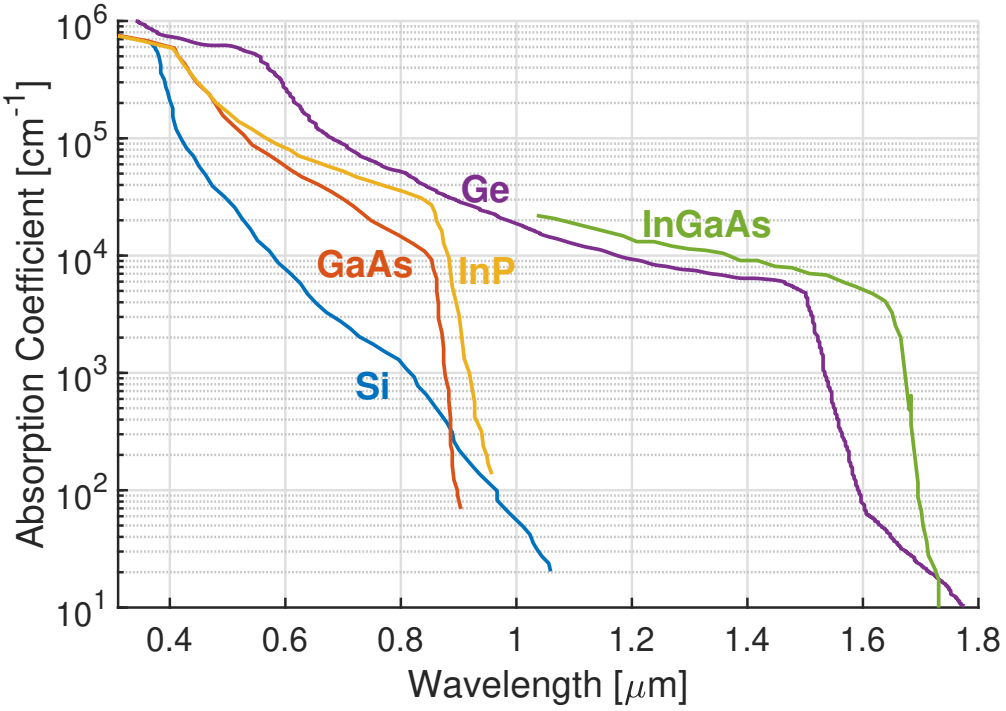


Figure 1.3: Absorption coefficient for several semiconductors commonly used for the fabrication of optical detectors [13].

1.1.1 Photomultiplier Tubes

Photomultiplier tubes (PMTs), invented in the late 1930s in the RCA laboratories, represented the first device capable of detecting single photons [14]. These devices rely on a photocathode as the source of primary photoelectrons. The photocathodes are made of low work function materials that convert an impinging optical photon into an electron through the photoelectric effect. This electron is then freed in a vacuum chamber. A high electric field is generated by applying a high voltage across the chamber. The field accelerates the electrons toward structures called dynodes. When a primary electron hits a dynode, secondary electrons are generated. The generated electrons are, in turn, accelerated by the electric field. Proper placement of dynodes along the path of the electrons will create a cascade of secondary emission events that multiplies the carriers. At the end of the multiplication chain, the charge is collected at the anode generating a macroscopic current pulse,which can be detected by ordinary electronics. A schematic view of a PMT is shown in Fig. 1.4. The PMT performance comprises a relatively high quantum efficiency (typically on the order of 30%) [15], a large sensitive area, low noise [16], and reasonable single-photon timing resolution (SPTR) (typical value of 200-300 ps and down to tens of picoseconds in the case of MCP-PMTs) [17], [18].
PMTs have some limitations:

- their structure is bulky

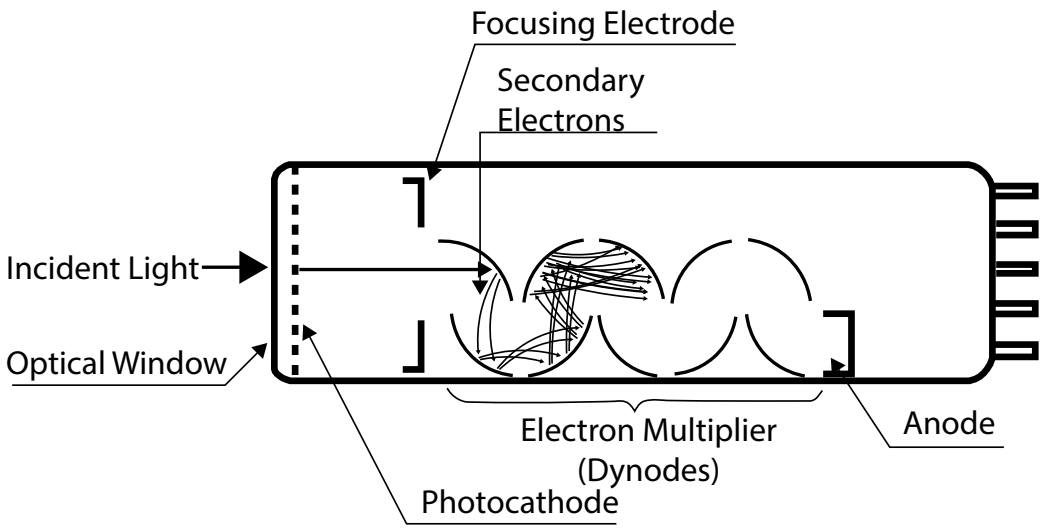


Figure 1.4: Photomultiplier tube schematic view.

- they need a very high bias voltage ($\approx 100V - 3k\ V$)
- they are sensitive to magnetic fields
- the photocathode sensitivity typically does not exceed 30%

All these limitations pushed the research to find valuable solid-state alternatives.

1.1.2 Superconductive Nanowire Single-Photon Detectors

Superconducting nanowire single-photon detectors (SNSPDs), first published in 2001 [19] are usually implemented using alloys (e.g., NbTiN, NbN) that show a critical temperature in the 1-11 K temperature range. When cooled below their critical temperature, they become superconductive. A bias current, usually on the order of 1 mA to 10 mA, is injected into the structure. This bias current is set just below the limit that causes a state switching of the device (i.e., the critical current). When a photon hits the superconductive material, it deposits its energy, creating a resistive spot (the *hotspot*). The hotspot region forces the bias current to flow around it, increasing the local current density. If the current density passes the critical value, the device quickly switches, becoming highly resistive (resistance on the order of 100 kΩ) [20], [21]. The current used for biasing the device is then redirected into a resistive load (usually 50 Ω), and a voltage pulse is captured by the front-end electronics [20]. The small value of the load is needed to allow the device to recover the superconductive state. Indeed, with a higher resistance, the current flowing into the device does not remain below the critical value, and the device cannot recover the superconductive state properly. The pulse generated by an SNSPD does not exceed 1 mV. This low signal value implies the use of a very high-performance

Table 1.1: Single-photon detector summary of typical performance.

Detector	Active area	Sensitivity range	Operating Tempera-ture	Noise	Sensitivity	Timing Jitter	Afterpulsing probability [%]	Array size
InGaAs/InP SPAD[a]	10-25 μm diam.	920-1700 nm	225-253 K	$\geq 500 cps$	≤ 40% PDP	≥46 ps	2-3 @20 μs hold-off	256×64
Ge-on-Si SPAD[b]	100 μm diam.	≤1550 nm	100 K	1.4×10^5 cps	36% @1310 nm PDP	310 ps @ 1310 nm	2 @20 μs hold-off	NA
SNSPD	18 μm diam.	≤ 1550 nm	≤11 K	≤ 10 cps	≥70% PDP	3 ps (typ. 15-20 ps)	NA	32 × 32
PMT	≥0.24 cm^2	115-1700 nm	RT	≤80 kcps @2600 cm^2	~ 30% QE	300 ps	NA	NA
Si-SPAD	≤ 500 μm diam.	250-1000 nm	RT	≤100 cps	~ 60% PDP	≥28 ps	≤1%	≤3.2 Mpxl

[a] according to [30]–[33];[b] according to [34].

front-end capable of raising this pulse to a level at which a low threshold comparator can detect it. The pulse is then translated into a digital signal which is sent to the downstream electronics. SNSPDs are capable of providing excellent performance over a broad set of wavelengths, featuring a photon detection efficiency above 90% [22] and a dark count rate lower than 0.01cps [23], together with a timing precision of 2.6 ps full width half maximum (FWHM) [24] and a count rate in the order of Gcps [25]. Nevertheless, these characteristics have never been demonstrated together on the same device [21], [26]. One of the main reasons for this separation is an evident technological trade-off coming from the fabrication and operation of this detector [27]. Moreover, it is worth noting that the sensitivity of these devices shows a dependency on the polarization of the light [28]. This characteristic is usually not desirable when the transferred information is encoded in degrees of freedom like path and time [26], [29]. Further, the need for a cryostat increases the cost of the setup substantially.

1.1.3 Single-Photon Avalanche Diode

Another kind of device available for single-photon detection relies on the photoelectric effect in semiconductors. This device is named single-photon avalanche diode, or simply SPAD. This device consist of a $p-n$ junction conveniently reversed biased that allows the detection of single photons thanks to an extremely high gain. A detailed description of this device is available in Chapter 2 of this thesis.

Table 1.1 reports a summary of typical performance of single-photon detectors.

1.2 Application Description

To provide the reader with a little background context, in this section, two applications relying on SPAD-based detectors are presented and briefly discussed.

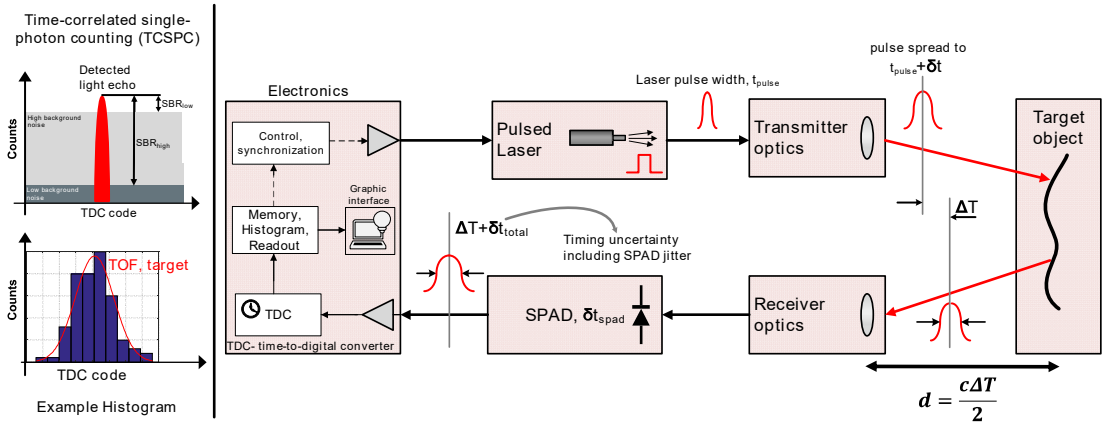


Figure 1.5: High-level block diagram of a dTOF sensing system. The image was extracted from [35].

The two applications presented in this paragraph are time-of-flight positron emission tomography (TOF-PET) and light detection and ranging (LiDAR). Both these applications rely on the time-of-flight concept, although they require different specifications.

1.2.1 LiDAR

General Concepts

Although the roots of this application go back to the 1930s, only after the invention of lasers in the early 1960s, light detection and ranging (LiDAR) started evolving [36]. The first experiments were performed to range distances, but the number of applications suitable for this technique started growing quickly. Initially, most of the LiDAR applications were for military purposes. However, from the early 2000s, the attention moved to automotive applications, such as ADAS (Advanced Driver Assistance Systems) [37]. More recently, large interest moved on the development of LiDAR systems for other applications, such as consumers, augmented and virtual reality, and environmental mapping to mention a few.

In general, LiDAR is a 3D imaging technique based on the measurement of the time-of-flight of optical photons reflected from an object. It consists of a topographic method to create a 3D graphical representation of a physical target. Its working principle is based on the illumination of a scene employing a laser and detecting the photons that reflect off the target. By calculating the time-of-arrival of these photons, the depth information z is extracted, and a representation model of the scene, with x, y, z information, becomes possible. The z information is calculated considering the time-of-flight (Δt) of the detected photons:

$$z = v\frac{\Delta t}{2} = \frac{c}{n}\frac{\Delta t}{2}, \tag{1.5}$$

where v is the speed of light in the medium, c the speed of light in vacuum, and n the refractive index of the medium.

In this thesis we will briefly discuss only one technique used in LiDAR for depth sensing: the direct time-of-flight (dTOF). The operating principle of this technique is as follows. A laser is pulsed using a clock signal synchronous to the detection system, and the light pulses are directed to the scene. The optical photons contained in the laser pulse undergo reflections and/or scattering along their path across the scene. Those photons that are reflected by the scene can be detected using a photodetector. There are several options for the kind of photodetector to be used. Several works demonstrated that a good solution is adopting multichannel digitel silicon photomultipliers (MD-SiPMs) (see 2.3.3) [35], [38]–[40]. The z information (connected to the x and y position in the array) is processed (on-chip or off-chip depending on the used system) to reconstruct a 3D map of the scene [41]. Fig. 1.5 provides a representation of the LiDAR principle of operation.

For a more detailed description of this application and its open challenges, the author suggests the review of [35], [36], [38], [39] to the interested reader.

1.2.2 Time-of-Flight Positron Emission Tomography

General Concepts

Positron emission tomography (PET) is the most relevant molecular imaging technique available today. The reason for this is its high molecular sensitivity and large tissue penetration [42]–[48]. Molecular imaging started evolving in the mid-90s, allowing the *in vivo* visualization and quantification of biological processes at the cellular and molecular level by using so-called *molecular probes* [49]. *Molecular probes* are defined as biocompatible image contrast enhancement agents that accumulate and stay in a specific target for a certain time. In the specific case of PET, the used molecular probes are labeled with a positron-emitting radioisotope. Many different molecular probes have been developed during the years for the most diverse objectives. Some examples are ^{18}F-fludeoxyglucose (^{18}F-FDG) or ^{62}Cu-labeled copper(II) pyruvaldehyde bis(N^4-methylthiosemicarbazone), Cu(PTSM), etc. [42], [43], [49]. Among the several molecular probes available, one of the best known is ^{18}F-FDG [42], [49], [50], widely employed in oncology. This sugar-like compound is used to detect cancer cells in high metabolism growing tumor masses.
PET is a so-called *emission mode* tomography, in which the subject under test emits the signal, and the PET scanner is used to detect it. As mentioned, the molecular probes used are β^+ emitters. When a solution of a molecular probe such as ^{18}F-FDG is injected into a patient undergoing a PET exam, the F-18 emits positrons inside the subject's body. The positron travels for a certain distance, called *positron range*, (~2 mm) before encountering and combining with an electron to form a positronium [51] (see Fig 1.6). This system is unstable, and after a time on the order of ~ 100 ps, the $e^- - \beta^+$ annihilation occurs. This causes the emission of two almost collinear γ-photons, each with an energy of E_γ=511 keV (the equivalent of twice the electron rest masses) and traveling in opposite direction [52]. The detection of these two γ-photons can allow the localization of the annihilation point. For this purpose, the subject

under test is placed in a ring scanner. The scanner is fully covered by detectors aiming to detect the gammas coming from the patient and extract the so-called *line of response* (LOR). The LOR is defined as the imaginary line connecting the two points in the scanner where a gamma pair is detected and ideally passing through the generation point (Fig. 1.7 *top-left*). The gammas can undergo Compton scattering along their path towards the detector ring. These events deviate the gamma from its original trajectory, potentially causing an error in the reconstruction of the LOR. During Compton scattering, the gamma deposits part of its energy (E_C). Thus the scattered γ-photon has an energy equal to:

$$E'_\gamma = E_\gamma - E_C. \tag{1.6}$$

Consequently, it is possible to identify the events that have undergone Compton scattering and discard them, by estimating the energy of the detected gammas.
Fig. 1.7 shows several possible scenarios which can take place during a PET acquisition.

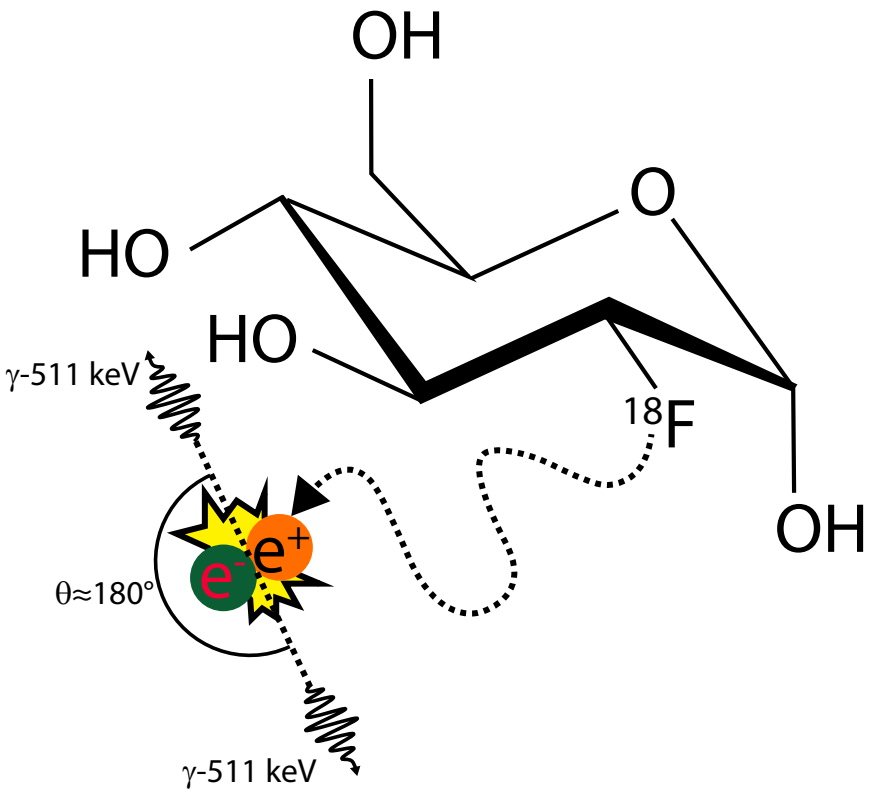


Figure 1.6: FDG molecule and positron emission. Annihilation and generation of gammas back-to-back.

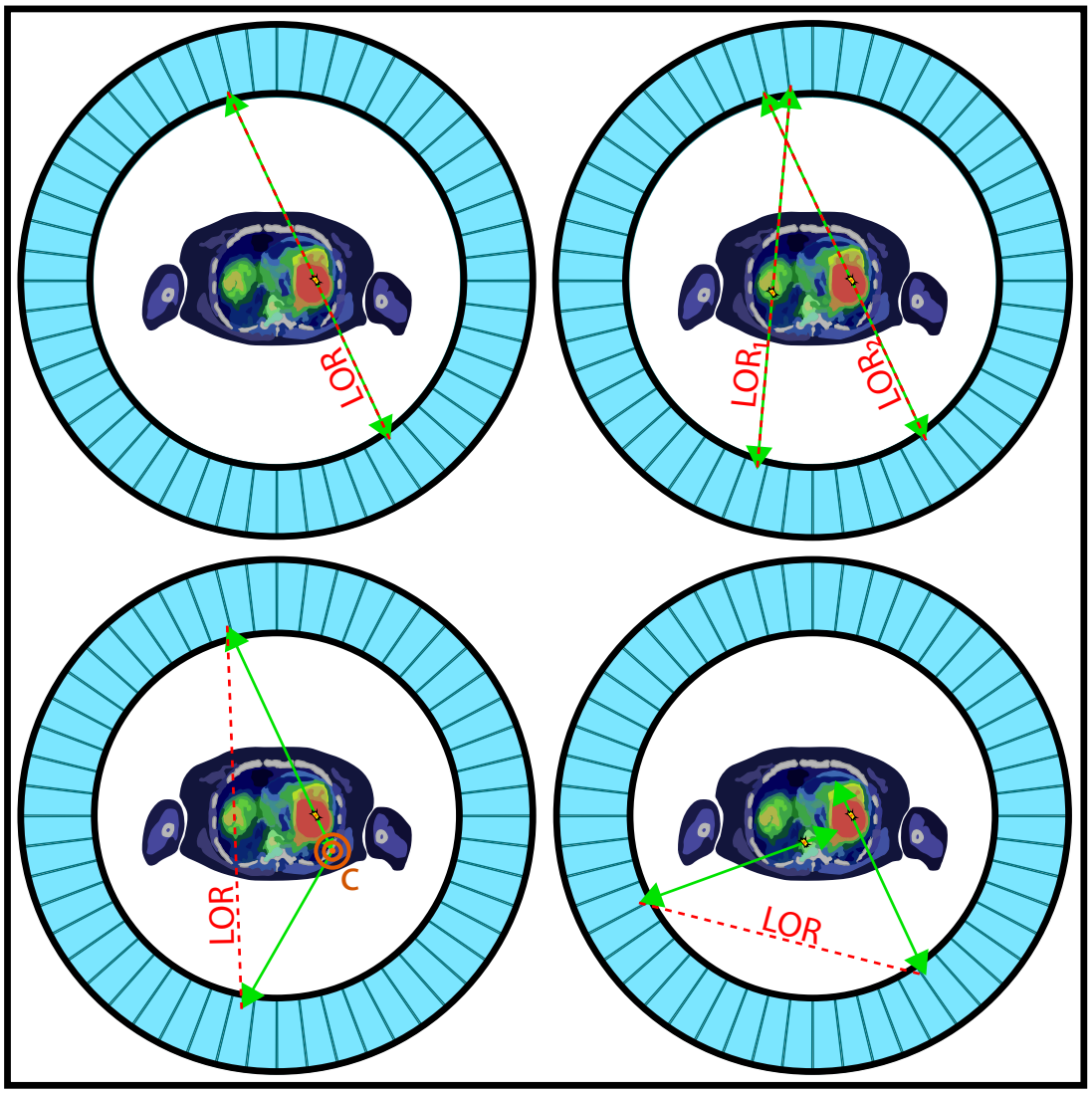


Figure 1.7: PET working principle. During the acquisition different scenarios can be identified: *Top left*: the proper coincidence happens when the two gammas interact with two detectors in the ring depositing all their energy, and the LOR passes through the annihilation point. *Top right*: a multiple coincidence takes place when multiple pairs of gammas are emitted by two annihilation points at the same (or very close in) time and they hit detectors that fall within the same field of view. *bottom left*: When at least one of the gammas undergoes Compton scattering (C), it can lead to a false LOR reconstruction. *Bottom right*: when two annihilations take place very close in time but just one of the two γ-photons of each one of them reaches the detector ring, while the other gets absorbed along the path, we talk about random coincidences.

The Benefits of Time-of-Flight

As clearly depicted on the right-hand side of Fig. 1.8, the probability of the gamma emission position along the LOR is equally distributed in a nonTOF-PET scanner. In this case, a large

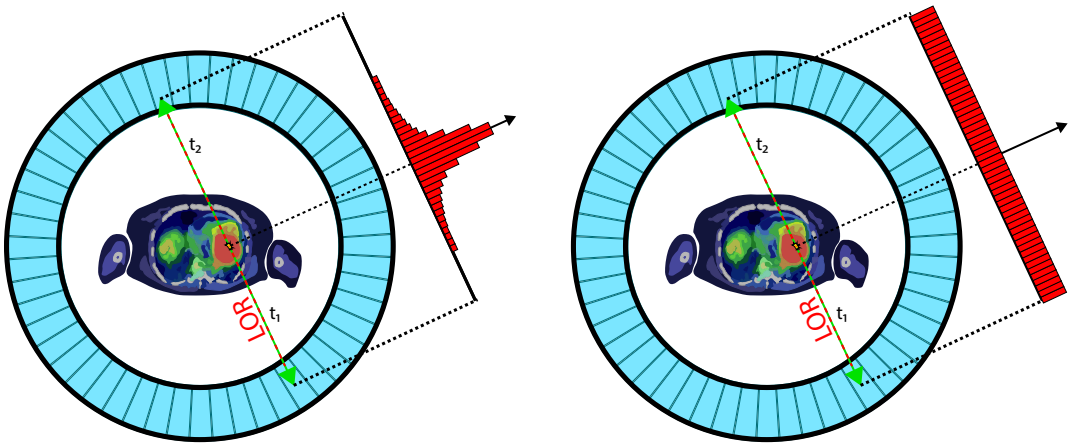


Figure 1.8: *Left*: Time-of-Flight (TOF) principle in PET. The use of the TOF information allows to improve the localization accuracy of the generation point along the LOR, improving the SNR in the reconstructed image. *Right*: in non-TOF-PET, the probability of where the annihilation took place is uniformily distributed along the LOR.

number of LORs is then needed to obtain a high resolution image with standard reconstruction algorithms. This means that a high computational load and long acquisition time are required. Since the early days of PET (the 1960s), the use of TOF information was recognized to be helpful to improve the SNR in reconstructed images and to reduce random coincidence event rates [53], [54]. However, the first systems incorporating TOF information were developed only twenty years later [55]–[60].
In TOF-PET, the difference in the arrival times of the two photons is measured with high precision, which helps localize the emission point along the LOR within a small region (Fig. 1.8, *left*). The uncertainty in this localization depends on the system coincidence timing resolution (CTR), which is measured as the full width at half maximum of the histogram of TOF measurements generated by a point source (i.e., the timing spectrum). The corresponding uncertainty in spatial localization (Δx) along the LOR is given by:

$$\Delta x = c\frac{CTR}{2}. \tag{1.7}$$

If Δx is the same as or similar to the positron range, then in principle, image reconstruction algorithms are not needed [61]. This value of Δx is on the order of 2 mm and corresponds to a travel time of ~ 10 ps [62], [63]. In general, the improvement of signal to noise ratio (SNR) is proportional to $\sqrt{D/\Delta x}$, where D is the diameter of the scanner.

TOF-PET Module: Principle and Limitations

The detection of the gammas generated by the $e^- - \beta^+$ annihilation can be performed with various types of detector, such as solid-state detectors, ionization chambers, inorganic scintillator-

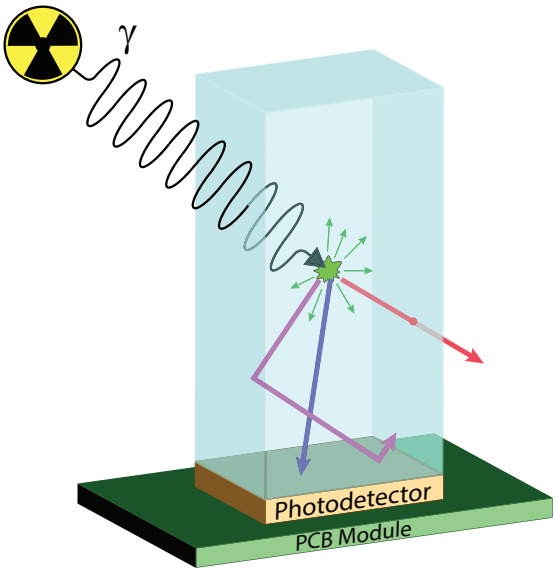


Figure 1.9: A PET module is typically composed of a scintillation crystal instrumented with a photodetector module. The crystal irradiated with gammas emits light that can be detected by the photodetector.

based detectors, etc. In PET, the detection widely relies on the use of inorganic scintillators because they offer the best compromise between energy, timing, and spatial resolution as well as a high detection efficiency for 511 keV γ-photons. These materials have the property of absorbing the incoming radiation and emitting optical photons following a fast exponential timing evolution with a decay time constant τ_d typical of each scintillator [64]. The emission is isotropic inside the scintillating material, and the amount of light is proportional to the amount of energy deposited in the crystal by the incoming radiation. The proportionality factor is called *light yield* of the scintillator. Moreover, the light is emitted within a specific range of wavelengths, named emission spectrum [65], [66]. The emission spectrum as well is characteristic of the scintillator used, and needs to be taken into account when designing a detection module (see Table 1.3).
The light burst (or pulse), generated by a scintillator, reveals when a gamma interacts in the crystal, and it can be detected by instrumenting one or more surfaces of the scintillator with a photodetector. Moreover, if the photodetector can quantify the amount of detected light, it is possible to estimate the energy of the incoming gamma, knowing the *light yield*. Fig. 1.9 shows the schematic view of a typical module used in PET applications. In this simplified scheme, the scintillator is optically coupled to a photodetector and a PCB module is used to read out and process the signal.
Originally, PMTs were the most widely used photodetectors, but nowadays, silicon-based

Table 1.2: Comparison table of typical detectors used in PET applications.

Detector	maximum sensitivity [%]	Time resolution [ps]	Noise	Availability	System Integration
PMT	~30 (QE)	300	≤0.3	high	indirect
MCP-PMT	≥20	~20	~10 nA (leakage)	high	indirect
aSiPM	50 (PDE)[b]	125-250[c]	50-150 kHz/mm^2 [b]	high	indirect
dSiPM	~30 (PDE)	101-113[a]	96 kHz/mm^2	low	direct

[a] according to [70]; [b] according to [71];[c] according to [72];[d] according to [73];[e] according to [74].

devices such as Silicon Photomultipliers (SiPMs) have moved into a prominent position. This trend is due to their more compact structure, lower supply voltage, and robustness to magnetic fields. The light sensitivity of both PMTs and SiPMs does usually not exceed 30% at short wavelengths (e.g., 300 nm) in commercial devices [67], [68] and peaks at around 50–60% in the visible for SiPMs [69]. Table 1.2 shows a comparison of detector used for PET applications.

The amount of light generated during the scintillation process in inorganic scintillators is in the range of 8–60 k photons/MeV (e.g., the LYSO light yield is around 30 k photons/MeV) [65], [66]. However, in standard configurations, only a low percentage of this light reaches the photodetector. The maximization of the light extracted from the crystal becomes thus crucial, given that both energy and time resolution strongly depend on the amount of detected light. Moreover, extracting more light enables a more accurate estimation of the energy deposited in the crystal by the incoming radiation, allowing better discrimination of events that underwent Compton scattering. Concerning the timing performance, the CTR has been demonstrated to be proportional to the inverse square root of the number of detected photons [68], [75], [76]. The limitations on light extraction in inorganic scintillators are discussed in detail in Chapter 5, together with a novel solution developed during this thesis.

Table 1.3 shows a comparison of the main parameters of common scintillators used for PET application.

Together with a reduced amount of light, another limitation of this kind of modules is shown in Fig. 1.10. The γ-photons travel at c both in air and inside the scintillation material. On the other hand, optical photons are affected by the medium's optical properties. For example, inorganic scintillators used in PET applications have a relatively high refractive index (≥1.8), which influences the propagation of the light photons when they enter the material. Generally speaking, the propagation speed of an optical photon of a particular wavelength λ, in a medium with refractive index n is equal to:

$$v(\lambda) = \frac{c}{n(\lambda)}. \tag{1.8}$$

Hence, light propagates slower than the gammas inside the scintillation material.

Furthermore, the γ-photons interaction with matter is a probabilistic event that follows an exponential probability function [77]. Therefore, a gamma can interact and deposit its energy anywhere inside the scintillator. In Fig. 1.10, two extreme cases of gamma interaction are shown in a simplified model: the gamma immediately interacts when it enters the scintillator (*bottom*); the gamma interacts right at the interface with the photodetector (*top*). In the two

Table 1.3: Parameters of Common inorganic scintillators utilized in PET [49].

Material	Linear attenuation [cm^{-1}]	Light yield [ph/keV]	Refractive index	Emission-peak wavelength [nm]	Primary scintillation decay time [ns]	Primary scintillation rise time [ps]
NaI(Tl)	0.34	38	1.85	415	250	NA
BGO	0.96	8-10	2.15	480	300	NA
LYSO	0.83	27.6	1.81	420	45	90
$LaBr_3$(Ce)	0.46	63	1.9	380	16	280

presented cases, the photodetector will estimate the arrival time of the gamma with the time at which the emitted light is detected. This difference will be:

$$\Delta t_{err} = \frac{L_{scint}}{v(\lambda)} - \frac{L_{scint}}{c} = L_{scint}\frac{n(\lambda)-1}{c}, \tag{1.9}$$

Where L_{scint} is the length of the scintillator. It is easy to understand how this creates a timing estimation error that sets a limit on the maximum time resolution achievable on one crystal. This uncertainty is linked to the so-called depth of interaction (DOI) and it is related to the size (L_{scint}) and the optical properties of the scintillator used. Therefore, the adoption of smaller scintillators reduces this uncertainty, but at the same time it decreases the module efficiency by reducing the probability that the gamma interacts in the crystal. On the other hand, the introduction of DOI estimation algorithms can help mitigate this problem with longer scintillators, without compromising the gamma detection efficiency of the module.
Moreover, as explained in [49], [79], [80], the measurement of multiple timestamps together with a higher photodetector granularity significantly reduces the timing uncertainty, improving the CTR.

1.2.3 Relevant Detector Parameters in PET

In PET applications there are four main parameters that should be taken into account during the detector design phase:

- energy resolution
- timing resolution
- spatial resolution
- sensitivity

The energy resolution of a detector is measured by exposing the detector to a known γ source, and building a histogram from the measured signal. This histogram, known as gamma source energy spectrum, will show the probability the profile of the energy deposited in the detector by the γ-photon. Fig. 1.11 shows a typical energy spectrum.
The energy resolution is then defined as:

$$ER = \frac{FWHM_\gamma}{E_\gamma} \times 100(\%), \tag{1.10}$$

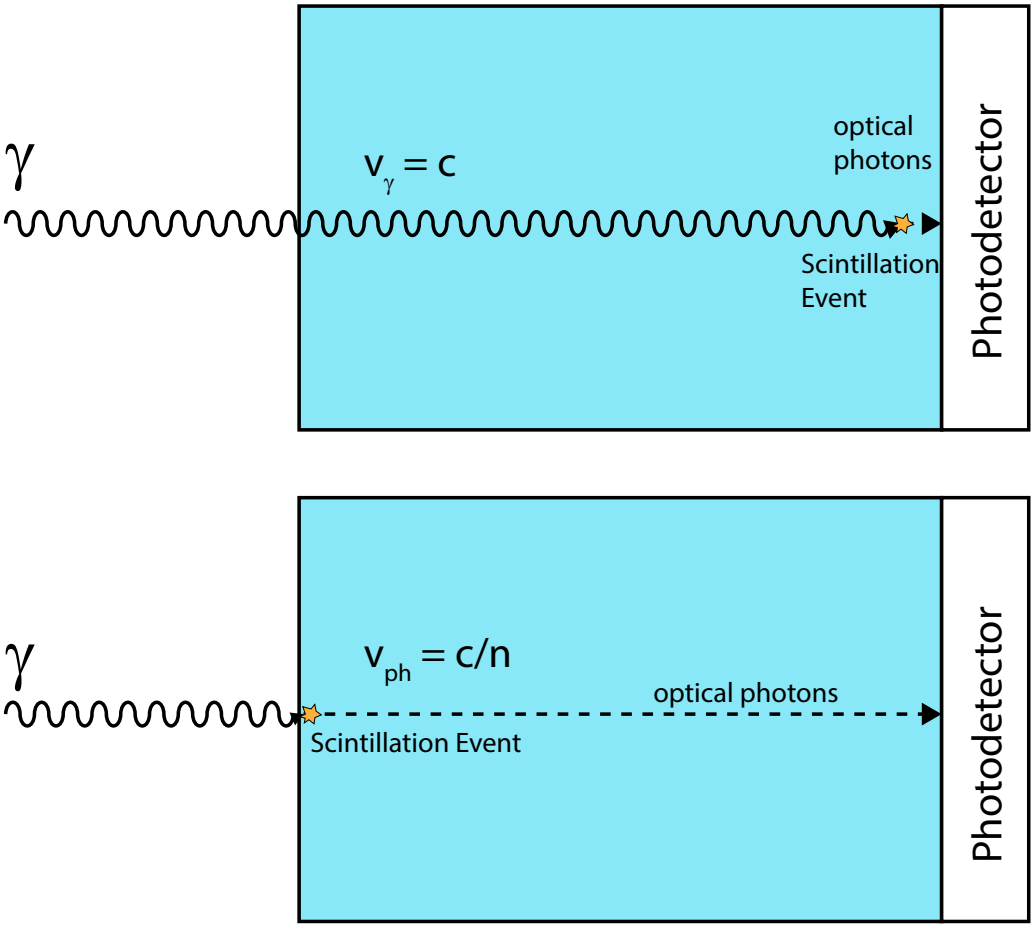


Figure 1.10: Intrinsic time limitation of scintillator-based detectors. Figure adapted from [78].

where ER is the energy resolution, E_γ is the energy of the incoming γ-photon, and $FWHM_\gamma$ is the FWHM of the energy spectrum photopeak (Fig. 1.11). In PET this parameter is used to determine an energy-window and filter out the Compton scattered events and the events that did not deposit the full 511 keV energy in the detector material. Thus an ER as small as possible is preferable to better reject the lower energy gamma-rays, due to Compton scattering. To optimize the energy resolution, a detector should maximize the histogram counts in the photopeak region. This maximization can be achieved with a proper selection of the material used for the detector, i.e., a small Compton cross section. Moreover, also the optimization of the optical interface between the scintillator and detector can improve the light extracted from the crystal, thus improving the final ER (see Chapter 5). Further, the estimation of the deposited energy is linked to the detector sensitivity and capability in counting single photons (minimal count losses). Thus, an optimization of the PDP and dead time of the device helps improving this first parameter.

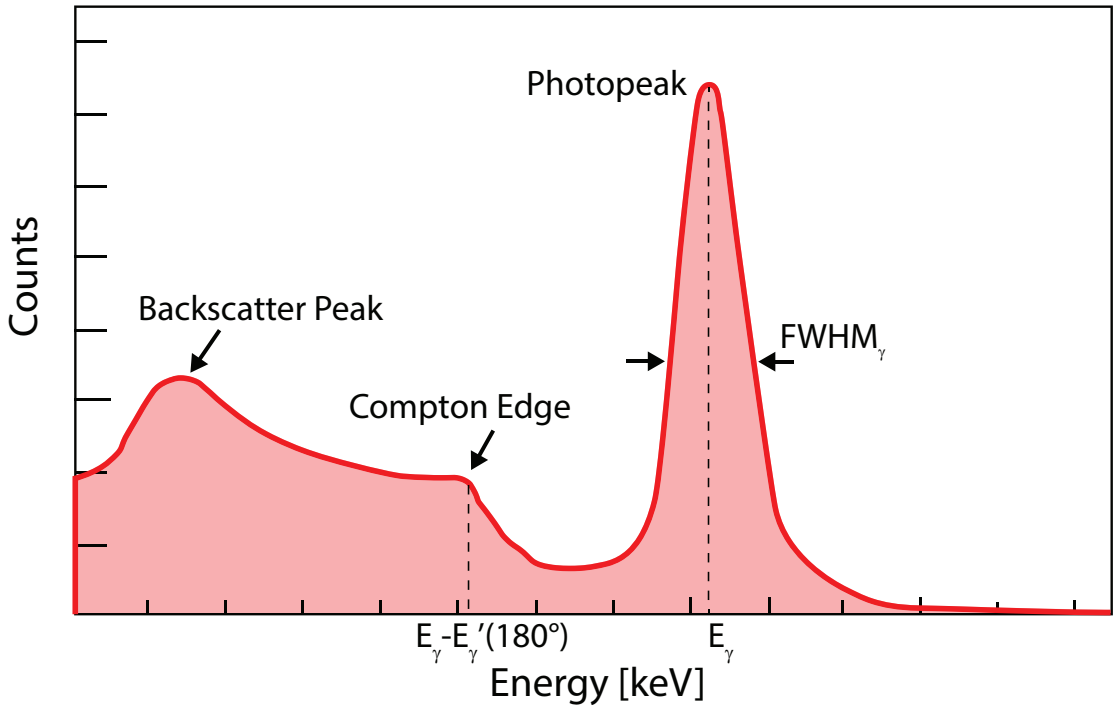


Figure 1.11: Representation of the energy spectrum of a detector and its energy resolution.

Another important parameter to consider is timing resolution. Let us analyze a coincidence experiment in which two detector face each other and a radioactive source emitting back-to-back γ-photons is placed between them. In this configuration, all the detected events are named as *singles* and labeled with a timestamp value after being filtered by an energy windowing scheme as described before. If we build a histogram with the time difference obtained by events in coincidence, we obtain a distribution characterized by a certain spread measured in terms of FWHM, the system CTR. This information can be used to set a time window used to filter background events, such as some random coincidence. Moreover, as shown in Section 1.2.2, the timing measurement provides information on the location of the emission point. It is known that reconstruction algorithms based on the iterative methods converge faster with smaller CTRs. Moreover, as stated in [61], with extremely high CTR the image can be obtained without a reconstruction algorithm. Many detector module parameters influence the CTR, such as the sensitivity of the detector, the size of the scintillator (as mentioned in the previous section), the timing jitter of the single detectors, etc.

The third parameter is the spatial resolution of the detector module. As shown in the [81], the spatial resolution (Γ) for a point source located at a distance r from the center of the scanner ring can be quantified as:

$$\Gamma = 1.25\sqrt{\frac{d}{2}^2 + s^2 + (0.0044R)^2 + \frac{(12.5r)^2}{r^2+R^2}}, \tag{1.11}$$

where d is the crystal width, s the *positron range*, R the detector ring radius. This uncertainty can cause the so called parallax error, which in turn causes a nonuniform resolution along the PET scanner radius. It is clear how the crystal and photodetector size directly influence the PET scanner spatial resolution.

Finally, the last parameter which we should consider is the system sensitivity (S) that corresponds to the probability that the system detects a pair of back-to-back gammas generated from $e^- - \beta^+$ annihilation events. This parameter is defined as:

$$S = \epsilon^2 \times G_a \times AF, \tag{1.12}$$

where ϵ is the single gamma module detection efficiency, G_a is the geometrical acceptance, and AF is the absorption factor that determines a loss in sensitivity because of absorption in the subject used.
Parameter ϵ depends on the probability that a γ-photon interacts in the scintillator (P_γ), which depends on the geometrical characteristics of the crystal and its stopping power, on the light extraction and collection efficiency of the detector, and on the photon detection efficiency (PDE). The latter measures the probability that an impinging optical photon is detected. This parameter strongly depends on the wavelength of impinging photons and on the geometrical parameter of the photodetector used:

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

where $PDP(\lambda)$ is the photon detection probability of the detector and FF its fill factor. These parameters will be thoroughly discussed in the next chapter. Moreover, the PDP is a key parameter that has been taken into account during the design and testing of the several devices in the present thesis.

We move now the attention on how the described system parameters translate in the implementation of a scintillator-based detector, focusing on the parameter needed by the photodetector. As discussed, it is important that detectors used for PET applications have high gamma-ray detection efficiency. As the scintillator is instrumented with a photodetector, the gamma-ray detection efficiency depends on the photon detection efficiency (PDE) of the photodetector. In case of a pixelated device like a SiPM (see Section 2.3), this parameter depends on the photon detection efficiency (PDP) of the single SPAD, and on the fill factor (FF) of the whole array (see Chapter 2). In this application the amount of signal could be limited. For this reason, it is important that the level of noise of the detector is as low as possible to improve the signal to noise ratio and improve both the energy and timing resolution. The latter is even more critical when other very low luminosity effects, such as Cherenkov luminescence [82], are considered. So, a good detector should limit the contribution of noise, such as dark count rate and afterpulsing (see Chapter 2.1.5). The timing resolution of the system is strictly linked to the one of the single photodetector [75]. For this reason a minimization of the timing jitter of the selected detector is important to improve the reconstructed image quality.

1.3 Why SPADs in Standard Technologies?

SPADs present some substantial advantages over the previously shown devices. First of all, they can be integrated into silicon dies of very small size, allowing compactness and high granularity of detection. Moreover, they are insensitive to magnetic fields, in contrast to PMTs. In addition, it was shown how it is possible to integrate SPAD with circuits in standard technologies. This compatibility makes them suitable for the implementation of large-scale arrays, reducing the costs, and speeding up the prototyping phase. Furthermore, the use of standard technologies with circuitry directly integrated side-by-side to the SPADs allows implementing sophisticated front-ends and complex functionalities on chip. Also, their capability of operating at room temperature, and the lower bias required, compared to PMTs, paved the path to a broad spectrum of applications.

1.4 Organization of the Thesis

This thesis is divided into seven chapters.The second chapter provides a background on SPAD devices. The third chapter is dedicated to metrology to provide the reader with all the concepts needed to characterize the devices discussed in this work. The fourth chapter describes the design, implementation, and characterization of SPAD devices implemented in two different deep-submicron standard technologies (180 nm CMOS and 55 nm BCDL). Chapter 5 reports a detailed analysis of the work carried out to design and implement a 3D-stacked FSI multi-digital SiPM realized in 180 nm CMOS technology. The sixth chapter describes a case study carried out to optimize of the optical interface of a scintillator-based module targeting positron emission tomography applications. The seventh chapter presents another case study in which SPADs were used for direct detection and time-of-flight measurements of minimum ionizing particles. A chip gallery of the devices developed during this thesis closes the manuscript.

1.5 Thesis Contributions

This thesis attempts to advance the state-of-the-art in SPAD-based sensors, primarily aiming to optimize the performance needed in time-of-flight applications, such as time-of-flight PET. Multiple design challenges have been addressed, focusing on sensors fully integrated with standard technologies, such as CMOS and BCD. Moreover, the challenges of designing SPAD devices with state-of-the-art performance in advanced sub-65 technology were also tackled.

The following part of the section reports the list of the contributions made in this thesis.

The **first contribution** includes the implementation of a SPAD-based sensor in 180 nm CMOS and 55 nm BCD technologies with unprecedented performance in terms of timing jitter and dead time. In addition, this work shows the critical influence of the front-end electronics on the timing performance of the device, and the influence of the doping level in the photocollector area on the PDP.

The **second contribution** includes the demonstration of a large-scale 3D stacked multi-digital SiPM implemented using front-side illumination architecture. This device is based on two tiers implemented in 180 nm CMOS technology and relies on TSVs as 3D connections. A full analysis of the detector structure and its electronic components is presented in this thesis, together with its first preliminary characterization results.

The **third contribution** focuses on the problem of light extraction enhancement in inorganic scintillators. In particular, we optimize a photonic crystal structure, implemented on the output face of the scintillator, with the introduction of a silica buffer layer. This strategy increases its stability and robustness. Moreover, with thoroughly performed characterization, it was demonstrated how this optimized structure substantially improves light extraction performance in BGO crystals.

The **fourth contribution** of this thesis is the use of SPADs to detect minimum ionizing particles and perform time-of-flight coincidence measurements in a high energy physics experiment. The test was performed at the CERN North Area beam test facility. The outcome of this test led to unprecedented results in terms of timing resolution, showing about a 3× improvement with respect to the state-of-the-art. This exciting outcome paves the way to a new generation of low-cost plug-and-play tracker detectors with extremely high spatial and timing resolution, meant to be used in beam test facilities.

2 SPAD-Based Sensors Background

2.1 Single-Photon Avalanche Diodes

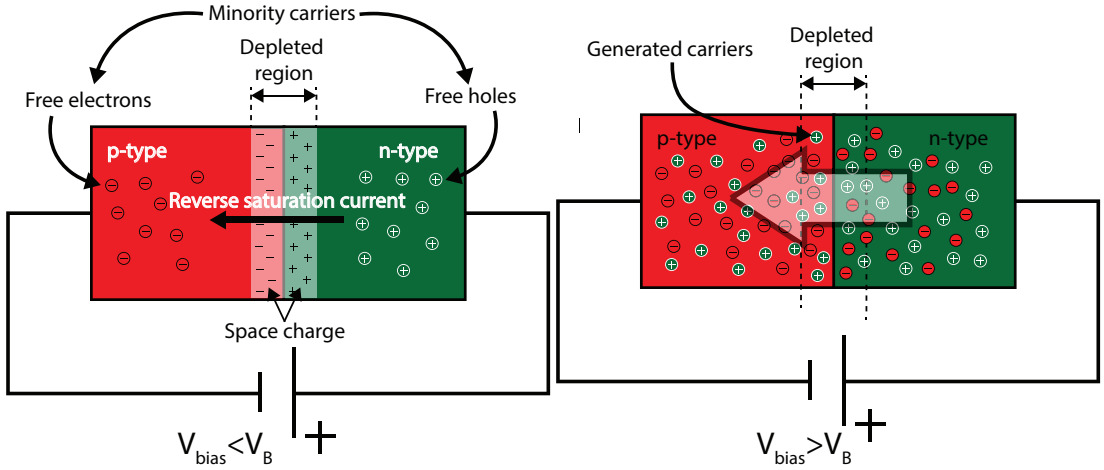


Figure 2.1: A *p-n* junction biased below the breakdown voltage (*left*). The image shows only the minority carriers. A so-called depleted region forms at the layers interface. In this situation only very little current flows, generated by the free minority carriers present in the crystal and driven by the electric field. When the bias is set above the breakdown, free carriers gain kinetic energy and are accelerated under the existing electric field. New carriers are generated by impact ionization. This mechanism causes an abrupt increase of charged carriers in the area of the depletion layer of the junction that vanishes. The absence of a barrier created by the depleted region allows a strong reverse current flowing through the device (*right*).

The detection of single photons, as mentioned in the previous chapter, has gained increasing traction among scientists. This interest is driven by emerging new applications (e.g. Raman spectroscopy, fluorescence spectroscopy, etc.) that require an extremely low light level to be detected. Initially, PMTs (see 1.1.1) were successfully used for this purpose, but their limitations in terms of size and supply voltage pushed researchers to find new alternatives that could accomplish the increasing need for performance. From the 1950's the impact ionization phenomenon in *p-n* junctions has been extensively studied, and several theories were elaborated to relate the ionization rate with the electric field at the junction [83]–[85]. The high gain of impact ionization based devices allows the detection of a smaller signal (i.e. smaller number of light photons) [86]. Only in the 1970s, it was proposed to perform single-photon detection with this devices [87], nowadays known as single-photon avalanche diodes (SPAD) or Geiger-mode avalanche photodiodes (G-APD).
This chapter reviews the basics of the avalanche process. Then, some standard SPAD structures are introduced and compared. Essential aspects, such as noise sources, light sensitivity, and timing performance are analyzed toward the end of the section.

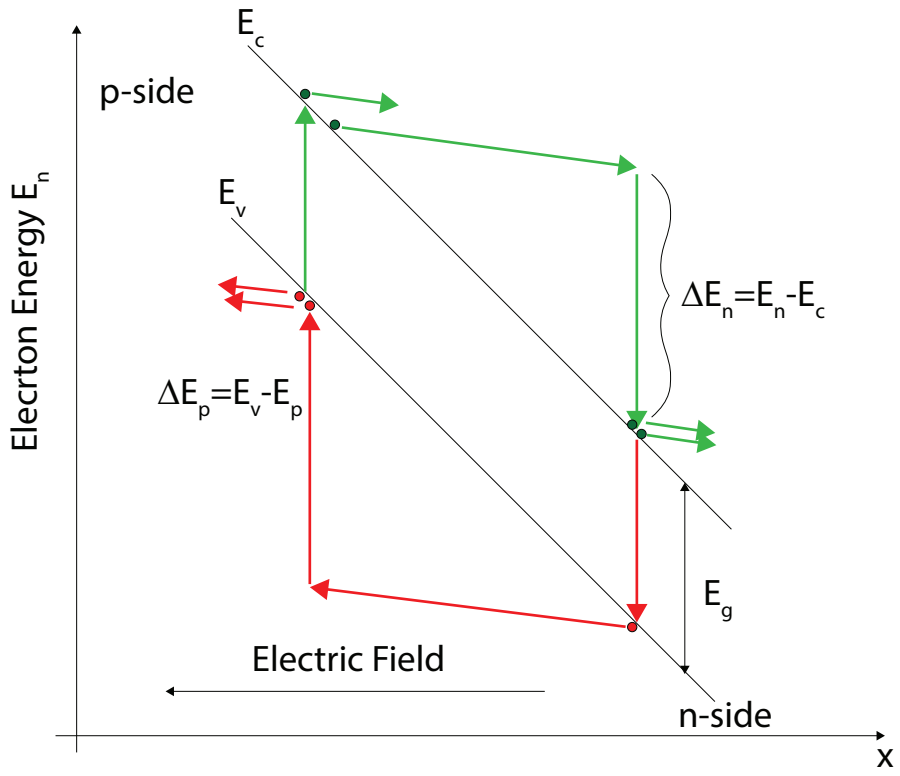


Figure 2.2: Impact ionization scheme for electron (*green dots*) and holes (*red dots*) in a *p-n* junction, where E_c and E_v indicate the conduction and valence bands, respectively, and E_g the energy gap. The strong electric field accelerates the free carriers, which accumulate energy (E_n and E_p). When the carriers accumulate an energy $\Delta E_n > 1.5E_g$ (for electrons) and $\Delta E_p > 1.5E_g$ (for the holes), they can cause impact ionization. This phenomenon frees new carriers that can be accelerated by the electric field.

2.1.1 Avalanche Detectors in Linear Mode

The Semiconductor breakdown is defined as the phenomenon that occurs when the electric field reaches the critical value in the material. This value is used to set the limit of the operating range of a device [86]. When the reverse bias applied across a *p-n* junction is increased, the electric field intensity rises, which leads to a significant increase of the current in the device (Fig. 2.1). Indeed, in a junction biased close to breakdown, a free electron drifting in the high field region is accelerated and gains kinetic energy. If the accumulated energy is not high enough, it is partially transferred to lattice vibrations by scattering events (i.e. non-ionizing collisions) that heat the crystal. Because of momentum conservation, an ionizing collision occurs only when the energy of the free carrier (ΔE_n and ΔE_p) is equal to or higher than $1.5E_g$ (Fig. 2.2) [88]. When an ionization event occurs, an electron-hole pair is freed into the crystal. Thanks to the strong electric field, these carriers are accelerated in opposite direction. The impact of these carriers with the lattice can generate more free carriers in the crystal. It is clear that this process is inherently a positive feedback loop created by this cascade of

ionization events, given by both electrons and holes accelerated by the electric field. This process is called avalanche multiplication. A photodiode biased below breakdown, but close to it, provides a linear amplification thanks to this mechanism. These devices are called avalanche photodiodes (APD). In APDs, if n carriers are generated by the photoelectric effect in the high field region, m free carriers will be generated through the avalanche process. Their ratio is defined as multiplication gain ($M = \frac{m}{n}$). Considering a specific incident optical power (P_{in}), the generated photocurrent ($I_{m,ph}$) can be expressed as:

$$I_{m,ph} = M I_{0,ph} = M \frac{q\lambda\eta}{hc} P_{in}, \tag{2.1}$$

where M is the gain and $I_{0,ph}$ is the photocurrent generated at low voltage (without avalanche multiplication) [86]. The multiplication process stops autonomously, and it provides a finite gain as the electric field is not strong enough to keep generating cascade events.

In silicon detectors, the ionization coefficients for electrons(α) and holes (β) are very different and their ratio defined as $k = \beta/\alpha$ is much lower than one. This difference in value implies that electrons have a significantly higher probability of ionization along their path with respect to holes. In addition, in APDs the avalanche reaches a finite value and it is affected by fluctuations caused by this ionization probability. In particular, the fluctuation of the electrons generated in an impact is not only amplified by further electron impacts in the subsequent multiplication path. The holes that are generated in the impact travel back and re-inject the fluctuation in a previous step of the multiplication path. This back-injection of fluctuations enhances the so called excess noise factor F, which depends on the k factor. The latter significantly increases as the electric field is increased. Therefore, F markedly increases as the bias voltage of the APD is raised for increasing the gain. This mechanism sets a main limitation on the maximum achievable gain, if we keep the fluctuation to an acceptable level. In general, typical APDs can achieve a gain of up to several hundred. Nevertheless, this gain is insufficient, and the APD generally is to be not suitable for single-photon detection.

2.1.2 Geiger-Mode Operation

When a *p-n* junction is biased above breakdown with a certain overvoltage (V_{OV}) or excess bias voltage (V_{ex}), no free carriers are present in the depletion region and therefore no current is flowing. This condition of equilibrium is unstable and just a single carrier, generated thermally or by photoelectric effect, can induce a cascade of impact ionization events. This results in a macroscopic and self-sustaining current flowing in the device. The aforementioned process takes the name of avalanche breakdown (Fig. 2.3 on the *left*) and if not quenched properly can damage the device due to the heat generated at the junction. The breakdown current is not divergent and keeps a finite value (Fig. 2.3 on the *right*) thanks to a feedback effect. This feedback is due to the mobile space charge generated during the process that reduces the electric field acting on the carriers, thus stabilizing the multiplication to a self-sustaining level. The avalanche process will stop only when the electric field at the junction is lowered

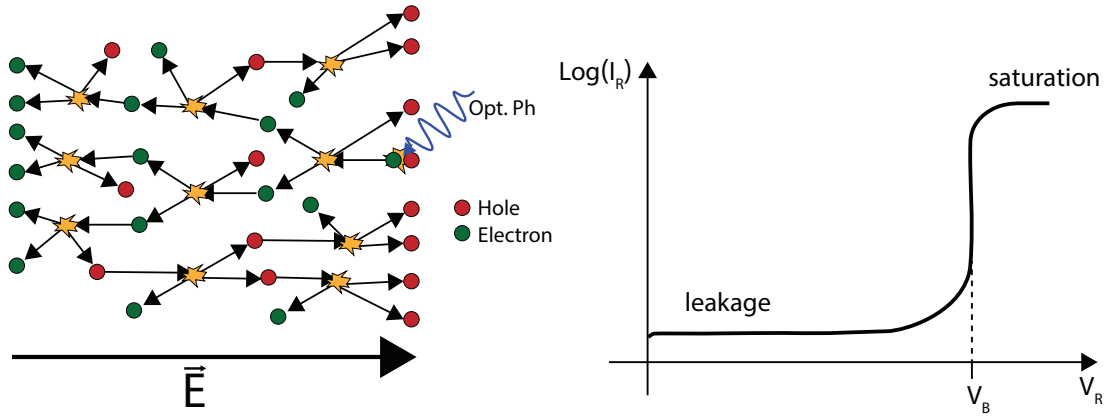


Figure 2.3: An electron-hole pairs is generated through the photoelectric effect by an impinging optical photon. The carriers are accelerated in opposite directions by the strong electric field. When the particles hit the lattice they generate new electron-hole pairs through impact ionization. This positive feedback mechanism is known as avalanche breakdown (left). The generated current stays at a finite level because of feedback created by the mobile space charge that reduces the electric field acting on the carriers. The current above breakdown saturates at a self-sustaining level (right).

below the critical value ($\approx 3 \cdot 10^5$ V/cm in Silicon). The field is effectively reduced by lowering the voltage across the junction below the breakdown. This mechanism is called quenching, and it can be implemented simply by placing a large enough resistor in series to the device. The system is rendered bistable when used with a quenching resistor because it will stay in an equilibrium state until an optical photon triggers an avalanche. Then, thanks to the quenching resistor, the SPAD is brought back to a quiescent state. This operating regime is called Geiger-mode (Fig. 2.4) because of its similarity to that of the Geiger-Muller counters of ionizing radiations. In this regime, a single photon can trigger a multiplication process (an avalanche) that causes a macroscopic current to flow in the device. This current, flowing through the quenching resistor, generates a voltage pulse that causes a voltage drop across the device, thereby quenching the avalanche. The voltage pulse generated by the avalanche and started by a single photoelectron has an amplitude equal to the excess bias voltage applied to the device. With sufficient excess bias, the generated pulse is easily detectable. Therefore, these devices are called single-photon avalanche diodes (SPADs) [89]–[91].
A more detailed description of the different kind of front-end circuits will be presented in the following sections.

2.1.3 SPAD Structures

As described in Section 1.1, the light absorption properties of silicon strongly depend on the wavelengths of the impinging photons. Indeed, different wavelengths are on average absorbed at different depths (Fig. 2.5). Therefore, several SPAD structures were designed to optimize the detection efficiency at a certain wavelength, in favor of specific target applications.

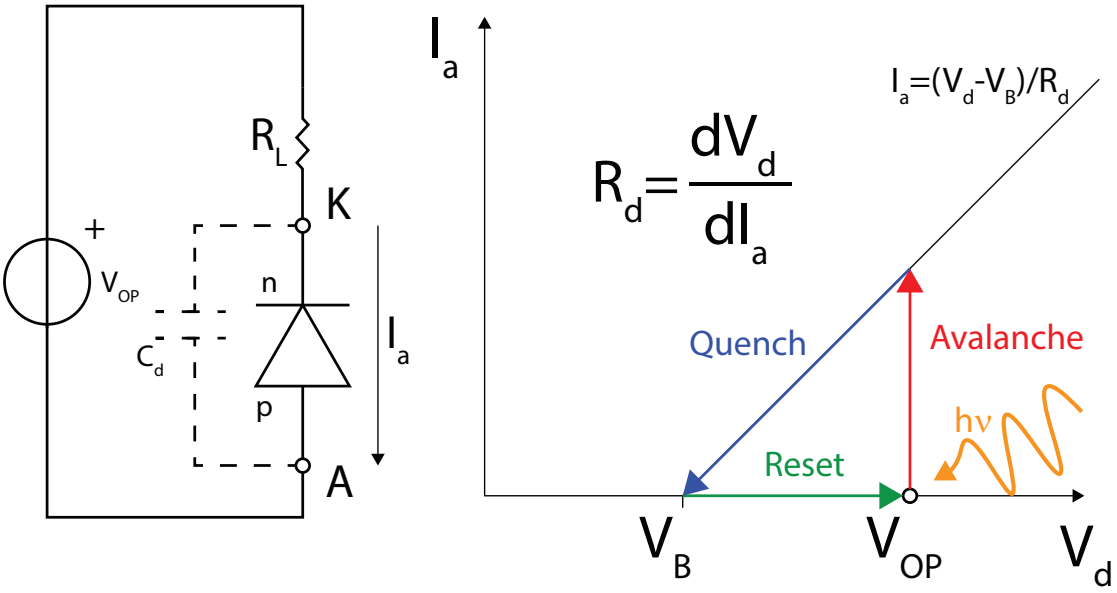


Figure 2.4: Geiger-mode operation scheme. A SPAD is connected in series with a quenching resistor used to quench the SPAD when an avalanche event occurs (left). In this configuration, when a photon triggers an avalanche, the avalanche current flowing into the load resistor will reduce the voltage across the SPAD. Once the avalanche is quenched, the resistor also allows restoring the voltage across the device, recharging the capacitance C_d, and making it ready to fire again (right).

SPAD in Custom Technologies

The first silicon SPADs were implemented using a dedicated custom process. An example of an early design is shown in Fig. 2.6 (*left*). This structure is called reach-through APD (RAPD) and it is based on a $p^+ - \pi - p - n^+$ cross section [92]–[94]. In this structure, the high field region is relatively narrow and takes place at the junction $p - n^+$. When the reverse bias is increased, the depletion region extends inside the p layer until it *reaches through* the π (nearly intrinsic) layer. Since the π layer is very wide, the electric field across it is very low. For this reason, the size of the π region is responsible for the high voltage needed to properly bias the RAPD device. Moreover, the π region helps in collecting the charges toward the high field junction (the multiplication region). The large size of the field region contributes to increasing the sensitivity of this device. However there will be a trade-off with timing performance. Alternatively, a planar structure is easier to implement and, most importantly, it is more similar to standard CMOS processes. Planar structures based on custom epitaxial technologies were introduced in the 1980s [95]–[99]. An example of this kind of SPAD can be found in Fig. 2.6 (*right*). In this structure, the junction is relatively shallow and placed between the n^+ and p^+

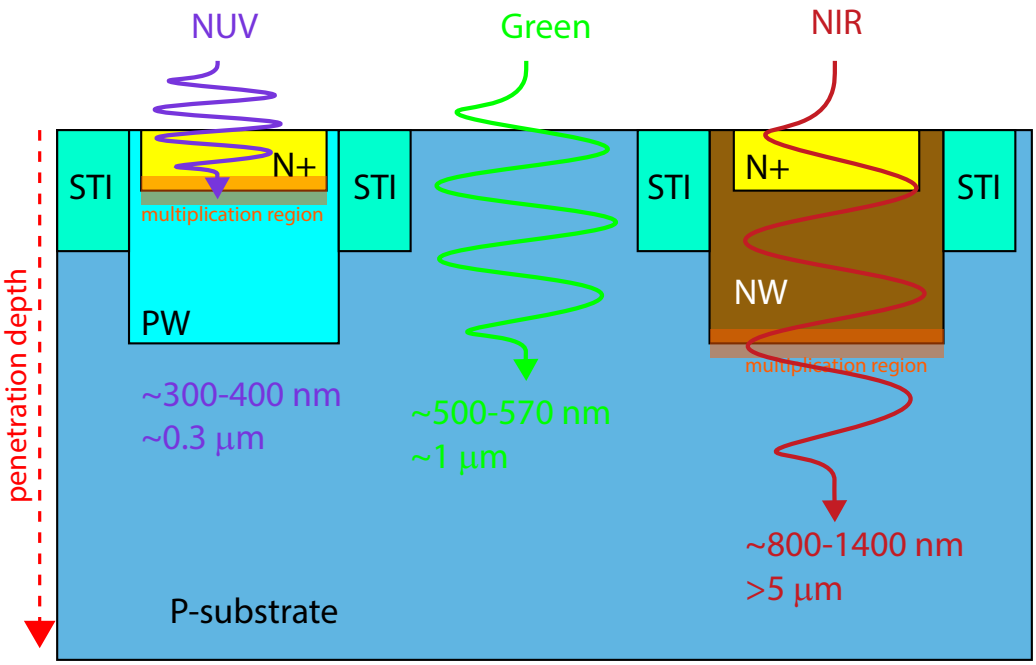


Figure 2.5: Light absorption in silicon-based SPADs with different junction depths.

layers. Since these two layers have relatively high doping, a guard region (or ring) is designed using an n^- layer to prevent premature edge breakdown. The use of a low doped n-type layer implant surrounding an $n+$ region was initially proposed in [100] and became one of the most common techniques for the implementation of SPAD structures to date. The two contacts for anode and cathode are both provided at the surface of the die. In this SPAD the sensitive area corresponds to the area of the high field junction and it is usually named *active area*. This is the only photo-sensitive region of the device. Contrary to the RAPD structures, these planar devices show much lower bias voltage requirements and at the same time a very good timing performance [99]. These characteristics, together with the effort to improve the sensitivity and noise performance, brought this device to evolve over the years [95]–[97], [99], [101]–[116]. As an example, in one of the most recent works published on the topic, the authors report a SPAD structure implemented in custom technology with enhanced NIR sensitivity, achieving a peak PDP above 70% at 650 nm (45% at 800 nm), a timing jitter of 95 ps FWHM, and a median DCR of about 2 $cps/\mu m^2$ (50 μm dia.) [116]. For a more comprehensive description and historical evolution of these devices, the author recommends the works [117] and [118].
Nowadays, many custom devices are available and they show very good performance. Nevertheless, the use of custom technology comes with some drawbacks, such as higher cost and the fact that it is not possible to integrate circuits side-by-side with the SPADs, on the same die. This limits the realization of large arrays that embed logic. Therefore, these technologies are mostly used for the implementation of single devices or of the so-called analog silicon photomultipliers (aSiPMs). This kind of device will be discussed in the following chapters.

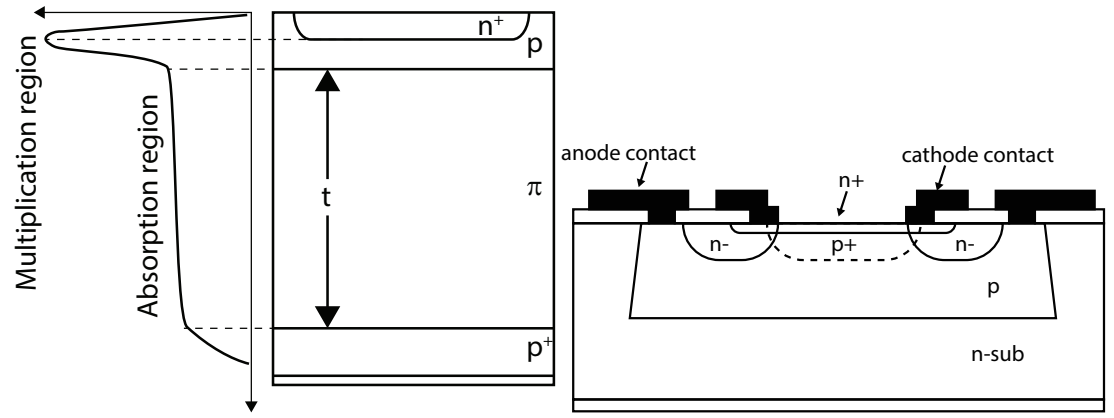


Figure 2.6: *Left*: Cross-section and electric field profile of reach-through APD. The thickness t of the intrinsic region can reach several tens of micrometers. The breakdown voltage of this device is high because of the large t, and for this reason, the bias voltage needed to operate it is to be very high. *Right*: Cross-section of a planar SPAD structure. In this case both anode and cathode contacts are placed on the surface of the wafer.

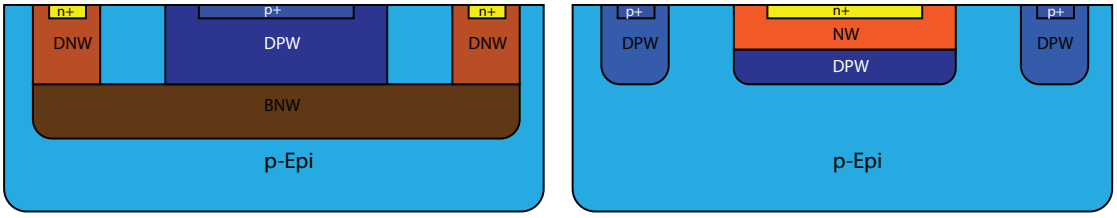


Figure 2.7: The substrate *isolated* SPADs are surrounded by a n-type deep layer (BNW) that electrically isolates the SPAD structure from the substrate (*left*). In the substrate *non-isolated* SPADs the anode of the SPAD is connected to the substrate (*right*).

SPAD in CMOS Technologies

Up until 2002, SPAD devices were implemented only in custom technologies, as described in the previous section. However, the interest in exploiting the full potential of industrialization pushed a quick development of planar SPAD structures into CMOS technologies. The use of a standard technology brings many advantages, such as cost reduction, availability of transistors and other devices to be integrated with SPADs, and the possibility to scale up SPAD-based architectures, thereby creating complex systems with sophisticated functionalities for the most diverse applications (see Section 2.3). Nevertheless, complications arise from the use of well-established commercial technologies to design and implement SPAD devices. These processes are often not very flexible, and only in sporadic cases do the foundries agree to introduce new mask layers dedicated to these devices in their process flow. Indeed, it is complex to add additional layers, while maintaining the same thermal budget as in the standard process flow. These complications prevent complete control of the process, thereby dramatically limiting the design freedom. However, the possibility of using deep submicron nodes to implement SPADs pushed many research groups to design their own devices. They were able to show

excellent performance, and eventually, these results convinced companies to introduce SPAD devices in their production lines [119].
The first SPAD device implemented in a standard high voltage 0.8 μm process was presented in 2003 [120]. While, the first array of fully integrated SPADs was published in [121]. In 2006, the first submicron SPAD was presented by [122] in a 0.35 μm HVCMOS technology. This inflection point resulted in further developments of SPADs in advanced technologies. Indeed, just a few years later, in 2007, the implementation of SPAD devices in deep-submicron technologies was demonstrated in [123]. After the demonstration of the feasibility of implementing SPAD devices also in relatively advanced technology nodes, the attention moved to the realization of large-scale arrays integrating timing [124]–[127] and processing [128]–[134] circuitry. For the interested reader, comprehensive review works on SPADs in CMOS technology can be found in [135]–[137].
During the following years, many works set new benchmarks on the SPAD performance achieved in CMOS technologies. Reference [138] tries to sort out the demonstrated SPADs using a figure of merit to rank devices based on their performance. In general, the SPADs developed in CMOS technologies can be split into two main categories: *isolated* and *non-isolated*. These two terms refer to the cross-section of the device and the use of specific wells to isolate the device electrically from the substrate.
These two general cross-sections are shown in Fig. 2.7. The substrate isolated SPAD was first introduced in [139] as a solution to limit the electrical cross-talk between devices. This technique presents many advantages with respect to its *non-isolated* counterpart. Indeed, when it comes to the integration of electronic circuits next to the SPAD, the use of a substrate isolation technique helps in protecting the circuit from the high voltage applied to the cathode terminal and allowing a substantial distance reduction between the SPAD and MOSFET of the front-end. Most of the commercially available technologies use a $p-type$ substrate. So, all the isolated SPAD structures shown in commercial technologies present an anode contact in the center made with a $p+$ implantation and a cathode contact implemented with an $n+$ implanted on top of an *n-well* region that contacts a deeper layer. The presence of an n region that completely embraces the SPAD forms a reverse-biased junction with the substrate. This *parasitic diode* guarantees electrical isolation from the substrate. Since the substrate of commercial technologies usually has a resistivity of few tens to few hundreds of $\Omega \cdot cm$ (corresponding to a doping level below or equal to $10^{16}\ cm^{-3}$), the breakdown voltage (BV) of this junction is relatively far above the one of the main SPAD junction, guaranteeing the correct operation of the device. Moreover, the enclosure of the SPAD in an isolated well makes the sensitive area and the high field region well defined. In agreement with what was discussed in the previous section regarding the optical properties of silicon, these devices do usually have a high sensitivity in the visible spectrum with a peak that varies from 300 nm to 600 nm, depending on the position of the main junction. However, their sensitivity at longer wavelengths is limited.
On the other hand, the substrate *non-isolated* SPADs are structures that implement the cathode at the center of the device with an $n+$ implantation and an anode contact with a $p+$ on top of a $P-well$ that surrounds the center cathode area. As it is easy to understand, these

kinds of SPADs do not have anything that prevents carriers from moving from/to the substrate region. For this reason, such devices are defined as substrate *non-isolated*. The main reason for using these devices relies on their enhanced sensitivity in the near-infrared (NIR) [140]. Indeed, in these devices, the carriers freed by photons that interact deep in the silicon can still be collected and trigger an avalanche. Although a high NIR sensitivity is very appealing in many applications, these devices are not easy to be integrated together with electronic circuits. Often, a considerable distance between the SPAD and the MOSFET is required. Moreover, because of their large absorption region, these SPADs show worse timing performance with respect to the *isolated* ones. The proposal of developing 3D implementations (as discussed in the following sections) was the natural step to overcome the issues.

The SPAD guard ring and Periphery

As mentioned when describing the SPAD operating principle, the guard ring is the region dedicated to premature edge breakdown prevention at the border of the main junction. On the other hand, the periphery of the SPAD is designed to provide a proper contact to the main junction on its deeper side in the substrate. Indeed, in SPAD structures as the one depicted in Fig. 2.7, the periphery of the SPAD is implemented using a DNW ($left$) and a DPW ($right$), respectively.
In general, these two regions are insensitive to light and, although required in the design, limit the overall active area of the device. Therefore, the most intuitive solution could be to reduce their size as much as possible. However, an improper design of these two areas can have catastrophic consequences on the SPAD performance or even prevents its Geiger-mode operation [136].
The classical guard ring can be considered as a region of the same dopant type as the center contact of the SPAD, used to smooth the transition to the region with opposite doping type. However, it is crucial to consider that the guard ring also forms a secondary junction and that this junction could also enter the Geiger-mode regime. Thus, careful designers should accurately choose the doping level of the layer used to implement this region, so that this secondary breakdown is reasonably far from the one of the main junction.
Fig. 2.8 (a) shows the standard guard-ring region of an isolated SPAD with a shallow $p+/DNW$ junction, implemented with PW diffusion at the edge of the $p+$ region [141]. In this case, the breakdown difference between the secondary junction (formed between the PW guard ring and the surrounding region) and the main junction will set the maximum excess bias voltage value at which the structure can be operated. Another critical point to be considered is the size of the guard ring. Indeed, as demonstrated in [136], we can observe an increment of DCR when reducing the size of the guard ring. To understand better what is happening, we need to analyze the electric field distribution for several excess bias voltages. As also shown in [136], the edge of the shallow $p+$ layer is protected inside the guard ring region as its size remains sufficiently large. Therefore, if we excessively reduce the guard ring size, the electric field increases inside that region to fully deplete it, causing an increment of noise given by tunneling contributions.

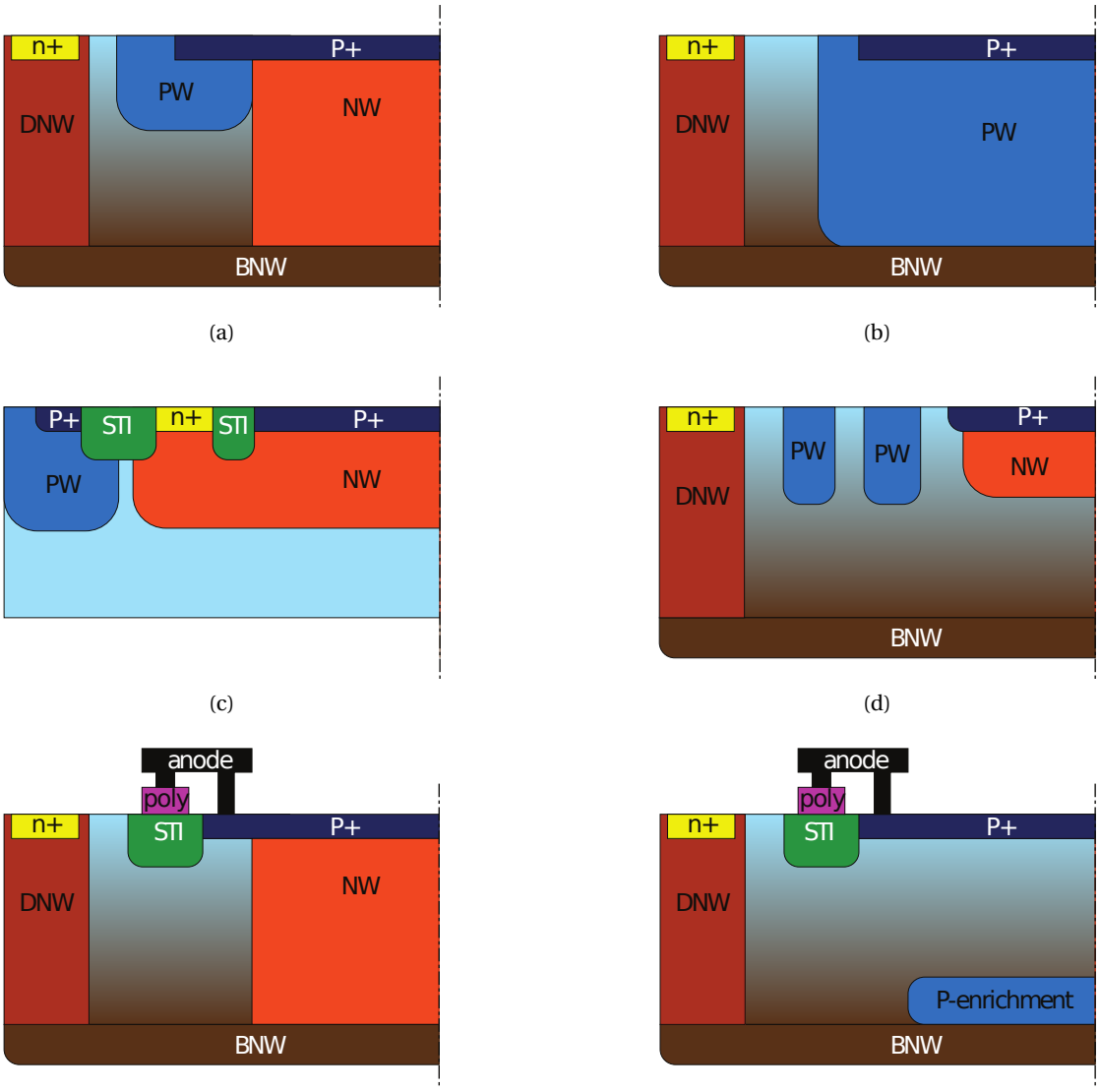


Figure 2.8: SPAD guard ring structures. *(a)*: diffusion guard ring [141] . *(b)*: virtual guard ring [142]. *(c)*: STI-based guard ring [143]. *(d)*: floating guard ring. *(e)*: polysilicon field plate-based guard ring [144].

A proper design of the guard ring region can also allow the operation of the SPAD when both main and secondary junction operate in Geiger-mode simultaneously [136]. Another strategy is the adoption of the so-called virtual (or implicit) guard ring (see Fig. 2.8 (*b*)) [95], [102], [145]. This structure was shown for the first time in 2011 by [142] in CMOS technology. This strategy relies on the retrograde doping of the buried n-well (*BNW*). This implant is available in all modern triple-well CMOS processes in order to isolate n-channel metal–oxide–semiconductor (MOS) transistors from substrate noise [142]. The effectiveness of this solution relies on two aspects. The first one concerns using STI to delimit the shallow implantation for the external

contact and to set the starting point of the guard ring to avoid interface traps [123]. The other aspect concerns the size of the intrinsic layer used to separate the center region from the peripheral contact. This parameter needs to be carefully tuned in order not to degrade too much the fill factor of the device as well as not to increase the DCR (similarly to the previous solution) or not being able to prevent premature edge breakdown.
Moreover, the size of this spacing will also be linked to the maximum operating voltage used on the SPAD. The higher the voltage, the larger the spacing. In standard CMOS technologies, a good size for this spacing has been demonstrated to be around 1 μm for excess bias voltages up to 6-7 V [136].
During the years, other kinds of strategies to implement guard ring structures were introduced. The first solution consists of applying a metal field plate (FP) in electrical contact with the diffused region on the oxide over the junction, surrounded by a metal ring (EQR) in contact with the substrate. This solution lowers the junction curvature field and improves the stability of the surface potential. As a result, the BV is higher and more stable [146]. A similar solution is proposed by [147] and more recently in [144]. In the first work, a polysilicon plate, placed on the edge center contact of the SPAD, is connected to an external voltage used to modulate the field beneath the polysilicon and mitigate the risk of edge breakdown. In the second work, the authors propose to place a polysilicon field plate (poly FP) on top of the edge of the $p+$ region and connect it to the anode voltage (Fig. 2.8 (*e*)). The field induced by this polysilicon plate in the silicon beneath smooths the electric field and prevents edge breakdown. In [143], the guard ring is implemented by bounding the $p+/n$ junction with a shallow trench isolation (STI) (Fig. 2.8 (*c*)). With this strategy, SPADs with 2 μm active diameter were demonstrated. Although this strategy allows increasing the fill factor, the reported DCR is much higher than in other designs. This noise is mostly attributed to the tunneling and field-enhancement generation effect given by the low breakdown of the structure's main junction ($\sim$ 11 V) [111]. Moreover, an additional noise source is given by possible surface generation from the STI interface ($Si-SiO_2$) [142]. The floating guard ring proposed in [148] consists of low doped implantation placed at a certain distance from the high-doping region (e.g., the p+) (Fig. 2.8 (*d*)). This diffusion decreases the electric field at the edge of the high doping implantation region, preventing the edge breakdown. In Fig. 2.8, a schematic view of the mentioned solutions is shown.
The peripheral area includes those implantations needed to create deep peripheral contact in the substrate and ensure a proper bias of the main junction. For example, for the isolated structures (Fig. 2.7), this area provides the cathode contact of the device. In the design of this region, it is essential to consider the depth of the layer used. Indeed, a reliable electrical contact needs to be created with the BNW to properly bias the cathode of the main junction. Another critical point is the doping level of the layer used. A too-low doped layer can create a resistive path through the implantation region, degrading the overall device performance. On the other hand, an excessively high doping level can considerably decrease the breakdown of the parasitic junction between this implantation and the internal region of the SPAD, potentially compromising the correct operation of the device, and raising the risk of edge breakdown.

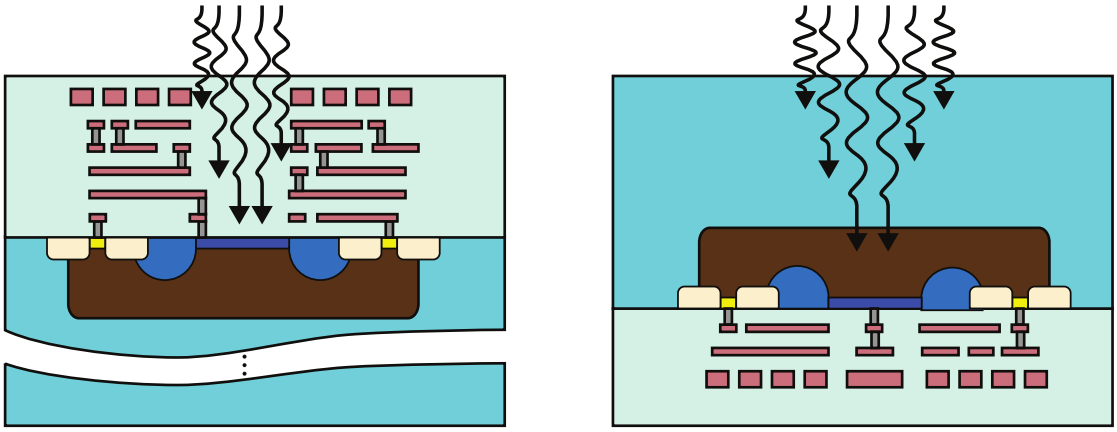


Figure 2.9: In front-side illumination the light reaches the photodetector from the top of the wafer, passing through the optical stack composed of oxide layers (*Left*). In back-side illumination the wafer is thinned down to a thickness of a few micrometers and illuminated from the back (*Right*). In this configuration, the optical photons need to travel through a thin layer of silicon.

More exotic SPAD structures have been proposed over the years trying to overcome some of the problems present in the standard structures. For the interested reader, the author suggests the following works: [149]–[155]

2.1.4 Front-Side and Back-Side Illuminated Sensors

As for the CMOS image sensors, also in the case of SPADs, the first and most widespread approach of operating the devices is the so-called front-side illumination (FSI). It consists in illuminating the device from the top of the die. The photons pass through the passivation layers and the full optical stack of the device until they reach the silicon (Fig. 2.9). This approach is straightforward for typical 2D architectures and it is the most frequently used one. Moreover, FSI guarantees high sensitivity of the devices, especially at short wavelengths, covered only by oxide layers, most transparent in the range of wavelength of interest (typically between 300 nm and 1000 nm). The main drawback of this approach is a limited fill factor (FF) of the structure that also embeds circuitry. A way to mitigate this limitation is described in Chapter 4. The second approach relies on the so-called back-side illumination (BSI). In this approach, the devices are illuminated from the back of the sensor. In this scenario, the photons interact directly in the silicon, freeing carriers that then diffuse up to the high field region of the SPAD and can trigger an avalanche. In standard fabrication processes, the thickness of the wafer is between 500 and 800 μm. With such thickness, only very few of the photons impinging on the silicon surface would ever interact close to the SPAD structure (usually no more than 10 μm from the surface of the wafer) and be able to trigger an avalanche. For this reason, wafer thinning is needed. In the structures demonstrated so far, the wafer is thinned down to a thickness that can go below the 10 μm allowing many photon interactions in the

proximity of the SPAD structure [156]–[161].

2.1.5 Sources of Noise

We can identify three primary noise sources in single-photon avalanche diodes: thermal generation, band-to-band tunneling, and afterpulsing. They need to be measured carefully to correctly characterize these kinds of devices and their performance [162]. The first two contributions are commonly merged and characterized as DCR (Fig. 2.10), while the third is usually considered individually and measured in terms of afterpulsing probability. The contributions mentioned above are also split into two categories based on their characteristics: uncorrelated (thermal generation and band-to-band tunneling) and correlated (afterpulsing) noise sources. This section provides an analysis of these sources of noise.

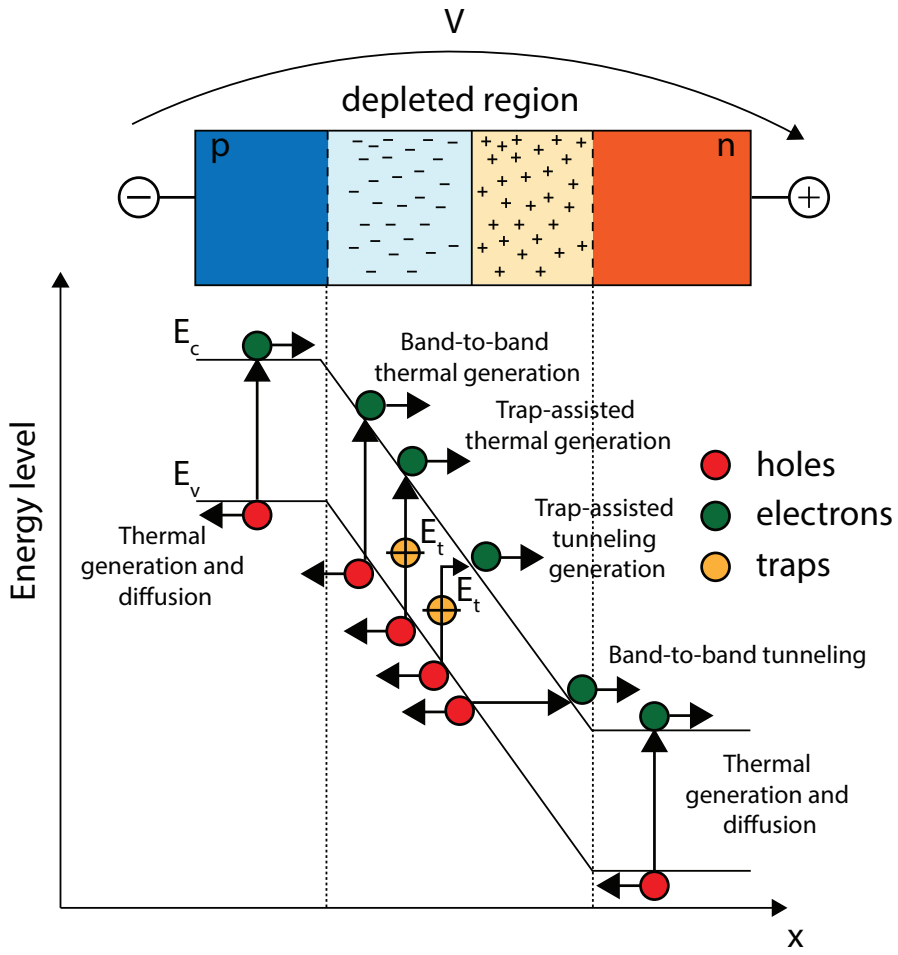


Figure 2.10: DCR contributions illustration for a reverse biased p-n junction. Neglecting the afterpulsing contribution, the DCR is composed of thermal and band-to-band tunneling contributions. In the figure E_t indicates the trap energy level.

Thermal Generation

The first source of noise in a SPAD discussed here is thermal generation from generation centers. This effect is composed of random avalanche triggers, which can be attributed to the natural thermal generation of carriers. Together with the band-to-band tunneling discussed in the next section, it contributes to the so-called *primary DCR* [136], [163]. This noise strongly depends on temperature, and it can vary by a factor of 2-4 over 10°C of temperature variation around room temperature. This variation is justified by the fact that this noise contribution is trap-assisted. The Schockley-Read-Hall (SRH) theory models the carrier capture and release with traps allowing states in the silicon energy gap (E_g) [26], [147], [164]. This theory explains how the presence of traps dramatically increases the probability of thermal carrier generation [147]. Consequently, the reader should note how the implantations and annealing processes performed by the foundries in the production phase will affect the noise performance of the devices [165]. If we assume the same release characteristics for both electron and holes, the carrier generation rate ($G_{tr}\left(\left|\vec{E}\right|\right)$) can be expressed as [147]:

$$G_{tr}\left(\left|\vec{E}\right|\right) = \frac{n_{tr}\left(1+\Gamma\left(\left|\vec{E}\right|\right)\right)}{2\tau cosh\left(\frac{E_t-E_i}{kT}\right)}, \tag{2.2}$$

where n_{tr} is the trap concentration, Γ is a correction factor based on the field strength, E_t is the trap energy level, E_i is the intrinsic Fermi level, τ is the mean capture lifetime [147]. If the release characteristic for electrons and holes are not the same, the model can be extended as shown in [111]:

$$G_{tr}\left(\left|\vec{E}\right|\right) = n_{tr}\frac{\left(\Gamma_e\left(\left|\vec{E}\right|\right)\right)r_e\left(\Gamma_h\left(\left|\vec{E}\right|\right)\right)r_h}{\left(\Gamma_e\left(\left|\vec{E}\right|\right)\right)r_e+\left(\Gamma_h\left(\left|\vec{E}\right|\right)\right)r_h}, \tag{2.3}$$

where r_e and r_h are the probabilities per unit time of emitting an electron or a hole, respectively, and $\left(\Gamma_e\left(\left|\vec{E}\right|\right)\right)$ and $\left(\Gamma_h\left(\left|\vec{E}\right|\right)\right)$ are enhancement factors that depend on the electric field, determined by the detector design and the biasing condition. From the previous expressions, it follows that this DCR contribution depends on the detector volume and increases with it. Optimizations of the process aimed to reduce the value of n_{tr} can drastically decrease G_{tr}, and as a consequence, the noise of the device. One of the most common causes of high noise is metal contamination during wafer handling, ion implantation, and high-temperature heat treatment [111].

Band-to-Band Tunneling

Zener theory extends to semiconductors and states that the average rate of carrier generations per cm^3 (ν_f) in a *p-n* junction is equal to the total number of oscillations carried out by

electrons in the valence band multiplied by a field-dependent probability factor [162]:

$$\nu_f = 4Naq\frac{E}{h}e^{-\frac{\left(\frac{\pi^2}{2h}\right)(2m^*)\left(\frac{1}{2}\right)E_g^{\frac{3}{2}}}{qE}}, \tag{2.4}$$

where N is the number of silicon atoms per cm^3, a the lattice constant of silicon (5.431 Å at 300K), q the electron charge, h the Planck's constant, m^* the effective electron mass, E_g the energy gap in silicon, and E the electric field. Thus, when the applied excess bias exceeds 1 V, it is correct to assume that every carrier generated in the high-field region triggers an avalanche. In this case, the internal field emission is expressed as:

$$N_f = \upsilon\nu_f, \tag{2.5}$$

where υ is the volume of the high field region. In [162] it is also demonstrated how this noise contribution is not strongly affected by temperature. The temperature dependence of band-to-band tunneling can be explained by taking into account the variation of E_g with temperature [166]. This dependence can be expressed by means of a temperature coefficient (γ):

$$\gamma = \frac{1}{\nu_f}\frac{d\nu_f}{dT} = \frac{1}{N}\frac{dN}{dT}. \tag{2.6}$$

The temperature coefficient is on the order of tens of mK. Thus it is possible to conclude that this effect is almost temperature independent with respect to thermal generation.
Furthermore, when we consider the dependency of this noise source on the electric field, band-to-band tunneling is usually the dominant contribution in SPAD structures where the junction presents high doping, resulting in a low breakdown voltage [147], [167], [168].

Afterpulsing

This phenomenon, already shown in [162], manifests itself as a train of pulses that follow a primary pulse generated by a noise carrier or a photoelectron. Deep level traps present in the crystal that allow energy levels close to the energy bands can capture carriers during an avalanche and hold them for a specific time [169], [170]. This time is called *trap's release lifetime* or simply *lifetime*, and its value can vary from tens of nanoseconds to several microseconds. The lifetime τ decreases when increasing the temperature following the expression [169], [171]–[173]:

$$\tau = \tau_0 e^{\frac{-E_A}{kT}}, \tag{2.7}$$

where E_A is the activation energy, defined as the energy difference between the trap's energy level and the bottom of the conduction band for an electron trap (or the top of the valence band for a hole trap), and τ_0 is the pre-exponential factor [174], [175]. After this time, the traps release the carrier, and if the electric field is sufficient, it can trigger an avalanche [172], [176]. Indeed, the release can happen either when the SPAD is fully reverse-biased or when the SPAD is quenched. In the latter case, the voltage across the diode is close to the breakdown voltage

one. In the first case, the released carrier can start another ionization cascade. Due to its correlated nature, the afterpulsing is detrimental for several applications and thus it must be minimized. Several research groups modeled the afterpulsing following a statistical approach [174], [177]–[181]. In general, there is more than one deep-level trap in a SPAD, and they can be modeled using different lifetimes. A good model turned out to be represented by means of the sum of four exponentials, looking at the fit of data obtained through the time-correlated carrier counting technique (TCCC) [111], [172], [173], [176], [181], [182]. In [183], the use of a power law was also shown to provide accurate results.
The afterpulsing probability is time-dependent. Indeed, depending on when the trap releases a carrier, it will have or not have the possibility to trigger an avalanche event. The total afterpulsing probability can be expressed as[181], [183]–[185]:

$$P_{ap}(t) = \int_{t}^{+\infty} n_{ap}(\tau) d\tau, \tag{2.8}$$

where n_{ap} is the afterpulsing probability density, and t the dead time of the device. This quantity can be expressed as the sum of several exponential decays, given that a SPAD usually presents different types of deep-level traps. If we consider only electrons, the afterpulsing probability density can be expressed as:

$$n_{ap,e}(\tau) = \sum_{i=1}^{N} A_{e,i} e^{-e_{e,i}\tau}, \tag{2.9}$$

where N is the number traps, $A_{e,i}$ is a suitable prefactor, and $e_{e,i}$ is the emission rate associated with the *i-th* trap. Following the same considerations, we can write a similar equation for holes:

$$n_{ap,h}(\tau) = \sum_{i=1}^{N} A_{h,i} e^{-e_{h,i}\tau}. \tag{2.10}$$

Several factors influence the afterpulsing probability in a SPAD. From the technological point of view, a very high wafer quality may reduce the number of deep level traps and, consequently, the number of trapped carriers. Moreover, during the fabrication process the flow needs to be designed to have a minimal number of steps and careful heat treatment to limit, as much as possible, the number of dislocations caused by implantations. Another crucial point that influences afterpulsing probability is the operating condition of the device. In particular, the afterpulsing is strictly related to the current amount that flows in the device during the avalanche process. Indeed, a significant current value implies many carriers flowing in the device's active region, increasing the probability that silicon traps can capture these carriers. It is then clear how it is very important to limit this current as much as possible in order to reduce afterpulsing. Without any specific integrated circuit (discussed in the following sections), the reader should note that a trade-off exists on the detector side. Indeed, one way to limit the amount of charge in the detector consists in decreasing the excess bias voltage. However, a lower excess bias will cause a reduction of sensitivity.
Moreover, a large parasitic capacitance connected to the terminals of the SPAD requires a large amount of charge to change the voltage. For this reason, a proper front-end design

integrated on-chip is recommended to improve the afterpulsing performance. Furthermore, as mentioned before, the lifetime is influenced by temperature, and in particular, a reduction of temperature can have an adverse effect on afterpulsing. Indeed, the lifetime increases with decreasing the temperature, and below 77K, the afterpulsing is observed in groups where the first pulse is triggered by a carrier generated by noise or optical photon and all the others are triggered by trap-released carriers [162].

2.1.6 Light Sensitivity

The light sensitivity of SPAD is characterized by its *photon detection probability* (PDP). This quantity can be defined as the probability that a photon impinging on the device's active area triggers an avalanche. As explained in the previous sections, the light penetration in silicon and the impact ionization are statistical processes. The PDP is normally used to evaluate and compare the performance of single devices and it does not take into account the fill factor [136], [147], [186]. The probability density function (PDF) of light absorption in silicon can be modeled as an exponential function of the depth:

$$f_\lambda(z) = \begin{cases} 0 & if\, z = 0 \\ \mu e^{-\mu z} & if\, z \geq 0 \end{cases}, \tag{2.11}$$

where μ is the mean penetration depth of light in silicon. Since silicon is an indirect bandgap material, light penetration changes rapidly with the wavelength (see Fig. 1.1). The PDP is computed as the integral of the number of electron-hole pairs generated at a certain depth z multiplied by the collection probability $p_{coll}(z)$ and the avalanche triggering probability $p(z)$:

$$PDP(\lambda) = \int_{-\infty}^{+\infty} T(\lambda)\, f_\lambda(z)\, p_{coll}(z)\, p(z)\, dz, \tag{2.12}$$

where $T(\lambda)$ is the transmission coefficient function of the wavelength at the optical interface and is defined as $(1 - R(\lambda))$, where $R(\lambda)$ is the reflection coefficient. It is important to mention that an optimization of the optical stack present on top of the device is mandatory in order to reach the maximum detection efficiency. The optical interface is usually composed of several layers of oxide and a final passivation layer. The latter can be optimized and used as an antireflection coating (ARC) to avoid optical photons being reflected back because of total reflection or Fresnel losses. The presence of layers of oxide on top of the device, if not optically optimized, often causes an interference pattern more evident when the light used is collimated. This pattern can be included in the PDP model through the transmittance and influences the final value of this parameter.

The collection probability can be explained by looking at Fig. 2.11 and Fig. 2.12 *(a)*. In this simplified 1D model, three regions can be identified: two neutral regions and the depletion region. In the depletion region, the high electric field makes sure that all the carriers in that volume are collected. Instead, the carriers generated in the neutral region need to diffuse to reach the high field region, and for this reason, the probability that they are collected will be

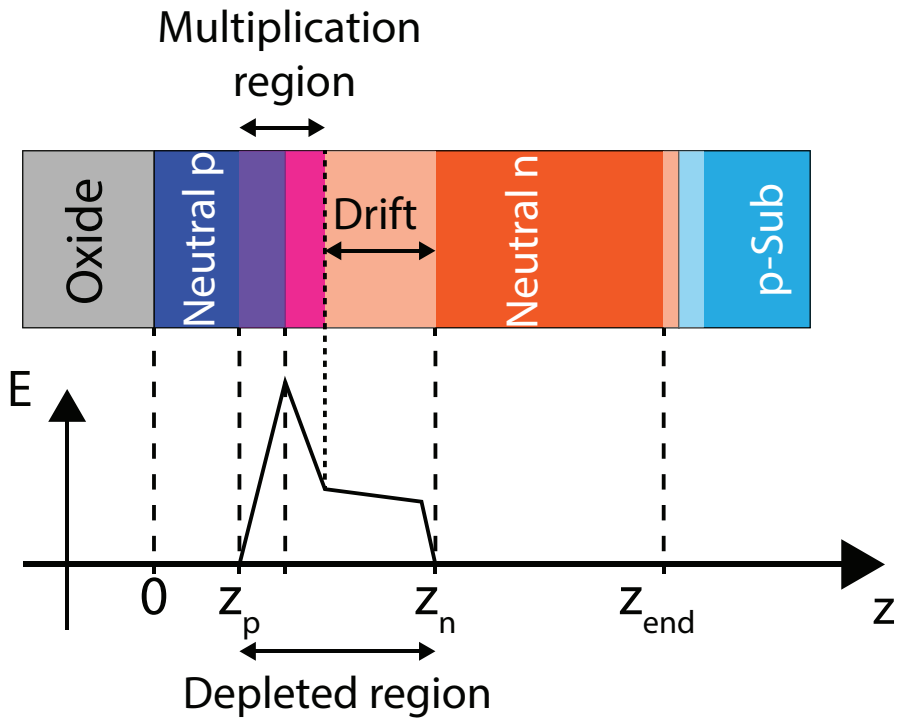


Figure 2.11: 1D SPAD cross-section model. In a 1D SPAD cross-section we can highlight a neutral p region, a multiplication region, a drift region, and a neutral n region.

reduced (Fig. 2.11). Therefore, we can express this probability as:

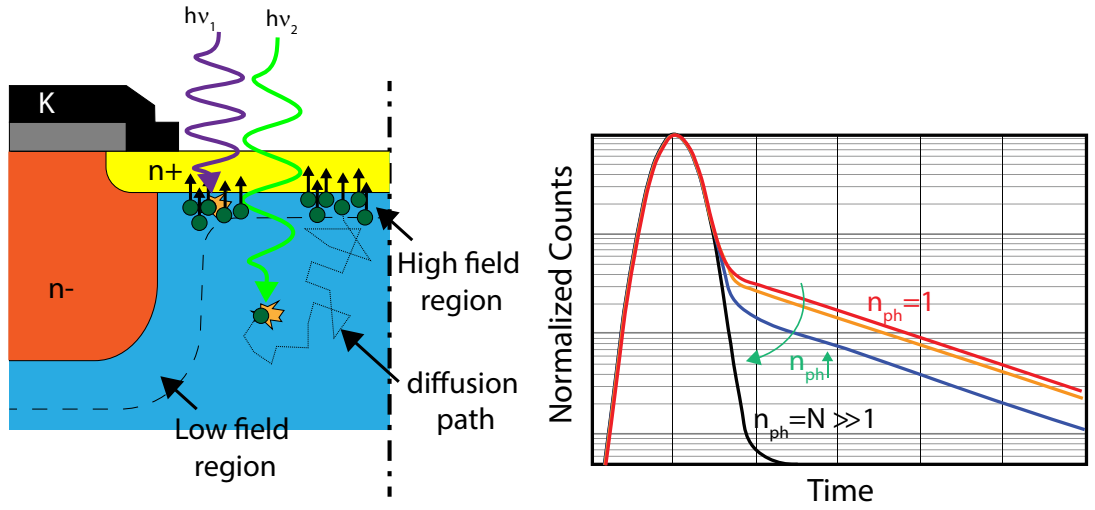
$$p_{coll}(z) = \begin{cases} p_{coll,e} & if\, 0 < z < z_p \\ 1 & if\, z_p < z < z_n \\ p_{coll,h} & if\, z_n < z < z_{end} \end{cases}, \tag{2.13}$$

where z_n is the limit of the drift region on the n side, z_p is the limit of the depletion region on the p side, and z_{end} is the end of the cross-section (Fig. 2.11). The term $p(z)$ accounts for the fact that not every carrier reaching the high field region triggers an avalanche [187]. From the work reported in [187], [188], this probability can be expressed as:

$$p(z) = p_{t,e}(z) + p_{t,h}(z) - p_{t,e}(z)\,p_{t,h}(z), \tag{2.14}$$

where $p_{t,e}(z)$ and $p_{t,h}(z)$ are the triggering probability for electrons and holes, respectively. This model considers the two events as independent and considers the probability that just one of them triggers the avalanche. This probability depends on the ionization coefficient that is strongly dependent on the electric field and the history of the particle [189]–[191]. Combining the equations 2.11 and 2.12 we obtain:

$$PDP(\lambda) = \int_0^{z_{end}} T(\lambda)\,\mu(\lambda)\,e^{(-\mu(\lambda)z)}\,p_{coll}(z)\,p(z)\,dz. \tag{2.15}$$



(a) (b)

Figure 2.12: *(a)*: Schematic view of a SPAD cross-section. The point where a photocarrier is generated through the photoelectric effect will determine if it can start immediately an avalanche or if it needs to diffuse until the high field region where it gains enough energy to trigger an avalanche. *(b)*: general profile of the IRF of a SPAD. The profile is composed of two main parts: peak and tail. The peak is caused by the statistical nature of the avalanche process and can be accurately fit with a single Gaussian profile. The tail follows an exponential decay and it is determined by all those photocarriers generated outside the high field region and that need to diffuse before reaching it. The exponential tail is also influenced by the number of photons (n_{ph}) that reach the active area of the detector at the same time (or reasonably close in time). In this case the photocarriers that interact outside the high-field region are *masked* by those that interact directly inside it. This effect distorts the IRF curve and if the number of photons (N) is relatively high, the tail is suppressed.
Images readapted from [111], [147]

In equation 2.15, the first term provides the number of photocarriers that reach or are generated in the high field volume of the device in proximity of the junction while the second corresponds to the probability that they have to start an avalanche. The first part can be identified as the quantum efficiency of the detector (QE). Thus, we can rewrite the previous expression as:

$$PDP(\lambda) = QE(\lambda)P_t, \tag{2.16}$$

where P_t is the overall avalanche triggering probability. Several factors influence the PDP of a SPAD. Indeed, P_t is related to the applied voltage, and it increases with it. Nevertheless, due to the carrier ionization coefficient saturation, the avalanche triggering probability tends to saturate when the field strength passes $5 \times 10^5\ V/cm$.
Another parameter influencing the sensitivity of a SPAD is linked to the doping profile of the SPAD cross-section. A more detailed discussion is presented in Section 3.2.

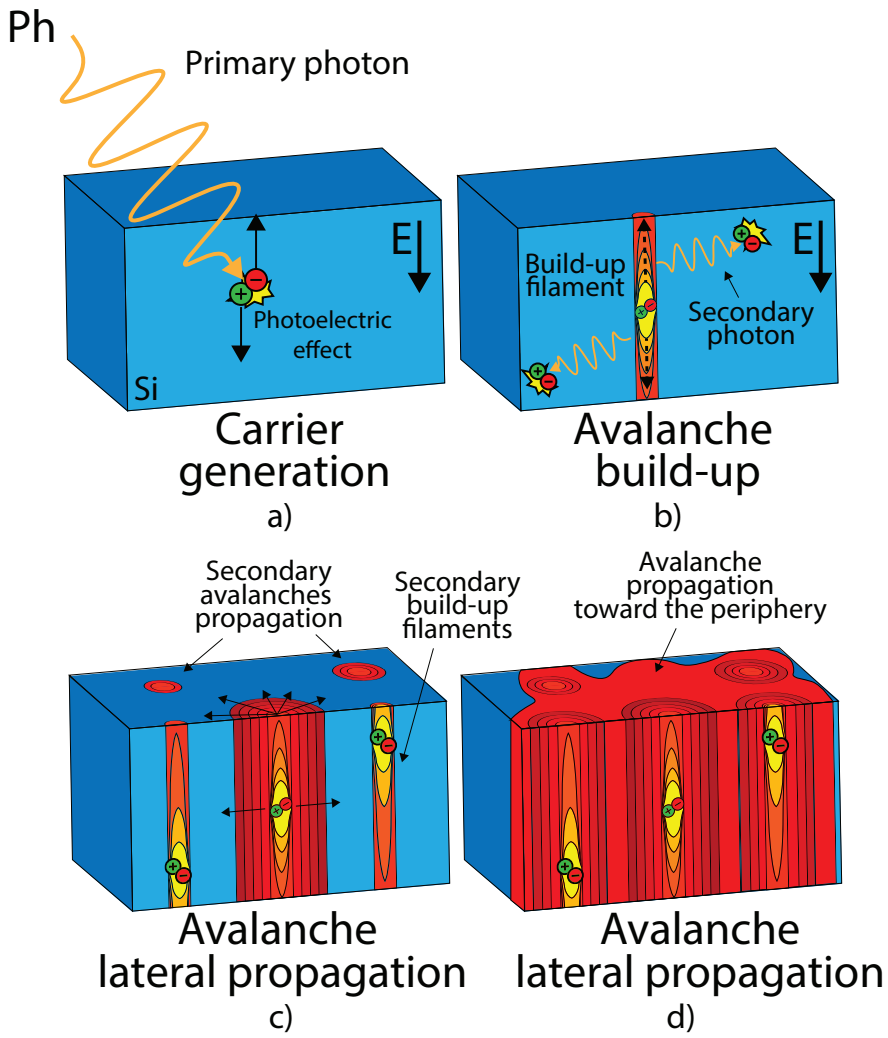


Figure 2.13: Simplified scheme of the avalanche phases. An optical photon generates a electron-hole pair interacting in a silicon volume, and the carriers are accelerated by a strong electric field (E) (*a*). The accelerated carriers start causing a ionization event, creating an avalanche filament (*b*). In this phase the charge increases in a region close to the seeding point and expand along a filament, parallel to the E. This phase is called avalanche build-up. Moreover, secondary photons are generated and they can interact creating new seeding points. The fast growing of the charge in the filament increases the gradient of charge density with respect to the surrounding substrate. This increment, induces a spread of the avalanche charge, termed lateral propagation (*c*). The avalanche keeps propagating until it expands to the entire area of the considered device (*d*). The current flow continuous until the avalanche is properly quenched.

2.1.7 Timing Performance

The timing precision of a SPAD is usually defined as single-photon time resolution (SPTR), or timing jitter, and expressed in terms of the full width at half maximum (FWHM) of the

instrument response function (IRF). This parameter expresses the precision of a SPAD in detecting the arrival time of single photons. In Fig.2.12 *(b)*, the standard shape of the IRF is shown. The envelope of the function has two main characteristic components defined as the peak and the tail [103], [188], [192]. As shown in Fig. 2.12 *(b)*, the time delay at which a photon is absorbed and the corresponding avalanche signal is detected, is not constant. This delay depends on the cross-section region where the photoelectron is generated (Fig. 2.12 *(a)*). In general, we can express the detection time ($T_{detection}$) with the following expression [26]:

$$T_{detection} = T_{transit} + T_{build} + T_{prop}. \tag{2.17}$$

The previous expression is composed of three terms and their statistical fluctuations. Let us analyze them one by one. To explain the first term ($T_{transit}$) of the equation 2.17, we need to have a closer look at the structure of a general SPAD cross-section and the profile of the electric field across it (Fig. 2.12 *(a)*). The avalanche can be triggered only in the region where the electric field is high enough (i.e., above the critical value). For this reason, we can split the depletion region in two parts: the high field multiplication region and the low field drift region (Fig. 2.11). In the first region, a carrier can promptly trigger an avalanche. On the other hand, if an optical photon is absorbed in the second region, the generated photoelectron needs to drift for a certain portion of the space charge region until the multiplication region before it can trigger an avalanche. The time needed by those photoelectrons to drift is defined as transit time ($T_{transit}$). Since the photon absorption, as explained in the previous sections, is a statistical process, the $T_{transit}$ has an inevitable spread that directly contributes to the device jitter.

The second term (T_{build}) summarizes the sources of randomness of the impact-ionization mechanism. Indeed, when a photoelectron is in the multiplication region, the avalanche current can grow in different ways. This statistical fluctuation can add an unavoidable delay to the signal until it reaches a set threshold. We usually refer to this mechanism as build-up [193]–[195]. Moreover, the impact-ionization is a very directional process that takes place parallel to the electric field. For this reason, other than the described initial build-up that takes place along a line parallel to the electric field and containing the interaction photon absorption point, we can distinguish another phase: the lateral propagation (T_{prop}). The lateral propagation is the last term of the Eq. 2.17. In this phase, the current starts to spread across the whole SPAD orthogonally to the direction of the electric field vector. Two main mechanisms lead to the lateral propagation of the avalanche current. The first relies on the multiplication-assisted diffusion due to the carrier density gradient between the initial build-up filament and the lateral regions [196]. Because of this strong gradient, the carriers diffuse laterally. This process also has random fluctuations because it relies on random phenomena (diffusion and impact-ionization) and because the propagation depends on where the initial filament is located (where the photon is absorbed and a photoelectron triggered the avalanche process). The other contribution to lateral propagation is given by the emission and absorption of secondary photons that can trigger another avalanche in the same detector but at a different location [197]. For this reason, the temporal response of the detector will be influenced by the

amount and location of secondary photons absorbed. Fig. 2.13 shows a simplified scheme of the avalanche growth.
The overall statistical dispersion of $T_{detection}$ ($\delta t_{detection}$) is commonly expressed in terms of FWHM of the IRF distribution, and its peak is commonly fitted with a Gaussian function. In some cases, carriers can be generated by the interaction of optical photons in the neutral regions. In this situation, the carriers slowly diffuse, and they may not reach the drift and multiplication region, where they can trigger an avalanche. Moreover, since these carriers do not undergo acceleration given by the electric field, they can experience a long random delay. This phenomenon causes the tail of the IRF.
The SPAD jitter has a dependency on the applied bias voltage and, in particular, an increment of the excess bias improves it. This improvement is due mainly to the build-up and lateral propagation contributions. In addition, this is due to the increased efficiency of the triggering and propagation phenomena.
Nevertheless, the gain of increasing the excess bias has a limitation set by a trade-off with the noise level that increases as well.
The size of the device is also an influencing parameter for the jitter performance. Indeed, it is easy to understand that the device's size will influence the avalanche's lateral propagation, degrading the jitter. This issue can be overcome by using a proper low-threshold front end that can detect the avalanche current when it is still at the beginning of the build-up process [198]. Thus, proper engineering of the electric field profile along the cross-section and proper front-end electronics can help get the best jitter performance. For the reader, the author suggests the works published in [199], [200] in 2021.
Several SPAD front-end strategies are described in Section 2.2. An accurate description of the jitter measurement protocol and setup is provided in Section A.7.

2.2 SPAD-Circuit Interface

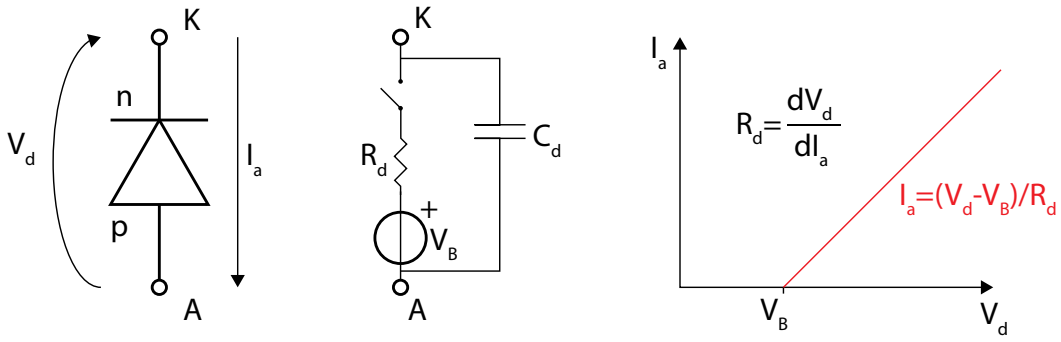


Figure 2.14: Simple SPAD equivalent circuit.

Fig. 2.14 shows a simple SPAD equivalent electrical model, where C_d represents the junction

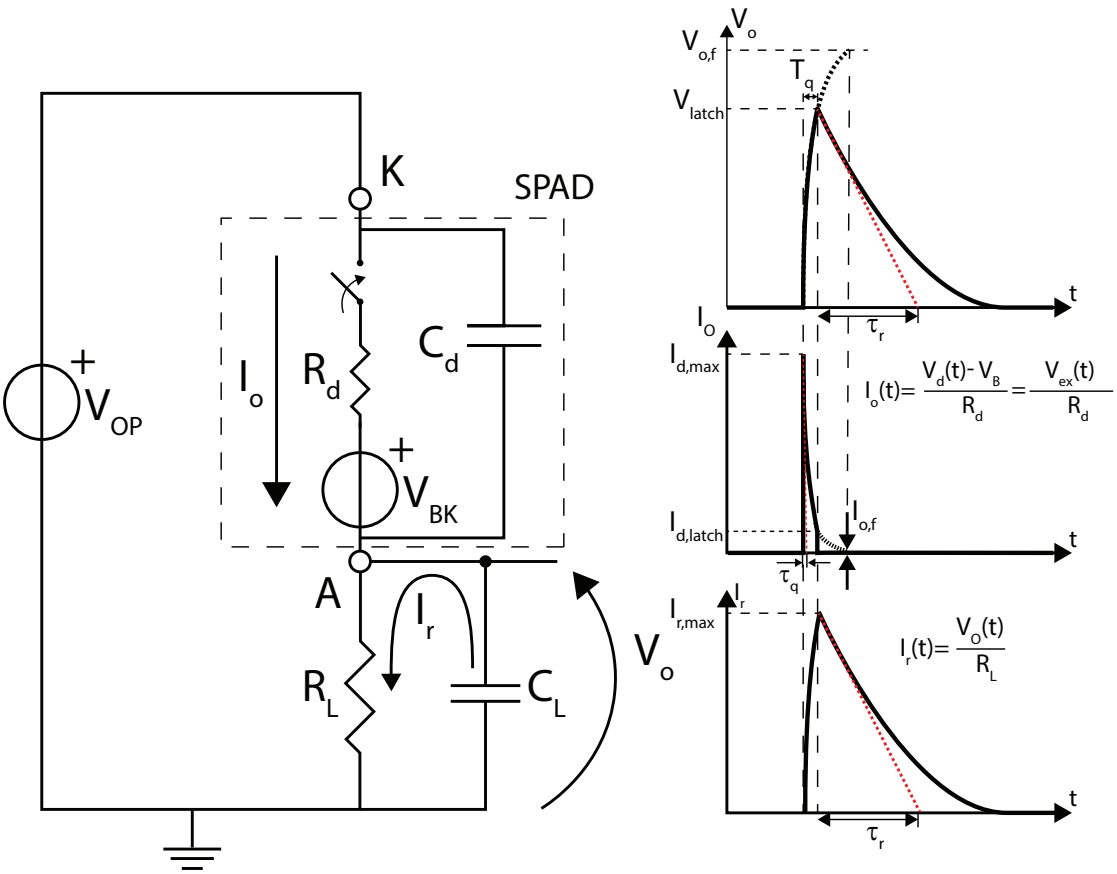


Figure 2.15: Passive quenching equivalent circuit (*left*) and corresponding output current and voltage profiles (*right*).

capacitance (typically on the order of 0.1-1 pF), and R_d is the diode resistance. R_d is given by the series of space-charge resistance of the avalanche junction and the ohmic resistance of the neutral semiconductor region crossed by the current [163]. Its value is generally in the range of 100-500 Ω. These parameters both depend on the SPAD structure and its size. V_B represents the breakdown voltage, and the switch is used to model the avalanche triggering given by optical photons that hit the active area. The front-end has a critical role in determining the performance of the device. Indeed, an inadequate front-end design can result in poor performance in terms of power consumption, count rate, afterpulsing, and timing jitter. In this section, several SPAD front-end architectures are reported and discussed.

2.2.1 Passive strategies

The easiest way to operate a SPAD device is to connect a resistive load in series to its cathode (or anode) and apply a high voltage (above breakdown) across the two components (Fig. 2.15). In this way, when a suitably biased SPAD fires, the avalanche current across the resistive load (typically larger than 100 kΩ) will quench it [163], [201]–[203]. This technique is called *passive*

quenching. The left-hand side of Fig. 2.15 shows the standard passive quenching and recharge structure. This technique's principle of operation can be split into two phases: the quenching and the recovery (or reset) transitions.

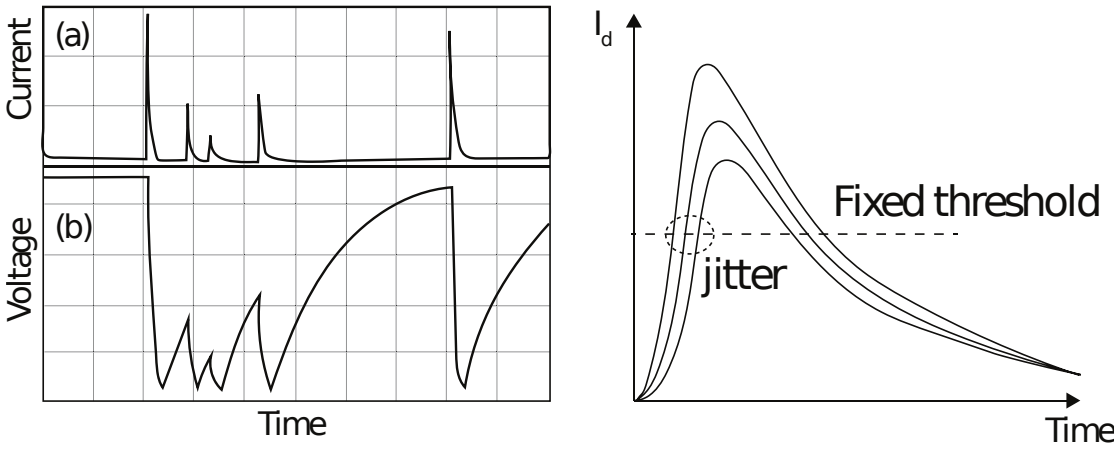


Figure 2.16: *Left side*: diode current (a) and voltage (b) waveforms of a SPAD biased with a passive quenching circuit (PQC); observe the retriggering of the SPAD during the recovery transition and the correlated non standard avalanche current pulses. *Right side*: comparison of diode current pulses occurred during the recovery transition and corresponding time-jitter in evaluating the photon detection time. [204].

Quenching Phase

During the quenching phase, the avalanche current charges the capacitances so that the output voltage V_o and I_d follow an exponential trend toward their asymptotic values $V_{o,f}$ and $I_{o,f}$ [163]:

$$I_{o,f} = \frac{V_{OP} - V_{BK}}{R_d + R_L} = \frac{V_{ex}}{R_d + R_L} \cong \frac{V_{ex}}{R_L}, \tag{2.18}$$

$$V_{o,f} = R_L I_{o,f} \cong V_{ex}, \tag{2.19}$$

where V_{ex} is the excess bias voltage applied to the device. In Eq. 2.18 the approximation is justified by the significant difference in value between the diode resistance R_d and the resistive load R_L used for quenching the device. The quenching is an exponential function with time constant τ_q:

$$\tau_q = (C_d + C_L)\frac{R_d R_L}{R_d + R_L} \cong (C_d + C_L)R_d. \tag{2.20}$$

When $V_o(t)$ approaches V_{ex}, the intensity of the current flowing through the diode ($I_d(t)$) decreases, as well as the number of carriers that cross the avalanche region (Fig. 2.15 *right-center*). As explained in the previous sections, the avalanche is a statistical process, and the probability that none of the carriers crossing the high-field region trigger an ionization cascade starts to be significant for current values on the order of $I_{d,latch} \cong 100\ \mu A$ (commonly used

value for the so-called *latching current*), rapidly increasing when this value decreases [163], [205]. Above this value, the avalanche has a self-sustaining behavior, and self-quenching below it. Hence, if I_d is small enough, with a high probability, all the carriers will leave the depletion region after a random time. Thus, no more ionization events will take place, and the avalanche will be fully quenched. As a first-order approximation of the quenching time (T_q), it is possible to consider the process as an exponentially-decaying current that stops when it goes below the *latching current* value:

$$T_q = \tau_q ln\left(\frac{I_{d,max} - I_{o,f}}{I_{d,latch} - I_{o,f}}\right). \tag{2.21}$$

Some considerations are needed to choose the correct value for the quenching resistor. First, we need to ensure that the asymptotic current $I_{o,f} \ll I_{d,latch}$. In this case, the current passes the latching value with a reasonable slope, and the avalanche is correctly quenched after a well-defined time lapse. On the other hand, if $I_{o,f} \cong I_{d,latch}$, the quenching occurs with a progressively more extended time and more significant time dispersion [163], [205]. If instead, $I_{o,f} > I_{d,latch}$, the avalanche will not be quenched, and a certain steady current flows through the device. In the latter case, the power consumption can cause excessive heat generation with consequence permanent device damage. For this reason, large resistance values need to be used (typical values range from 50 kΩ to 500 kΩ). However, a too high value of the quenching resistance brings some problems that will be discussed in the next part of the section.

Recovery Phase

The recovery (or reset) phase is when the device recovers the total reverse bias across its terminals and is thus reset. It corresponds to the opening of the switch in Fig. 2.15. During the recovery phase, C_d and C_L slowly discharge through the load R_L so that the bias voltage across the diode exponentially recovers with a time constant τ_r:

$$\tau_r = R_L(C_d + C_L). \tag{2.22}$$

The voltage is considered restored after $\sim 5\tau_r$ within 1%, considering the exponential transient. The reader should note that the dead time of the SPAD is not well defined using passive quenching and that the triggering probability of a photon hitting a device's active area increases progressively as the bias voltage across the SPAD rises. Hence, a photon can trigger an avalanche before the SPAD bias is fully restored (Fig. 2.16 *left*). In this case, the detection efficiency is low. In addition, a SPAD triggering during the recovery transition has degraded performance because of the time-varying excess bias and, consequently, a different peak avalanche current. Also, additional jitter is introduced because of the pulses with different amplitude (Fig. 2.16 *right*).
Another critical point to be highlighted is the influence of the quenching resistance R_L on the recharge time. Although a high value of R_L is beneficial for the quenching of the device, it has an adverse influence on the recharge phase. Indeed, τ_r increases proportionally with R_L. This could be non-negligible in the case of a high count rate. Indeed, as shown on the *left* of Fig.

2.16, the pulses may not fully recover after an avalanche event, and, at a too high frequency, this will imply severe count losses.

This straightforward quenching strategy is still adopted also in the case when the SPAD is an isolated device, and the quenching is implemented by means of an external circuit with discrete components. However, the use of external discrete components has several disadvantages in terms of performance. Indeed, the wire bonding and long connections will increase the parasitics dramatically at the output node of the SPAD (the output capacitance can easily reach 10 pF). As explained in the previous sections, this can adversely influence the afterpulsing probability and the jitter. Indeed, a large amount of current will be needed to charge the parasitic capacitance and quench the SPAD. This large current will increase the number of carriers that can be trapped during an avalanche.
Moreover, the signal will be detected when the amount of current reaches a level that goes well beyond the one developed at the beginning of the build-up phase, increasing the timing dispersion. For this reason, both in custom and CMOS technologies, integrated structures were adopted. In the case of custom technologies (mainly used for the implementation of analog SiPMs), the quenching load is typically implemented through a poly-resistor and an external low-threshold comparator is used to detect the avalanche pulses early on. On the other hand, when CMOS technologies are used, the load is usually implemented using a transistor biased in linear mode. The latter, although it enables tuning the value of the load, limits the maximum value of the excess bias which can be applied to the SPAD (Fig. 2.20 *(a)*). Most of the limitations and problems highlighted in this section were overcome using more sophisticated solutions. These solutions, including active strategies, are discussed in the following section.

2.2.2 Active strategies

Several limitations affect the standard passive quenching and recharge technique, discussed in the previous section. To overcome these issues, for the first time in 1975, *active quenching* was proposed in [89] (Fig. 2.17). Initially, this technique relied on discrete components, as in [202]. Afterward, nuclear instrumentation modules (NIMs) became commercially available [211]. These solutions are based on the detection of the avalanche and a feedback path that acts on the device's bias. They use a low impedance path to sense the avalanche current and implement the quenching and recharge branches by employing active components. In these kinds of systems, the voltage across the SPAD goes below the breakdown level because of the circuit action during the quenching phase. In this way, no fluctuation given by the statistical nature of the avalanche process influence the quenching time. Once the avalanche is quenched, the SPAD is kept off for a specific time, called *hold-off*. Then, the reverse bias across the SPAD is fully restored, again using active components (Fig. 2.18).
In the first implementations of this strategy, since the circuit is not integrated side-by-side with the SPAD, it might be possible that the system detects the avalanche with a non-negligible delay. This delay depends both on the technological properties of the SPAD (e.g., the peak

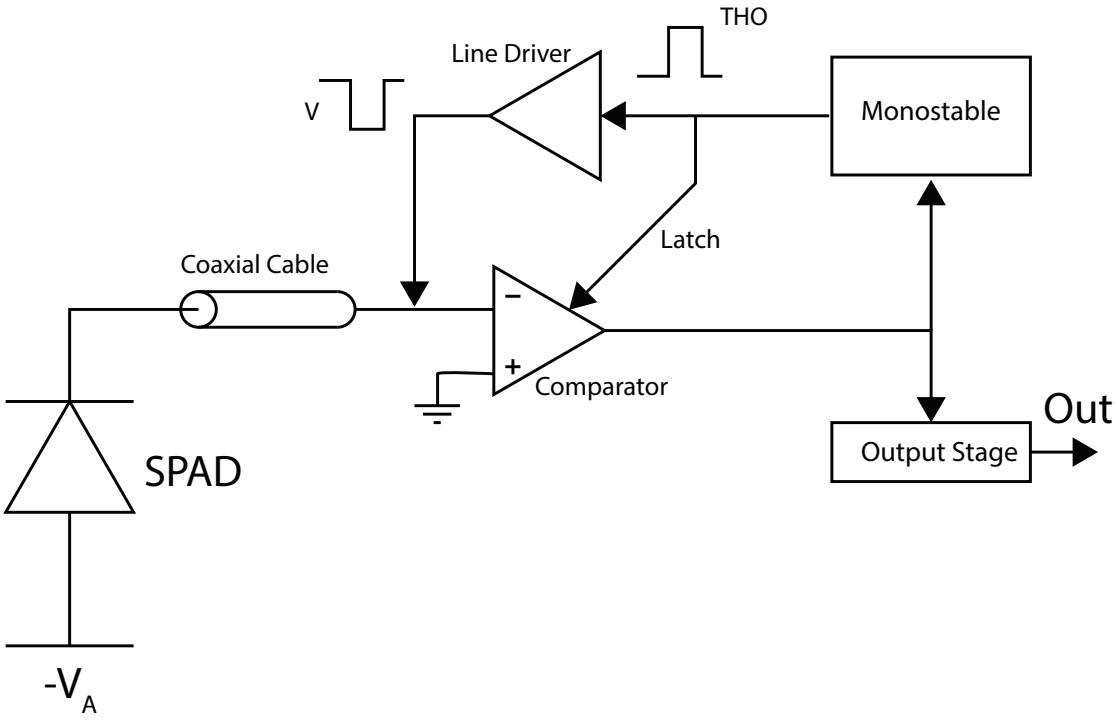
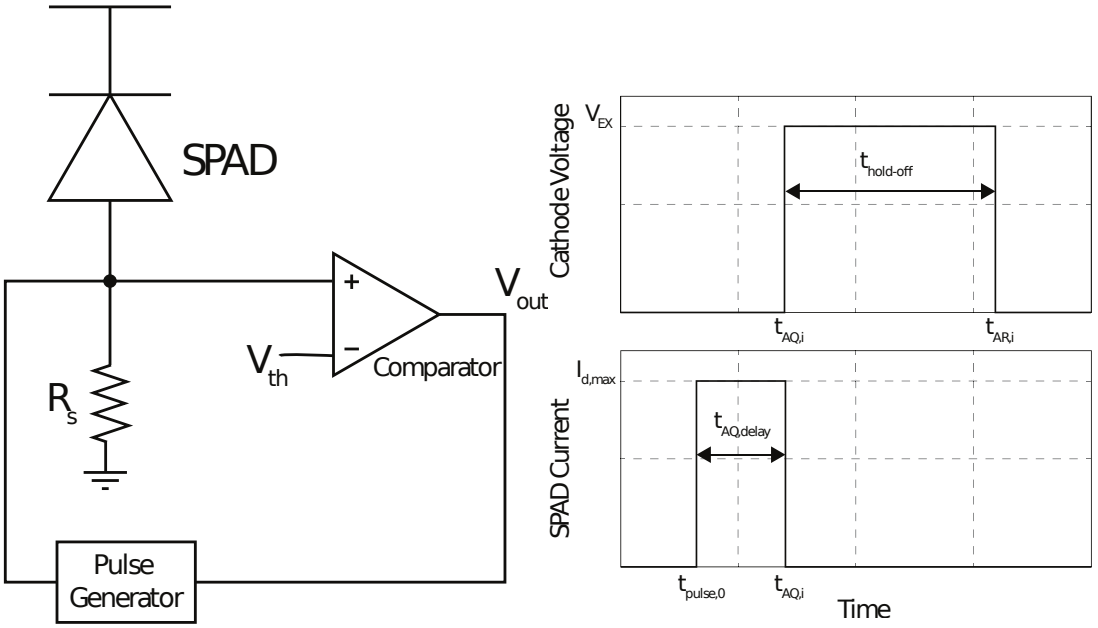


Figure 2.17: Early active quenching and recharge circuit implemented with discrete component to remotely detect the avalanche current and quench the SPAD [89], [204].

value of current) and the active circuit (e.g., sensitivity threshold, propagation delay). The use of a system which is not well designed and too long delays, can lead to significant problems. An example is represented by the afterpulsing probability that can increase significantly since the peak avalanche current ($I_{o,f}$ of Eq. 2.18) flows for a long time, and a large number of carriers can be trapped. In this situation, the charge flowing through the device can also be larger than the passive quenching case, making the use of an active solution unjustified [203]. It is important to note that, in this case, the avalanche charge does not depend on the parasitic capacitance on the SPAD output node, but it is linked to the delay time before the active circuit quenches the avalanche. After the quenching, the circuit keeps the SPAD below breakdown for a certain hold-off time. After that, a recharge pulse is sent to restore the bias voltage. The active reset is carried out through a low impedance path. Avalanches triggered during the reset phase can not be sensed until the reset phase ends. For this reason, the quenching time for avalanches that start during the reset phase is longer and can cause a higher afterpulsing probability. Thus, the reset time must be as short as possible. A direct influence on the reset time is given by the parasitic capacitance load on the output node of the SPAD. Indeed, the larger the capacitance load, the longer the time needed to discharge it, and the power consumption. The reader should note that the main disadvantage is given by the implementation using discrete components. Indeed, large parasitics and long delays can compromise the operation of this kind of system.
During the years, much effort has been spent on improving the performance of the active



(a) (b)

Figure 2.18: Active quenching and recharge simplified scheme (*left*), and corresponding voltage and current waveform across the SPAD (*right*).

approach [73], [203], [204], [206], [209], [210], [212]–[230]. Until 2010 only rarely a pure active quenching approach was used [73], [203], [215], and no fully integrated systems were demonstrated together with SPADs. More recently, some works show the fully integrated implementation of active quenching circuits [218], [220], [223], [227], [230].

2.2.3 Hybrid Strategies and Other Integrated Solutions

Recent implementations mainly rely on a hybrid approach [214]. This solution combines the advantages and to overcomes the issues of the two pure approaches, introduced in the previous section. The most common hybrid approach uses a combined active-passive quenching and an active reset.

In this approach, the quenching branch is composed of a passive load in parallel to a switch driven by an active feedback that controls the SPAD's recharge. Fig. 2.19 *(a)* shows a schematic view of the active recharge and quenching mixed solution. The operation of such a circuit can be split into five phases (Fig. 2.19 *(b)*): passive quenching, active quenching, hold-off, and recharges. When a photon triggers an avalanche, a current starts to flow across the passive load. When the voltage across the load reaches the threshold of the discriminator, the active quenching loop is activated, and the S_Q switch is closed. An additional current is injected in the A node, increasing the slope of the quenching pulse. Once the quenching phase ends, S_Q opens again. The voltage at the node A discharges slowly through the passive load. The control logic activates again after a certain time, usually tunable, closing the S_R switch to recharge the SPAD and restore the full reverse bias. The active reset branch enables choosing a

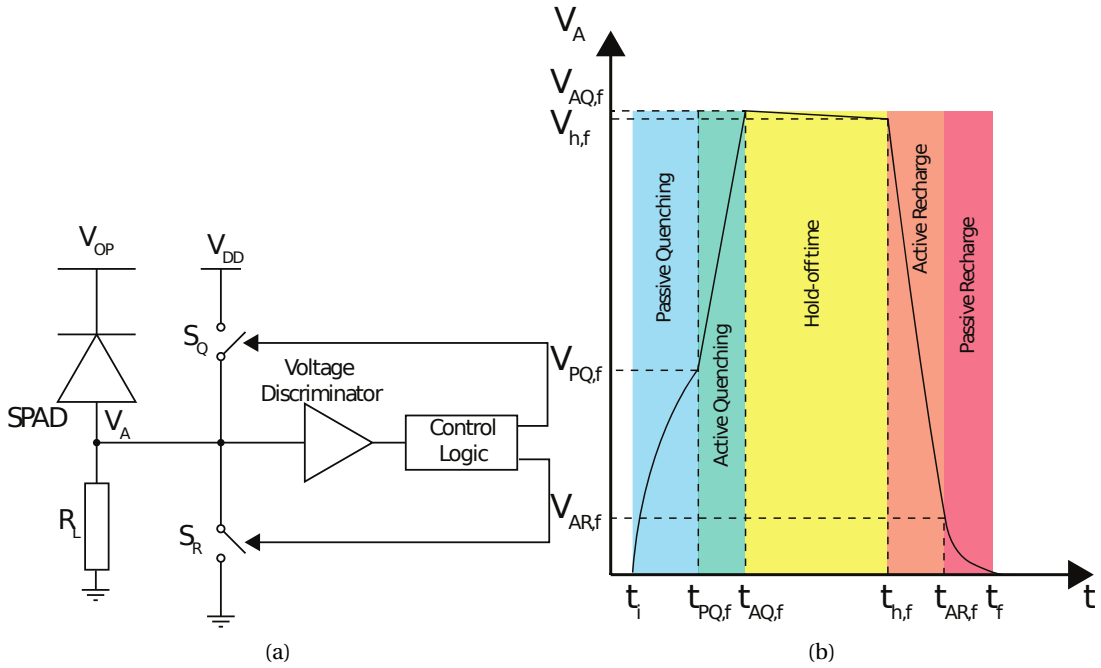


Figure 2.19: *(a)*: Mixed active quenching and recharge simplified scheme. *(b)*: Timing diagram of the SPAD anode voltage (V_A). When an avalanche is triggered, the passive quenching increases V_A. After a certain time $t_{PQ,f}$ the active reset activates and rapidly fully quenches the SPAD in a time $t_{AQ} = t_{AQ,f} - t_{PQ,f}$. The SPAD is then kept off for a certain hold-off time ($t_h = t_{h,f} - t_{AQ,f}$). During this period the node A can be sligltly discharged through R_L until the $V_{h,f}$ level. Then, the control logic activates the active reset with a pulse width of $t_{AR} = t_{AR,f} - t_{h,f}$. After that the active reset stops, and if a small voltage level ($V_{AR,f}$) remained on node A, it is discharged completely through R_L.

very high value for the passive load resistor to mitigate the trade-off explained in the previous section. Moreover, in this case, a high value of the passive load is desirable to avoid large changes in the voltage level during the hold-off time, while the additional capacitive load added when using transistors does typically not exceed the tens of femtofarads. Moreover, the use of deep sub-micron technologies can guarantee a fast circuit response and an additional improvement of the performance, in terms of afterpulsing [158], [210], [231], [232], power consumption [210], count-rate [158], [210], [232], and timing jitter [224], [232]. In contrast to pure active quenching, this solution limits the overall avalanche charge flowing through the device. Indeed, even if the reaction of the active quenching has some delay, the passive branch starts working immediately. Hence, the overall charge will be the same as in the passive quenching approach, although the quenching time decreases dramatically. However, given the area occupation of this circuit, the fill-factor is degraded when a 2D array is implemented. Nevertheless, solutions using simple circuits with fewer transistors have been recently shown (as is [232] that will be discussed in more detail in the following chapter). In addition, recent research allowed to explore alternative solutions for the implementation of arrays. The 3D

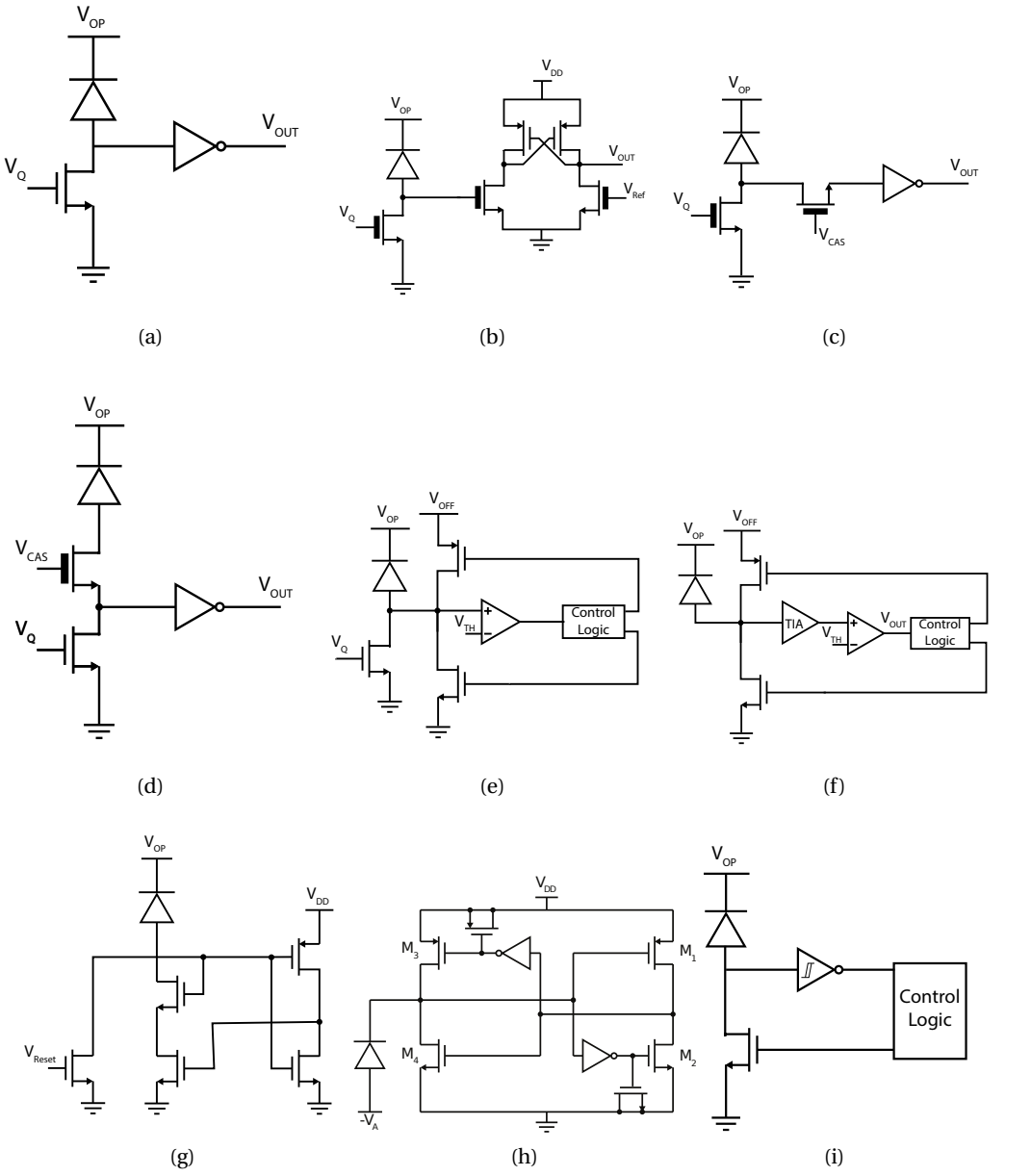


Figure 2.20: Several examples of SPAD quenching circuits: Standard passive quenching with nMOS as tunable resistive load *(a)*; passive quenching with the use of thick oxide transistors to increase the applied excess bias and level shifter to protect the lower voltage core logic *(b)*; passive quenching using a thick oxide transistor as tunable load and an output clamp transistor in series to the output *(c)*; passive quenching using a cascode structure to increase the applied overvoltage and protect the downstream electronics *(d)* [158]; active quenching and recharge structure using a low-threshold comparator for the early detection of the avalanche pulse *(e)*; active quenching and recharge with a transimpedance amplifier (TIA) to amplify the avalanche current signal *(f)* [206]; current-mode quenching technique *(g)* [207]; dynamic quenching *(h)* [208]; variable-load quenching *(i)* [209].

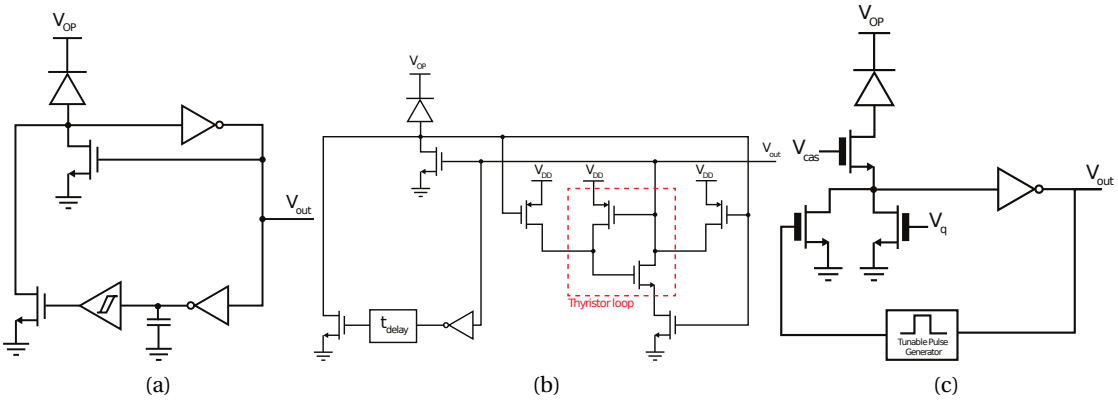


Figure 2.21: Compact quenching circuit *(a)* [210]; Thyristor-based quenching circuit *(b)* [210]; Cascode passive quenching with active recharge loop *(c)* [158].

integration, in particular, demonstrated to be an effective solution to overcome the fill-factor problem. Two flavors of 3D integration exist, named 3D FSI and 3D BSI. The next section will discuss these implementations.
Together with the approach just presented, other kinds of hybrid approaches are possible. Passive quenching with active reset represents an example of those [158], [224], [231]–[236]. This approach implements the quenching by only a passive branch, and only the reset has a feedback loop with active components. Another option is to implement a circuit with an active quenching and a passive reset [216], [237]–[239].

Together with the approaches discussed in this section, other solutions aiming at full integration were developed. In Fig. 2.20, the schemes of several example circuits of different quenching implementations are collected. An example is provided by the dynamic quenching presented in [208], [240], based on a thyristor implementation for the quenching mechanism (Fig. 2.20 *(h)*). In this circuit, when an avalanche is triggered, the high impedance seen at the cathode node of the SPAD starts a passive quenching process. When the cathode voltage starts going down, the PMOS M1 activates together with the transistor M4. The action of these two transistors quickly brings the cathode voltage of the SPAD to the ground voltage level of the circuit, quenching the device. After a delay equal to the propagation of the signal through the two blocks composed of an inverter and a MOS capacitance, the remaining two transistors turn on, resetting the cathode voltage to the initial value. In this case, the hold-off time is fixed, but the circuit could be modified to change it or make it tunable. More recently, a similar approach was proposed in [210] in a 90 nm technology node. In this architecture, the quenching is performed dynamically, modifying the impedance of a transistor used as a ballast resistor (Fig. 2.21 *(b)*). Indeed, the quenching transistor is turned off as soon as the thyristor-based front-end detects the avalanche pulse. The front-end is triggered when the output voltage goes below $V_{DD} - |V_{th,p}|$, where $V_{th,p}$ is the threshold of the pMOS transistors. It should be noted that in this implementation, the output node is kept at V_{DD} in steady condition. After the avalanche is quenched, the output goes to zero until a pulse propagating

through a tunable delay element reaches a reset transistor, recharging the SPAD.
In the same work [210], another architecture called *compact quenching circuit* (CQC) is also presented (Fig. 2.21 *(a)*). This architecture uses a similar approach as the previous circuit. Indeed, the avalanche is initially passively quenched by a transistor. When the avalanche is detected by an inverter used as a voltage discriminator, the quenching resistor is turned off, increasing the quenching resistance and then limiting the current. The signal propagates then through another inverter used as a delay cell and loaded with a capacitor. A Schmitt trigger is used to restore the fast transition edges and be less sensitive to noise during the transition of a slow signal. The output of the Schmitt trigger activates a reset transistor that restores the bias across the SPAD.
A different approach is presented in [207]. The work shows a *current-mode quenching* approach. The avalanche current is sensed through a current mirror, also used to increase the transistor's resistance in series to the SPAD. The current flow is eventually interrupted by turning off the series transistor, quenching the SPAD. However, the resistance is increased during the quenching phase, starting from a low value, which implies a longer quenching time. The current-mode circuits are typically faster than the voltage-mode ones. For this reason, this circuit should be able to provide faster transitions. However, the results are provided only in simulation and with a discrete components implementation. There was no fully integrated implementation of the mentioned current-mode quenching until now.
To conclude, the author would like to mention a different strategy to operate the SPADs, i.e. the gated operation. To have a better description of this operation mechanism, the author suggests the review [203] to the interested reader.

2.3 Array Architectures

The natural evolution of the SPAD-based systems is the implementation of arrays composed of several pixels. An example of array implementation in standard CMOS technology was shown for the first time in 2003 [128]. Over the years, many architectures have been designed and tuned to a variety of applications [137]. This chapter provides a brief overview of the main categories of SPAD arrays and a description of their operating principle.

2.3.1 Linear Array

This architecture places the SPADs along a line, and the pixel circuits are implemented next to them, outside the linear array area. In this configuration, all the pixels can operate in parallel. Moreover, additional electronics can be integrated beside the sensor (e.g., timing circuits, processing units) without compromising the sensor's fill factor. These devices are typically operated using a scanning approach, in which a moving element in the setup allows the device to scan the observed scene and create in post-processing a 2D image. Some examples of this kind of architecture can be found in [241]–[244] and [156], where the array was implemented using a 3D BSI architecture (see 2.3.4 for more details). Fig. 2.22 shows the architectural

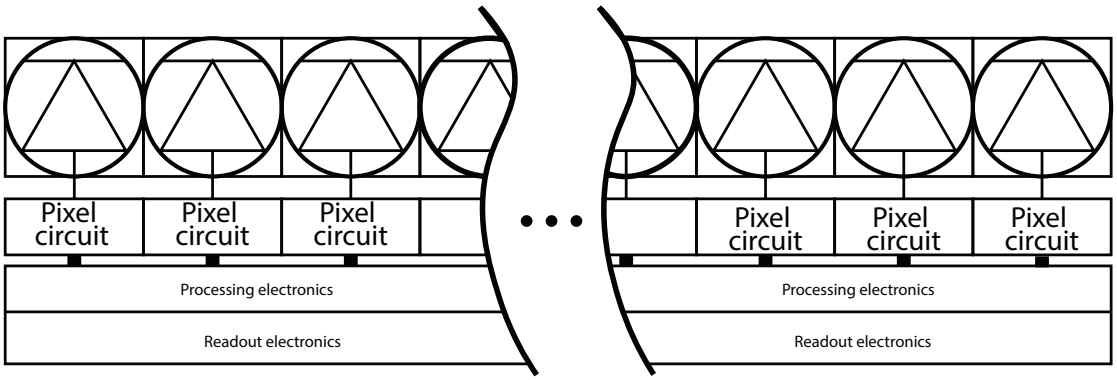


Figure 2.22: *(a)*: Linear array. The pixels are organized along a line. The front-end pixel electronics is placed beside the detectors and does not limit the fill factor of the array. Additional electronics can be placed next to the array to implement on-chip processing, readout and/or timestamping.

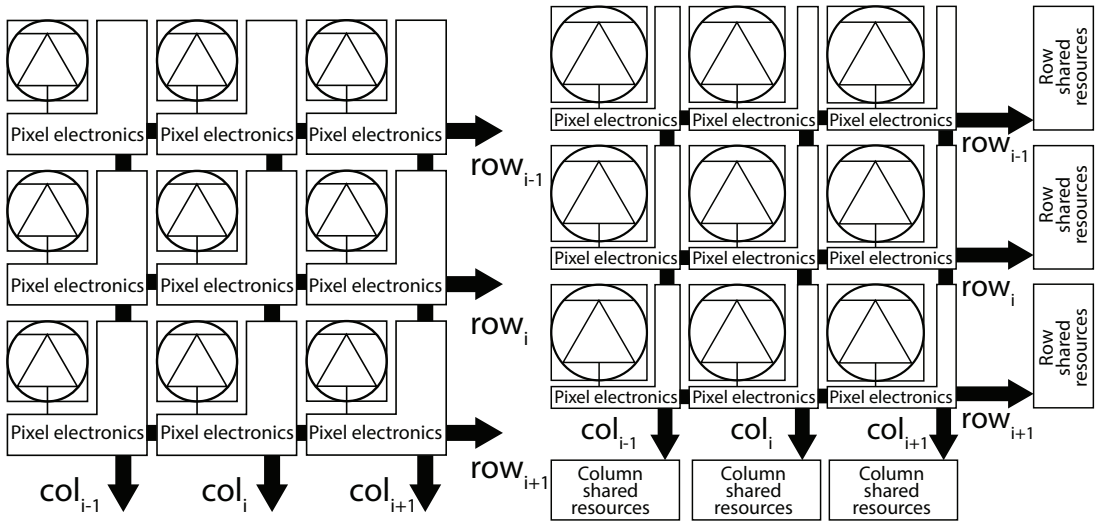


Figure 2.23: (*a*): 2D array with full independent circuit per pixel. (*b*): 2D array with resource sharing for column and/or rows. Figure adapted from [137]

scheme of a linear arrays.

2.3.2 2D Arrays

These systems are capable of acquiring a full 2D image directly. However, the complexity of the architecture is remarkably increased compared to 1D arrays. Indeed, in this kind of device, the SPADs are organized in a 2D matrix, and a pixel circuit for quenching and recharge is integrated close to each SPAD (Fig. 2.23). The presence of the circuit limits the overall fill factor of the array (Fig. 2.23 (*a*)). For this reason, in many cases, the circuits used are simple

and integrate a reduced number of transistors. Several architectures of 2D SPAD arrays are available nowadays [137].
Nevertheless, some main characteristics can be highlighted in all of them. Indeed, in 2D arrays, each pixel is independent and has its own pixel circuit (Fig. 2.23 (*a*)). Depending on the specifications and the target application, the designer can also increase the complexity of the pixel circuit, inserting additional elements (e.g., memories, counters, processing units, timing circuits) in order to acquire additional information directly on-chip, with consequent degradation of fill factor [126]. The data from each pixel are collected, preserving the information on the SPAD that generated the data. All supplies, control, and data signals can be shared across rows and columns, leaving just the need to integrate a line driver at the pixel level. This choice can significantly improve the fill factor (Fig. 2.23 (*b*)) [245]. This choice allows to reduce the size of the needed electronics and maximize the fill factor of the array, together with a potential reduction in its power consumption and area occupation. For a complete review of the developed solutions in the last fifteen years, the author suggests the review [137] to the interested reader.

2.3.3 Silicon Photomultipliers

The silicon photomultiplier (SiPM) (also called solid-state photomultiplier, SSPM, or multi-pixel photon counter, MPPC) is a family of solid-state devices integrating a high number of SPADs, also called microcells or pixels [14], [67], [69], [73], [246]–[250]. Large sensing size and both single- and multi-photon detection capability made this family of detectors very attractive for several applications, including high-energy and medical physics.
Throughout the years, thanks the improvement of the technologies and and increased interest in these kinds of devices, two main categories of SiPMs emerged: analog SiPM (aSiPM) and digital SiPM (dSiPM).
Next, we present an overview of these two categories, their characteristics, and their principles of operation.

Analog Silicon Photomultiplier

The analog silicon photomultiplier was invented in 1988 [251], and it launched a revolution in the way of performing light detection quickly supplanting PMTs in many applications [67], [69], [250], [252]–[258].
The aSiPM is an array of SPADs where each device has its own quenching resistor, forming a pixel. The pixel size can range from 100 μm [259] down to 10 μm to improve the dynamic range of the detector [260]. All the pixels are connected in parallel (Fig. 2.24). When a photon hits one SPAD generating an avalanche, a current starts flowing through the pixel. First, the avalanche is quenched through the resistor, and then the bias is restored across the SPAD (see Section 2.2.1 for more details on the passive quenching mechanism). In this way, the current pulses generated by the avalanche processes in the SPADs are summed up in an analog way. Thus, the aSiPM output signal is composed of the superposition of signals of all the single

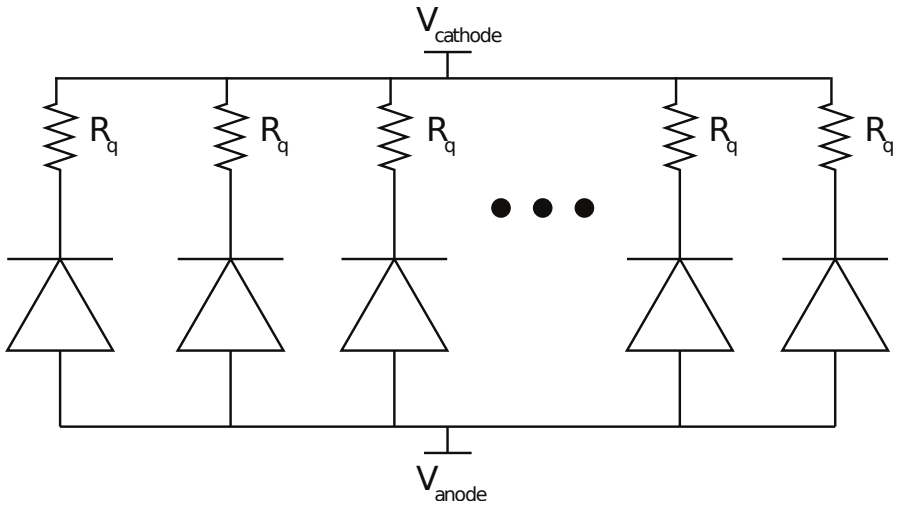


Figure 2.24: In an analog silicon photomultiplier (aSiPM) all the SPADs are connected in parallel. The output current is the analog sum of the currents of the single cells.

pixels that compose it. A schematic view of this architecture is shown in Fig. 2.24. If multiple photons hit the device simultaneously, the output signal will be higher and proportional to the number of detected photons (see Fig. 2.26). Indeed, aSiPMs make it possible to perform photon counting, relying on the output signal level. Knowledge of the number of detected photons is very useful in many applications where the amount of light detected contains essential information (e.g., PET).

In Fig. 2.25, the electrical model of an aSiPM is shown. As it is possible to see, the circuit is similar to the one for a single SPAD device, with the difference that in the aSiPM, additional passive components model the not-triggered SPADs including the inductance and capacitance (even higher than 100 pF in value [69]) of the connection grid [67], [69], [261]–[264]. These parasitic elements contribute to load the output of this device, limiting the bandwidth and the amplitude of the device's output and potentially compromising its timing performance. For this reason, a careful design of PCBs and front-end circuit is mandatory to preserve the properties of this device. Also, depending on the SiPM readout, in "voltage" or "current" mode, different frequency behaviors of the sensed signal can be observed [67], [261].

When an avalanche occurs in one SPAD, the diode capacitance C_d is discharged across the diode resistance R_d. As a result, a voltage drop is observed on the load capacitance C_q. This drop induces an external current pulse. The limit of the signal rise time is given by the following equation [262]:

$$\tau_r = (C_d + C_q)\left(\frac{R_d R_q}{R_d + R_q}\right), \tag{2.23}$$

where R_q is the quenching resistor. Once the avalanche is quenched, the SPAD voltage recovers following the time constant $\tau_{recharge}$:

$$\tau_{recharge} = R_d(C_d + C_q). \tag{2.24}$$

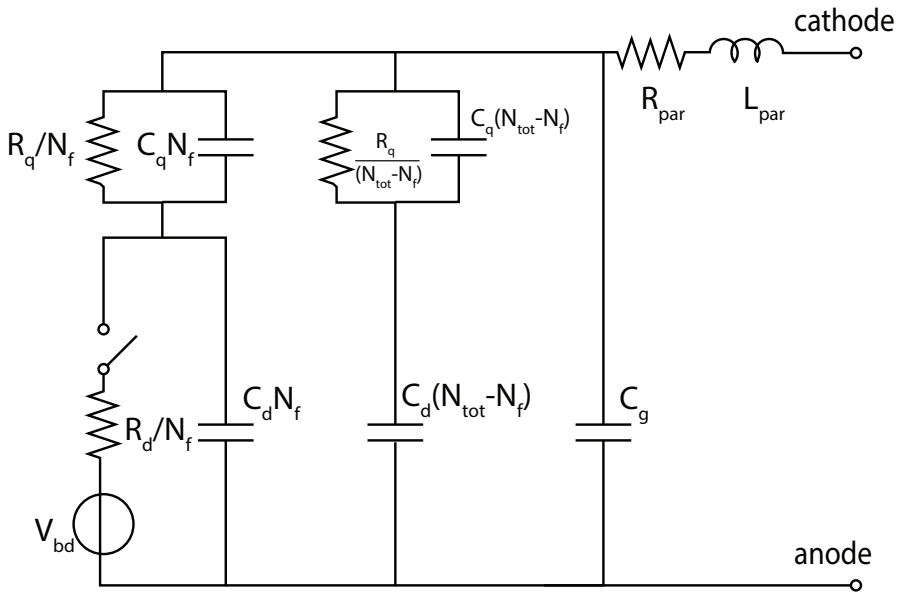


Figure 2.25: Typical electrical model for an analog silicon photomultiplier (aSiPM). R_d is the diode resistance, n_f is the number of SPADs that fired, V_{bd} is the SPAD breakdown voltage, R_q is the quenching resistor, C_q is the load capacitance linked to the quenching branch, N_{tot} is the total number of SPADs in the device, C_g is the grid and bonding pad parasitic capacitance, R_{par} and L_{par} are the parasitic resistance and inductance linked to the connections [67].

In an aSiPM, the output current produced by a single pixel is divided by the parasitic capacitance (C_g) (given by the grid and bonding pad) and the series connection of the passive capacitances of the inactive cells ($(N_{tot}-n_f)C_q$ and $(N_{tot}-n_f)C_d$, with N_{tot} the total number of pixels and n_f the number of SPADs that fired) [67], [69], [265]. For this reason, the parasitic capacitance can drastically decrease the single SPAD signal at the output. If a voltage mode is used for reading out the aSiPM output, the signal amplitude is typically on the order of tens to hundreds of millivolts, and its amplitude will depend on the number of optical photons detected (Fig. 2.26). The output signal characteristic highlights, once more, the importance of a proper readout in order not to degrade the device performance.

A detailed description of this kind of device is outside the scope of this thesis. See [67], [69], [250] for a review.

Digital Silicon Photomultiplier

The second category of silicon photomultiplier is represented by its digital version (dSiPM). This device, conceptualized and delivered as a product for the first time in 2009 [73], was born from the research of a perfect detector for applications like PET and to overcome all those issues related to its analog counterpart, i.e., low signal level, high output capacitance, time walk of the signal [74], [267].
In a standard dSiPMs, each SPAD has its own pixel circuit that can vary from design to design

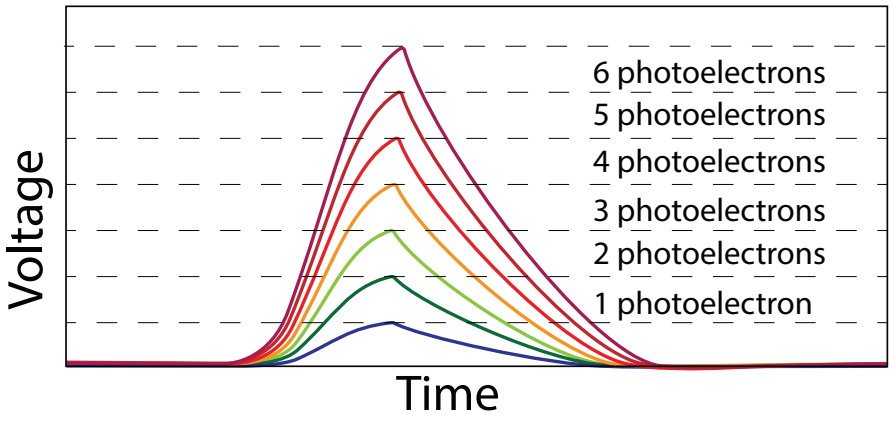
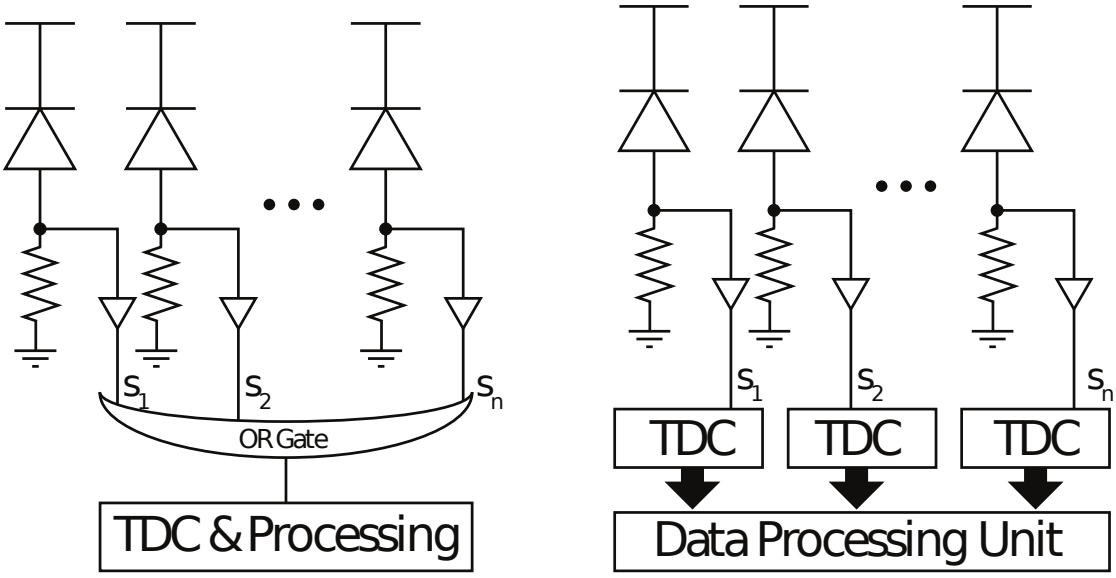


Figure 2.26: The analog silicon photomultipliers (aSiPMs) feature a discrete output when illuminated by brief low-level light pulses. The signal level at the output will increase together with the amount of detected photons. The increment shows quantization steps that can be used to determine the amount of photons detected in a single event.

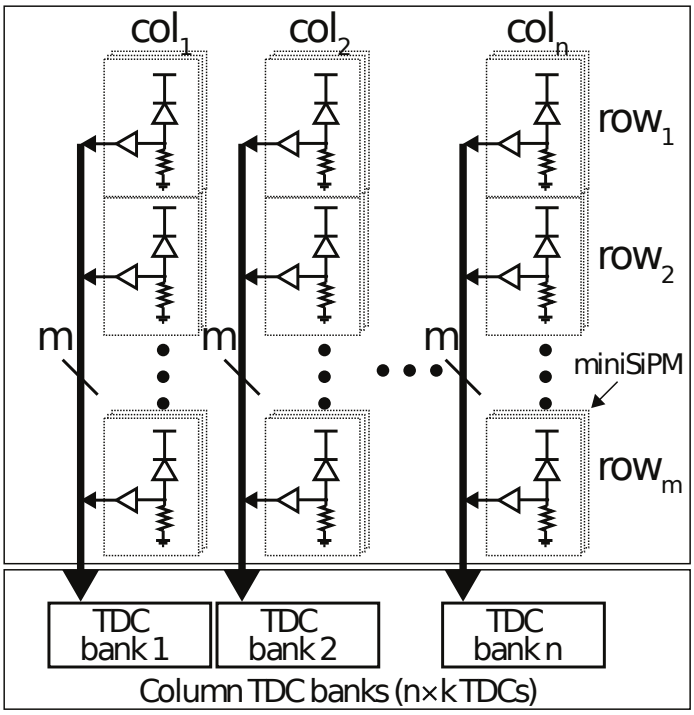
considering the design specifications (see Section 2.2). Their operation typically consists of the following steps. First, the output of the pixel is digitized and propagated using buffers. Then, all the digital SPAD signals are combined using a digital logic *OR* function. The output of this combinational gate is then fed to the downstream electronic circuit that adds additional functionalities to the system and reduces the need for external components, such as time-to-digital converters (TDCs), counting systems, and processing units. In Fig. 2.27 (*a*), the typical architecture of a dSiPM is shown. In this kind of architecture, the timing resolution for single-photon detection is mainly given by the SPAD jitter together with the circuit noise. In addition, imbalance of the propagation data path of the SPADs can create a systematic skew that, if not characterized and corrected, could degrade the timing performance as well. Thus, as explained in [80], a dependency of the timing performance on the architecture can be highlighted. Hence, a careful design is needed in the architectural choice, addressing all the design parameters that can compromise the device's performance.
Nevertheless, the dSiPM provides several advantages that brought researchers to develop improved architectures to solve/mitigate these issues [35], [38], [39], [80], [266]. Indeed, the early digitization of the signal at the SPAD level guarantees a very high SNR to the device. Moreover, the signal of one SPAD is not influenced by the other inactive pixels and keeps its amplitude and characteristics unchanged. Another advantage of the dSiPM is the possibility of providing multiple timestamps corresponding to each detected photon. However, in the architecture presented in Fig. 2.27 (*a*), this is not possible for two reasons. The first reason is the fact that a pulse entering in the OR gate will digitally *mask* all the other pulses arriving within its pulse width. This will in turn limit the number of pulses that the downstream electronics can discriminate. The second reason concerns the conversion time of the TDC. Indeed, this device typically requires some time between two successive conversions, sets an upper bound to its conversion rate. The natural solution to this problem could be to redirect the effort in designing a chip that follows the architecture described in Section 2.3.2 and



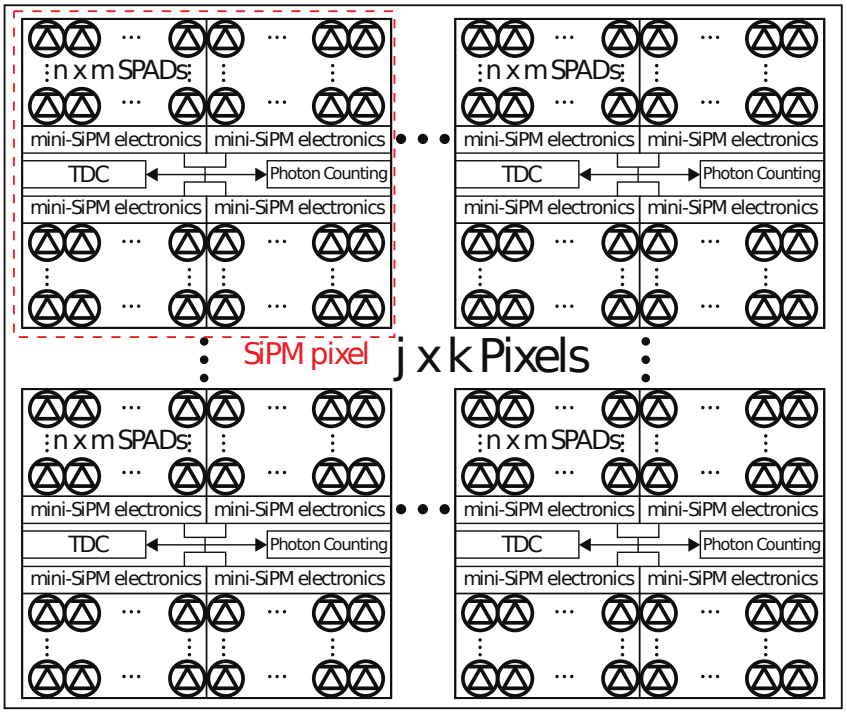
(a) (b)

Figure 2.27: (*a*): Standard dSiPM. (*b*): Ideal dSiPM architecture.

inserting the TDC at the pixel level (Fig. 2.27 (*b*)). Nevertheless, this solution will dramatically reduce the fill factor and dramatically increase the number of readout channels, together with the amount of data to be read out.
The multichannel digital SiPM (MD-SiPM) concept was introduced for the first time in 2012 [268]. This architecture overcomes the problems related to the original dSiPM implementation [80], [269]. In this approach, the array is segmented in smaller clusters called miniSiPMs (Fig. 2.28 (*a*)). The miniSiPMs are organized in cluster columns. Within the cluster columns, the miniSiPMs share a certain number of TDCs (a TDC every three pixels rows) to ensure improved timestamping capability and a high level of resource sharing. The structure presented can provide higher granularity in the spatial and temporal domains and keep the fill factor very high. The author suggests referring to [49], [80] for more details.
Another MD-SiPM architecture is presented in [266] with the chip named SPADnet (Fig. 2.28 (*b*)). This architecture does also rely on capturing multiple timestamps per event and takes advantage of the underlying *photon statistics*, particularly useful in applications such as PET. Between the new concepts introduced in this implementation, there are the concepts of *spatial compression* and *temporal compression*. The spatial compression consists in routing the signal of a group of SPADs to the input of an OR-function gate. As mentioned before, this implies that if one of the SPADs triggers during each other's deadtime, it will not be visible at the output. The output is then *compressed*. Thus, the higher the number of SPADs connected to the same OR gate, the higher will be the resulting compression. Although this strategy can lead to count losses at the output of the tree, it provides substantial advantages at system level. Indeed, it allows a large sharing of resources with consequent reduction of area and power consumption. On the other hand, temporal compression is based on the shrinking of the pulse propagating

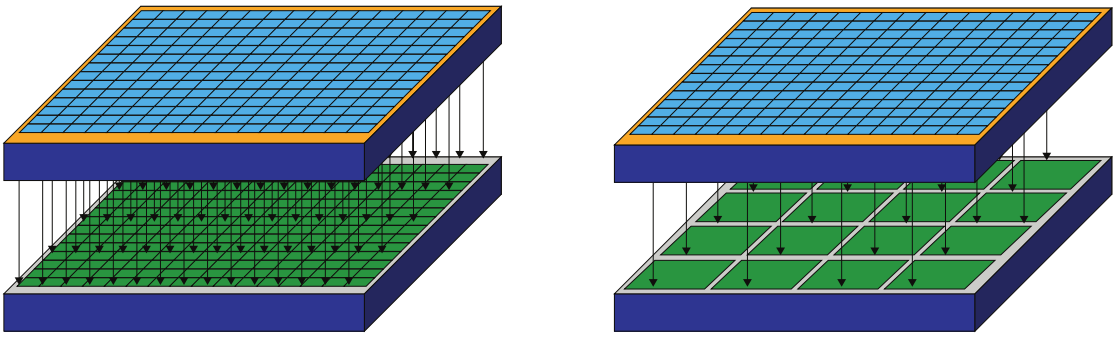


(a)



(b)

Figure 2.28: (*a*): MD-SiPM architecture proposed in [80] (*b*): MD-SiPM architecture proposed in [266].



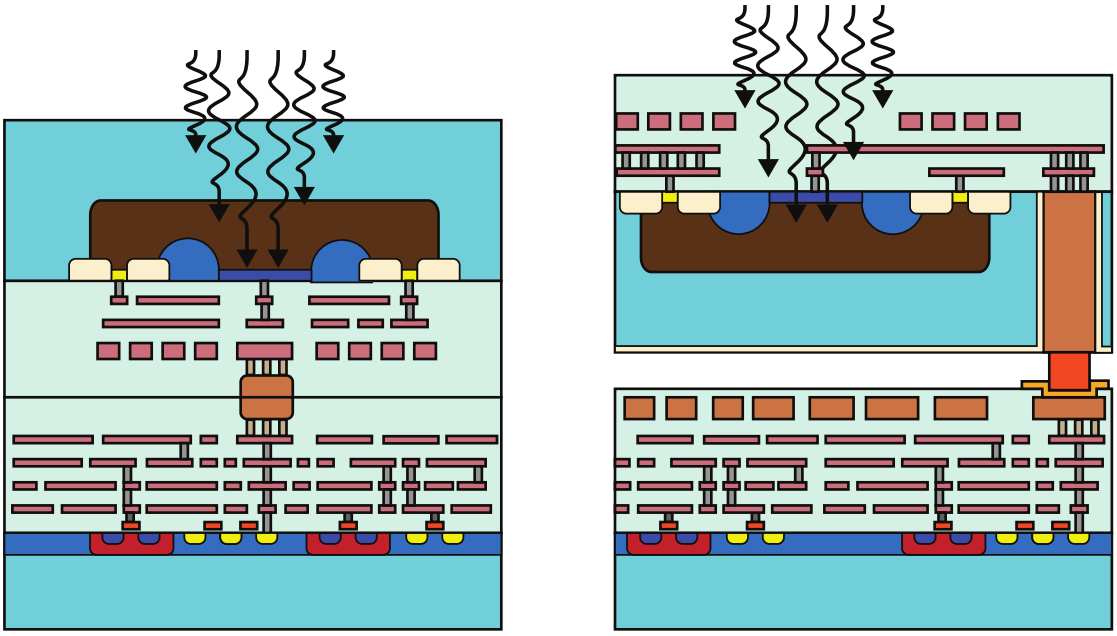
(a) (b)

Figure 2.29: (*a*): 3D per pixel connection approach. (*b*): 3D cluster matrix connection approach. The pixels are located on the top tier (*light blue*) while the electronics is embedded in the bottom tier (*green*).

along the data path using pulse generators. This technique reduces the possibilities that two pulses overlaps, mitigating the problems introduced by the spatial compression. So, it allows a reduction of the deadtime of the propagation tree and, as a consequence, a lower count loss at the output and a higher bandwidth of the tree itself [270]. an evolution of this techniques, based on a XOR-tree, was also presented more recently in [270]. For the sensor reported in [266], the array is segmented in multiple cluster, with a reduced number of SPAD pixels. Each cluster embeds a counting and a timestamping circuit. In this case, the array is organized uniformly, with the processing electronics distributed inside it and not at its periphery. For a more detailed description of this architecture, the interested reader can refer to [266]. Further development of the MD-SiPM architecture, together with an increased number of functionalities integrated on-chip and an increasing need for better performance, brought the designers to exploit using 3D integration techniques for their realization [35], [38], [39].

2.3.4 3D Integration

In the use of 2D monolithic integration architectures to implement SPAD-based arrays (see Section 2.3), many limitations can be highlighted. Indeed, 2D array architectures can be composed of large area SPADs but with limited space for electronic functionalities to boost the fill factor, or smaller SPADs with more electronics to add functionalities. This architecture presents a clear trade-off. Moreover, the fill factor of the overall structure can be improved by placing the large processing units at the periphery, making the photosensitive portion of the device as compact as possible [271]. In this case, the fill factor of the photosensitive area can be high. However, the overall fill factor of the detector is relatively low. This becomes a significant issue when the goal is to build a tile using multiple sensors, and influences the overall module efficiency. Usually, the target application and the required specifications drive the choice of the best compromise.



(a) (b)

Figure 2.30: (*a*): 3D BSI approach. (*b*): 3D FSI approach.

Nevertheless, one of the main advantages of using this kind of architecture (2D monolithic) is the possibility of adopting well-established technology processes for their design and fabrication. Furthermore, the use of standard technologies allows taking advantage of fast prototyping and reduced costs. Hence, this solution is a perfect candidate for all those applications in which the overall sensor fill factor and compactness of the detector is not a very stringent concern.

Recently, 3D-stacking technologies generated significant interest thanks to several demonstrations in commercial technologies which were able to overcome or mitigate the trade-off mentioned above. These architectures are generally composed of two layers (tiers) connected together, guaranteeing 3D heterogeneous vertical integration of a SPAD array over a CMOS chip. Indeed, one substantial advantage is that each tier can be independently optimized using dedicated processes (e.g., a CIS process for the top tier and an advanced low-power CMOS process for the bottom tier). Another potential advantage is represented by extremely low skews across large chips, thus enabling accurate timing over multi-megapixel SPAD image sensors. A further advantage is the potential trade-off solution between fill factor and functionalities, enabling the implementation of advanced functionality on-chip with negligible impact on the fill factor. Two main approaches can be used to implement 3D SPAD-based sensors, shown in Fig. 2.29. One option is to provide an independent 3D connection per pixel and connect each SPAD to the bottom tier, where the front-end and processing electronics are integrated (Fig. 2.29 (*a*)). This approach can be used to implement all the structures described so far in this chapter, just acting on the design of the circuit present in the bottom tier. The connection per pixel is currently the most used solution, and it allows a direct per pixel control

and the possibility to reach the highest image resolution and flexibility of the system. However, this solution implies realizing a high density of miniaturized 3D connections that can add complexity and cost to the system. Another possibility, recently proposed, is the clusterization of the pixels to be connected to a single 3D connection [272], [273]. This kind of approach stays in between the two SiPM families described in 2.3.3 and 2.3.3, creating a hybrid solution that tries to take advantage of both. SPADs are built on the top-tier and grouped as aSiPMs (but with few detectors) in this approach. Then, the mini-aSiPMs are connected to the front-end electronics (located in the bottom tier) through the implementation of a limited number of 3D connections [272], [273].
Similar to what was done in CMOS image sensors, two 3D integration approaches were proposed, namely the 3D FSI and the 3D BSI architectures.

Back-Side Illumination (BSI)

The first reported implementation of this architecture was shown in [156]. In this implementation, the SPADs are integrated on a top tier and operated using the BSI approach (see Section 2.1.4). The bottom tier embeds the electronic circuits from the SPAD front-end to the processing and readout units.
This kind of sensor is typically fabricated using two separate wafers: one for the so-called readout integrated circuit (ROIC) and the other housing the photodetectors. Hence, the two wafers are fabricated independently and often in different technologies and then combined in post-processing. The two tiers are stacked together, and the electrical connections are guaranteed by *hybrid bonding* [274], [275]. Moreover, to guarantee high detection efficiency (in particular in the infrared), the top tier is thinned down to thicknesses on the order of ~ 10 μm or less, to ensure that optical photons can penetrate the silicon substrate and interact in the proximity of the SPADs, where the collection efficiency is still high.
This approach was demonstrated to be very effective in implementing compact sensors aiming at applications where the wavelength range of interest is towards the NIR. Moreover, the use of hybrid bonding connections allows the implementation of a large number of miniaturized pixels [157], [159], [161].

Front-Side Illumination (FSI)

This approach differs from the one described before not only because, in this case, the SPADs are operated in FSI mode (see Section 2.1.4) but also for the 3D connection strategy used. Indeed, in contrast with the 3D BSI approach, in the FSI one, the integration is performed in a *face-to-back* fashion. The SPAD electrical contacts are located on the top-side of the top-tier wafer, while the electronics connections are placed on the other side. To ensure proper electrical connections, through silicon vias (TSVs) need to be used. A review of TSV technology can be found in [276]. As in the previous case, this approach enables the use of different technologies to fabricate the two tiers.
TSVs are not commonly available components in commercial technologies. In addition,

the implementation of structures as in Fig. 2.29 (*a*) requires a small size and pitch for the TSV connections. Together with their implementation in post-processing, these stringent specifications can increase risk, cost, and lead time.
The first example of such architecture was reported in [108], lately expanded in [277]. The FSI 3D approach will be discussed in more detail in Chapter 4.

3 Single Photon Avalanche Diodes in Standard Technologies

As a first step in the development of SPAD-based sensors for time-of-flight applications, we started with the design and implementation of SPADs and on-chip frontend circuits in standard technologies. The main purpose of this phase was linked to the need for optimization of sensitivity and timing performance, essential parameters of the main target application (TOF-PET).
Two standard technologies were taken as reference: 180 nm CMOS; 55 nm BCDLite. The choice of these two technologies was also driven by the intention to explore SPADs in deep sub-micron technologies and understand possible limitations in smaller nodes with respect to well-established technologies.
In advanced technologies, the higher relative doping levels and reduced annealing often cause an indeed much higher order defect concentration in SPADs, resulting in high noise and reduced sensitivity. Moreover, the front-end circuit substantially impacts the device's performance in terms of noise (i.e., afterpulsing probability), timing accuracy, and dead time (i.e., maximum achievable count rate and dynamic range).
The performance improvement is object of discussion in the following chapter, where we present a detailed description of the implemented SPADs and their experimental characterization.

The work presented in this section is based on [232], [278], [279].

3.1 SPADs in 180nm CMOS Technology

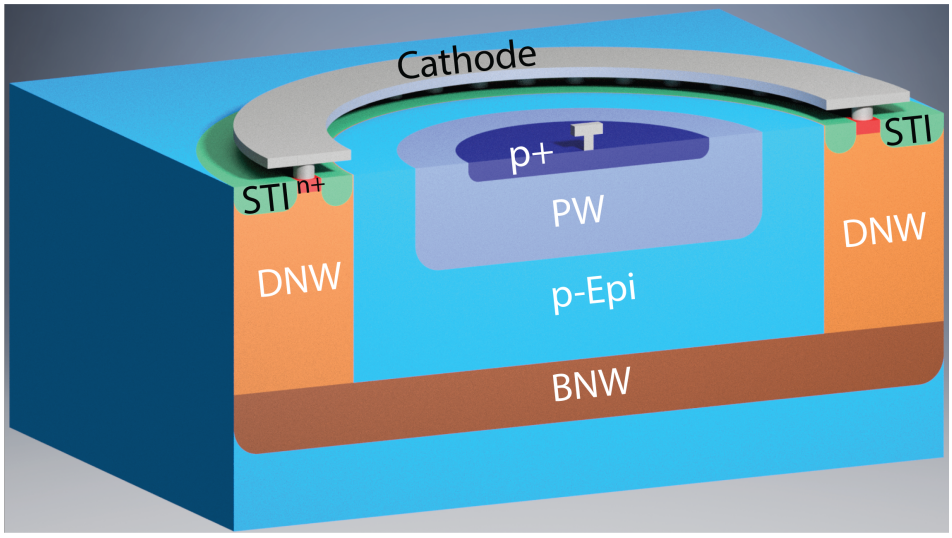
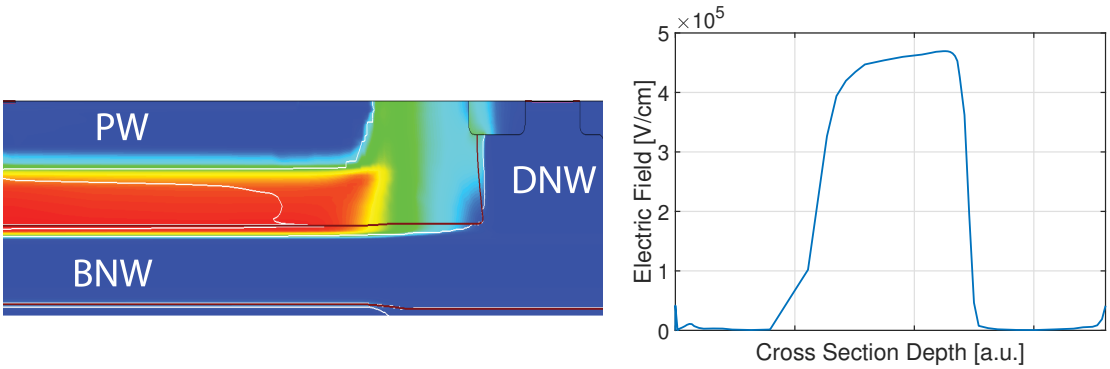


Figure 3.1: 180 nm CMOS SPAD cross-section. The SPAD presents a substrate-isolated structure, where the *buried n-well* (BNW) represents the cathode and the *p-well* (PW) is the anode. The high-voltage is brought to the cathode region thanks to a *deep-n-well* (DNW) contact [232].

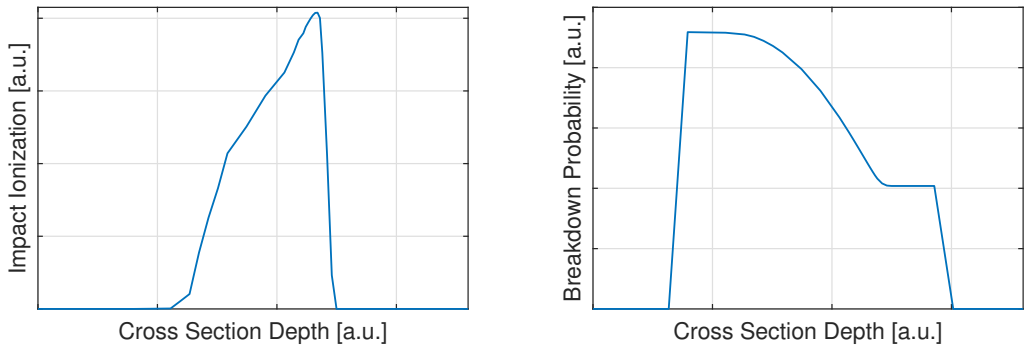
The first SPAD structure presented in this section is implemented in 180 nm CMOS technology.



(a) (b)

Figure 3.2: TCAD simulation of the 2D electric field distribution in the cross section (*a*). A quantification of the electric field along the vertical axis is also shown (*b*) [232].

This device's cross-section is based on a substrate-isolated type, where a *p-well* (PW) layer forms the anode of the SPAD and a *buried n-well* (BNW) layer creates the cathode contact. The latter is connected to the high voltage through a *deep-n-well* (DNW) (see Section 2.1.3 for more details). The SPAD presents a *p-i-n* structure, similar to [280]. Thus, a *p-Epi* layer between anode and cathode allows a relatively extended high-field region (Fig. 3.2, (*a*)). Fig. 3.1 shows the cross-section of the SPAD. In this SPAD structure, the DNW represents the periphery of the SPAD that provides the bias voltage to the BNW layer deep in the silicon. The guard ring has a virtual design (see Section 2.1.3) and employs a very low doping *p-Epi* region located between the PW and the DNW contact to smooth the doping transition and lower the electric field. TCAD simulations were performed to highlight the characteristics of the device.



(a) (b)

Figure 3.3: TCAD simulation of the impact ionization (*a*) and breakdown probability (*b*) along the cross section depth. The cross section depth corresponds to the one in Fig. 3.2 (*b*).

A representation of the 2D electric field distribution in the cross-section and its quantitative profile along the vertical axis is shown in Fig. 3.2. As highlighted by the simulation results, the device presents an area between anode and cathode, corresponding to the semi-intrinsic

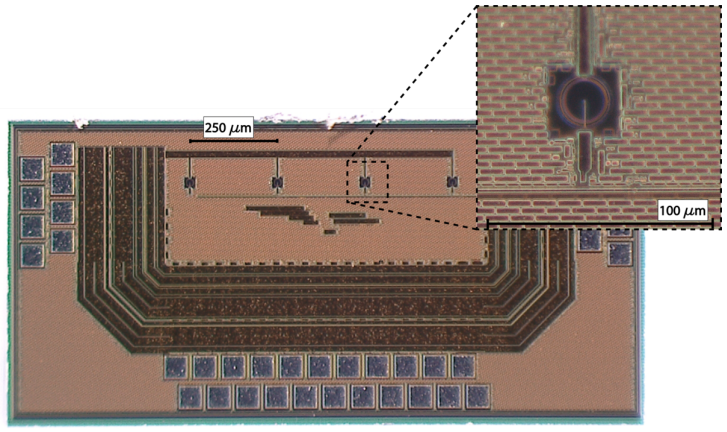
layer of the structure, where the electric field shows an almost constant profile above the critical electric field in silicon (about 3×10^5 V/cm). This structure guarantees a significant impact ionization and breakdown probability (Fig. 3.3) and increased sensitivity over a large wavelength spectrum. All the simulation results correspond to the SPAD operation at an excess bias voltage of 6 V.
The SPAD was implemented in three sizes: 25 μm, 50 μm, and 100 μm diameter (Fig. 3.4). This choice was driven by the interest in comparing the SPAD performance across a range of sizes. The three implementations are shown in Fig. 3.4. In these chips, the large number of pads is due to the need for several control voltages and power supply connections. The pixel circuit architecture will be analyzed in detail later in the chapter.

The SPAD breakdown voltage was measured at room temperature using the procedure explained in Section A.1, and it corresponds to a value of about 22 V. Fig. 3.5 shows the corresponding I-V curve under both dark and illuminated conditions. As seen in the graph in Fig 3.5, the breakdown voltage is overestimated in the dark. This behavior is a qualitative indication of the low noise level of the device, as discussed in Section A.1.

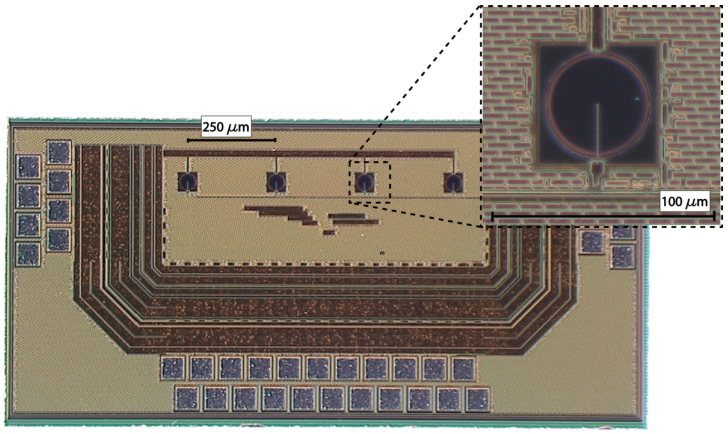
Next, we present the complete experimental characterization of the devices in terms of noise, sensitivity, and timing performance.

3.1.1 Chip Description

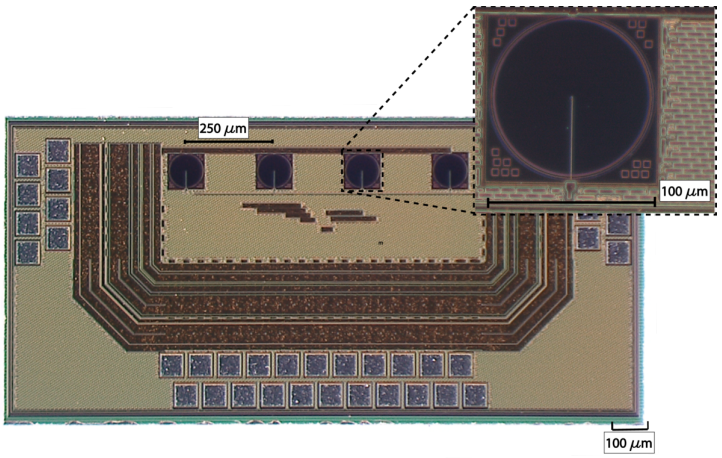
We designed three devices in separate dies. Each device comprises four independent SPADs with a dedicated pixel circuit, placed at a distance of 250 μm. Fig. 3.4 shows the micrograph of the three implementations, whereas the inset on the *top-right* of each micrograph shows three different SPAD diameters (25 μm, 50 μm, and 100 μm, respectively). In this implementation, each chip has a large pad ring to achieve the maximum controllability and observability of the system. For this reason, they present several bonding pads to ensure, in this prototyping phase, the possibility of fine-tuning the control voltages of the pixel circuit, together with the high voltage to bias the SPAD, digital VDD, ground, and ESD supplies. The SPAD front-end circuit, shown in Fig. 3.16, was inspired by [158], where the cascode transistor M_1 is used as a resistive divider, along with M_2 to enable high excess bias (up to 11 V) [281] in combination with thin-oxide MOS transistors in the remainder of the front-end.
The gate of M_1 is fixed at V_{cas}, supplied externally. When an avalanche is triggered in the SPAD, the voltage at the M_1 source rises, thus decreasing the transistor overdrive. When the voltage reaches $V_{cas} - V_{th,n}$ (with $V_{th,n}$ threshold of thick oxide nMOS transistors), M_1 turns off, boosting the impedance seen at the SPAD's anode. Thanks to the *body effect* acting on these transistors, the overdrive of M_1 is dynamically reduced, thus making it turn off faster.
Both passive and active recharge strategies are available in the pixel and can be used independently.
M_5, controlled by V_{pq}, is used to disable the passive quenching/recharge branch, represented by M_4. M_2 and M_3 form the active recharge branch, turned on by the feedback loop composed by the OR gate, Schmitt trigger, and tunable delay element. The feedback loop acts as a



(a)



(b)



(c)

Figure 3.4: Micrographs of the implemented chip embedding 25 μm (*a*), 50 μm (*b*), and 100 μm diameter SPAD. Each chip integrates four SPADs with an active pixel circuit [232].

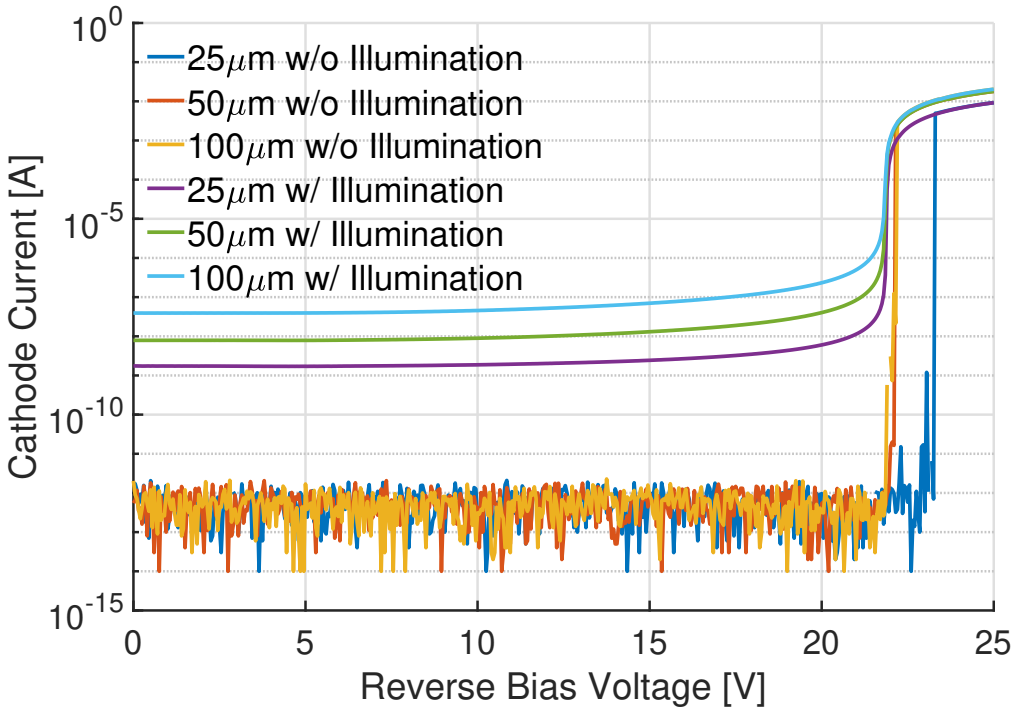


Figure 3.5: I-V curve measured on isolated SPAD samples with the same cross section and size. The measurements were performed both with and without illumination. The breakdown voltage is about 22 V. As can be seen in the measurements without illumination, the breakdown voltage results overestimated. This is an indication of the low noise of the device [232].

programmable-length monostable. The delay element is implemented using a current starved inverter (CSI) with a series voltage-controlled transistor for both pMOS and nMOS branches (Fig. 3.16 *right*). Controlling this delay and thus the hold-off time is essential to mitigate the afterpulsing, especially in relatively large SPADs. This mechanism determines both the pulse width at the output and, in large part, the dead time. An inverting Schmitt trigger was added to guarantee the stability of the monostable and to get sharp edges at the output signal. In addition, to improve the linearity of the CSI controls, a current mirror was included [282]. Finally, the output's slew rate was maximized, unlike in [281], using a custom buffering chain to the bonding pad. This solution ensured an output slew rate of approximately 1 V/ns.

3.1.2 Noise Performance

This section analyzes and discusses the noise performance of the presented devices. First, the DCR characterization is presented. Then, afterpulsing probability is extracted.

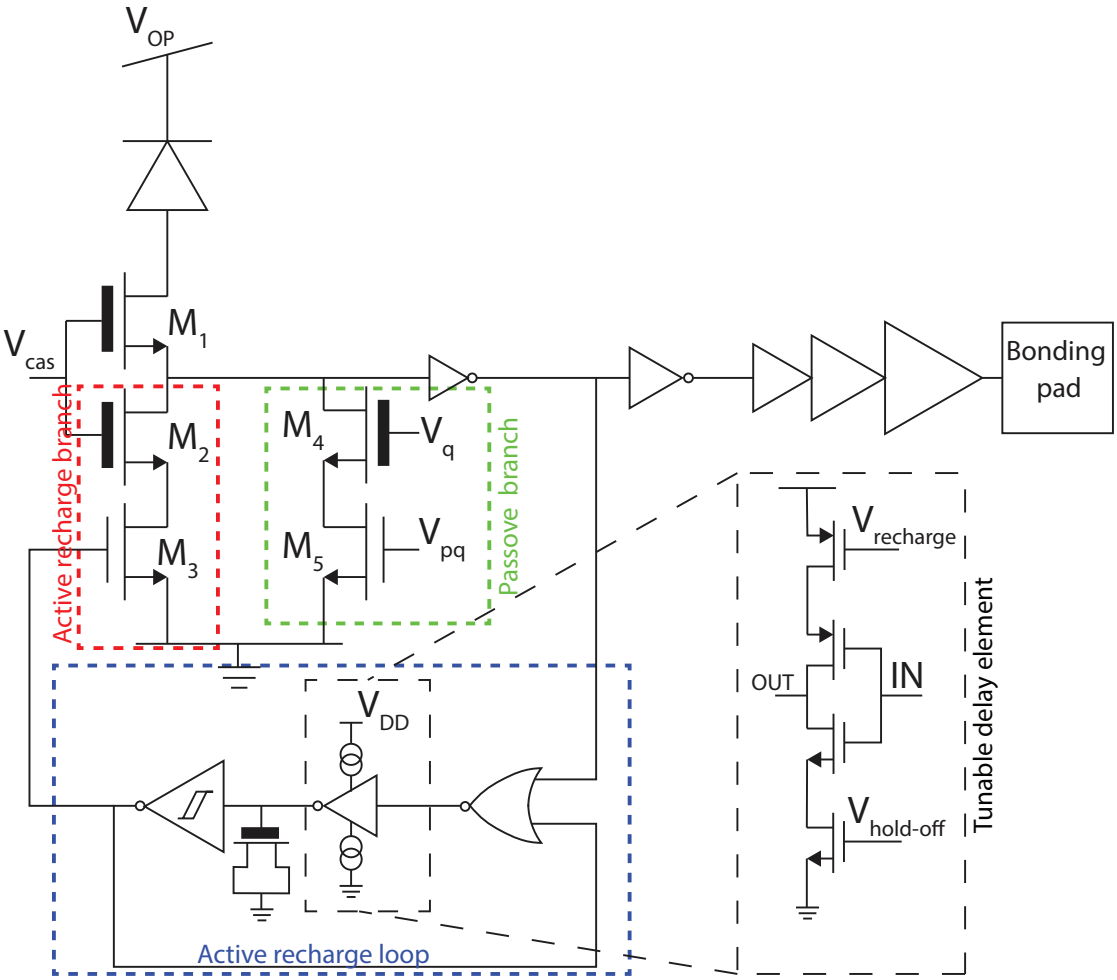


Figure 3.6: Simplified schematic of the SPAD pixel. A cascode structure is used to allow high excess bias voltage (up to 11 V). An active reset feedback guarantees the fine control of the SPAD deadtime. A buffering chain is implemented to output the signal on an analog pad.

Dark Count Rate

We measured the DCR at different excess bias voltages for each one of the three SPAD structures (Fig. 3.7). The measurement was performed at room temperature using an oscilloscope (Teledyne LeCroy WaveMaster 813 Zi-B) as explained in Section A.4. To a first approximation, the DCR is linear with the area of the active region. However, a super-linear behavior is generally observed in the normalized median DCR due to the increased probability of traps in larger SPADs, thus causing trap-assisted dark counts. Fig. 3.7 shows the results, where the median DCR over sixteen devices is 0.2 cps/μm^2 at 6 V excess bias and room temperature for the 25 μm diameter SPAD. As discussed in Section 2.1.5, the DCR consists of several contributions, and it usually shows a temperature dependency due to the SRH contribution. For this reason, DCR was also measured as a function of temperature in the -65°C to 40°C range using a climate chamber operated in a closed loop. Fig. 3.8 shows DCR in cps as a

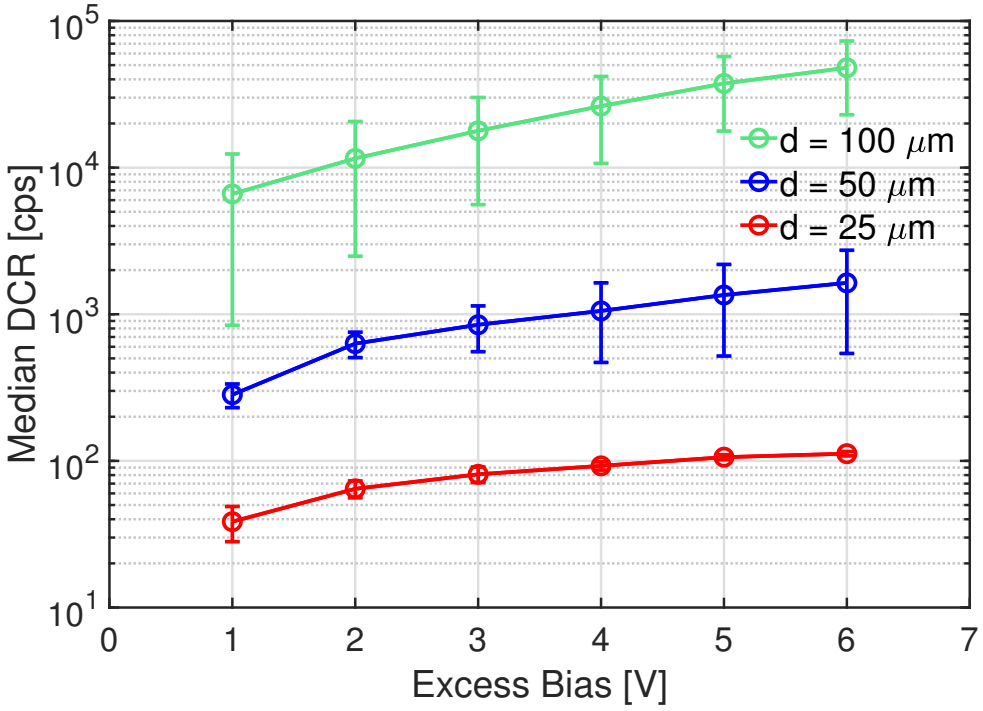


Figure 3.7: DCR median value measured at several excess bias voltages for different SPAD diameters (d). The results are obtained measuring 16 devices for each size at room temperature [232].

function of temperature for a range of excess bias voltages for the smaller (25 μm, (a), (c)) and the larger (100 μm, (b), (d)) SPADs. The figure also shows the breakdown voltage behavior over temperature for the same devices (e: 25 μm; f: 100 μm). These values of breakdown voltage were used to apply a precise excess bias level. By decreasing the temperature, the DCR decreases by about three orders of magnitude, reaching a value of 1.6 mcps/μm^2 at 6 V_{ex} for a diameter of 25 μm, operating at -65°C. The normalized DCR on the active area slightly increases to 4 mcps/μm^2 at 8 V_{ex} at -65°C.

From the graphs in Fig. 3.8 it is possible to observe the different trends of the curves. For example, by analyzing the charts (a) and (b), one can see how the 25 μm SPAD (a) noise shows a slope change at about -10°C. This change indicates that the main noise contribution is no longer given by SRH but by band-to-band tunneling.

On the other hand, the 100 μm SPAD shows a nearly constant slope also below -10°C. This trend indicates that in this case, the dominant noise source remains thermal generation. We can explain the dominance of thermal generation whereby considering the larger size of the SPAD and the higher number of dislocations that can be present in the device. The activation energy was also calculated to have more insight on the primary dark count contributions, using the Arrhenius equation:

$$DCR = A \cdot e^{-\frac{E_A}{kT}}, \tag{3.1}$$

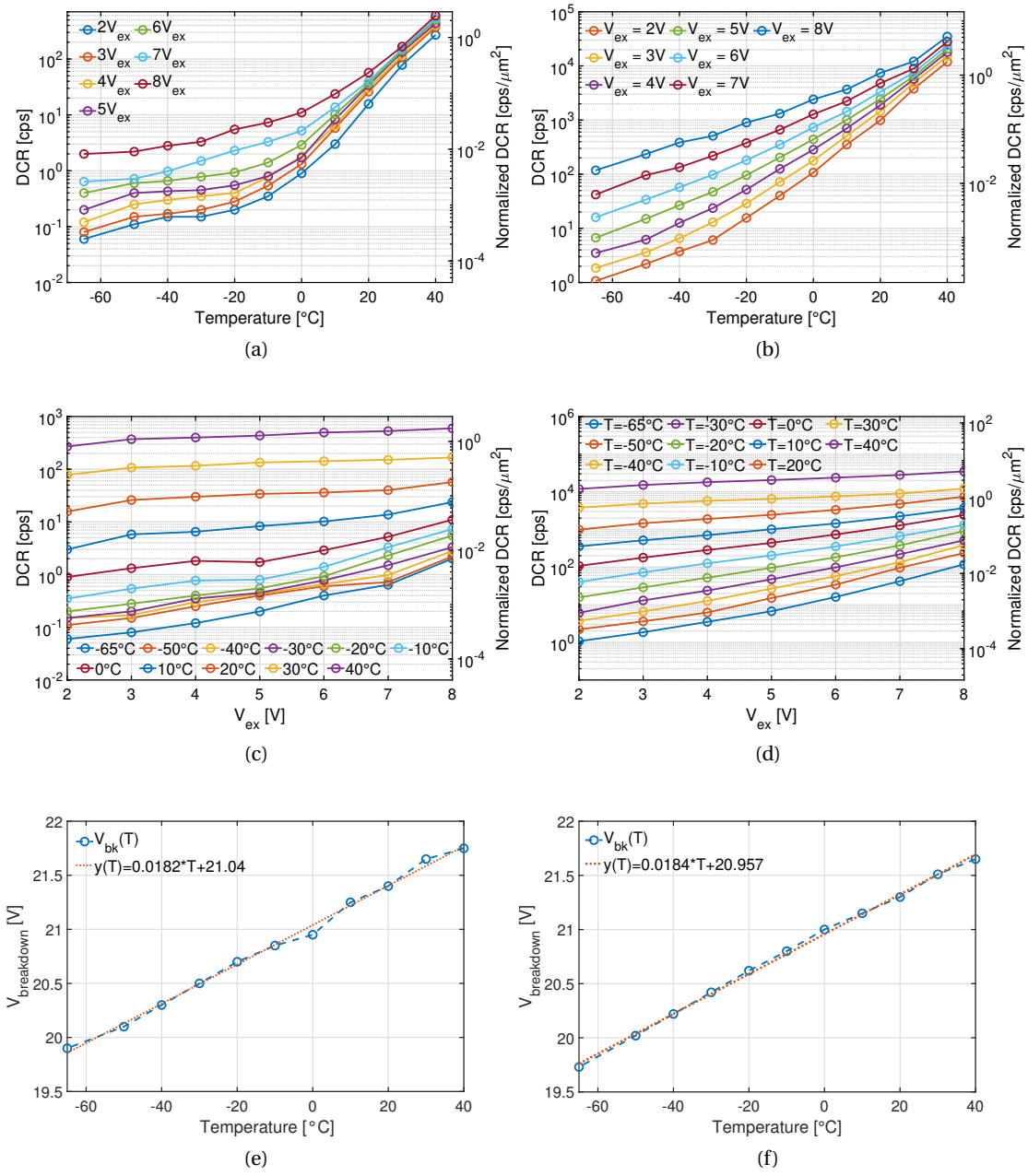
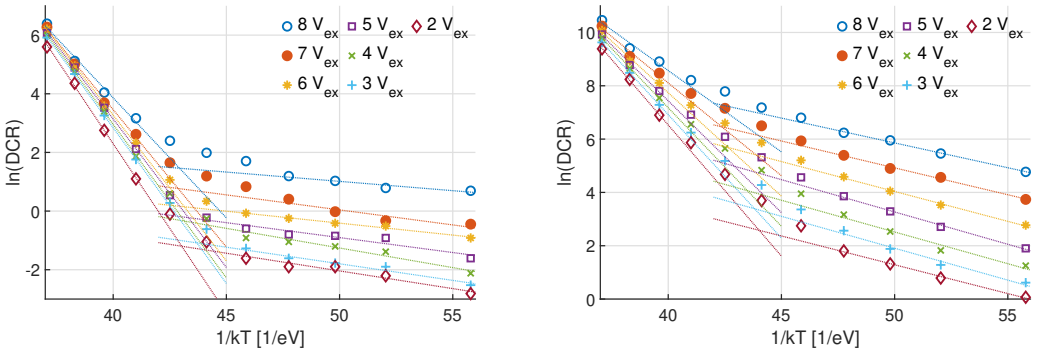


Figure 3.8: *Top* (*a*,*b*): DCR measured at different temperatures; *Middle* (*c*,*d*): DCR shown as a function of the excess bias voltage; *Bottom* (*e*,*f*): Breakdown voltage as a function of temperature. The results refer to a SPAD diameter of 25 μm (*left*) and 100 μm (*right*), respectively [232].

where k is the Boltzmann's constant, A is a pre-exponential factor, and T is the temperature in Kelvin [283], [284].

Fig. 3.9 shows the Arrhenius plots for the 25 μm (*a*) and 100 μm (*b*) SPADs at different excess bias voltages. In these plots, it is possible to see how the different trends change with



(a) (b)

Figure 3.9: Arrhenius plots of the 25 μm *(a)* and 100 μm *(b)* diameter.

Activation Energies				
	25 μm dia. SPAD		100 μm dia. SPAD	
V_{ex} [V]	$E_{a,1}$ [eV]	$E_{a,2}$ [eV]	$E_{a,1}$ [eV]	$E_{a,2}$ [eV]
2	1.120	0.121	0.983	0.216
3	1.061	0.113	0.926	0.239
4	1.047	0.134	0.889	0.238
5	1.012	0.102	0.838	0.242
6	0.994	0.081	0.776	0.224
7	0.937	0.102	0.696	0.202
8	0.818	0.063	0.610	0.186

Table 3.1: The activation energies extracted from the Arrhenius plot for the 25 μm and the 100 μm diameter SPADs, in the SHR-limited region and in the tunneling-limited region. The data are taken for several excess bias voltages.

the temperature. At higher temperatures, the fit line is steeper, indicating higher activation energy. This is due to the dominance of thermal generation. Conversely, while decreasing the temperature, we can observe that the activation energy decreases up to a value below $E_A = \frac{E_g}{2}$ (E_g = 1.12 eV), indicating the dominance of band-to-band tunneling [285]. We can also observe how the trend change in the data is more pronounced in the smaller SPAD. This behavior indicates, as mentioned earlier, that the thermal generation is affecting more the SPADs with a larger active area. This fact is also confirmed by the work published in [286].

Afterpulsing

The afterpulsing probability is another very important parameter (see Section 2.1.5), especially when one wants to minimize dead time through active recharge (Section 2.2.2), so as to increase the maximum count rate in SPADs while keeping pile-up in check. As explained in more detail in Section 2.1.5, this effect is due to some carriers generated during the avalanche process that may be captured by deep-level traps [174], [181], [185]. These carriers are then released after a statistical delay that depends on the lifetime of the traps [174], [181]. If a free

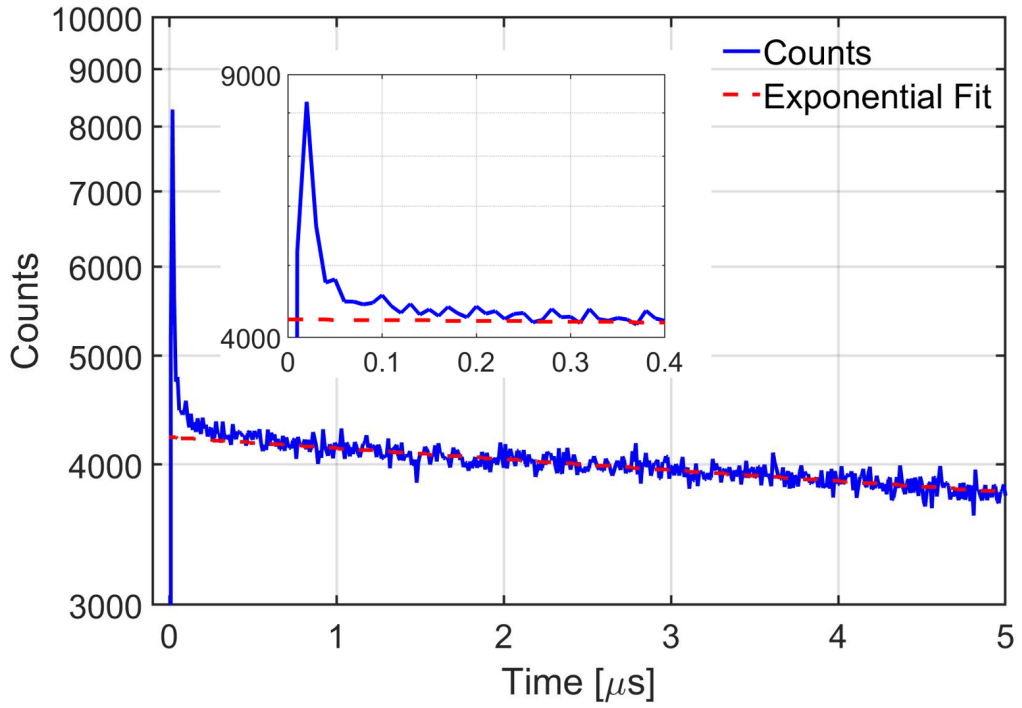


Figure 3.10: Inter-arrival time distribution measured with a pulse width of 11 ns on the SPAD of 25 μm size. The measurement is taken at a temperature of 25°C [232].

carrier is released in a region where the electric field is sufficiently high it can trigger another avalanche. In general, the probability that this event occurs is more frequent with short dead times.

The afterpulsing characterization for the presented SPAD devices is performed by histogramming the pulse inter-arrival time, as described in Section A.5, in this case, under dim and uniform illumination. Fig. 3.10 shows the measured inter-arrival time between pulses generated by the 25-μm SPAD at 6 V_{ex}. In this experiment, the SPAD dead time was set to about 11 ns using the integrated active recharge circuit described earlier.

In Fig. 3.11, the same experiment has been repeated for several control voltages of the active reset to obtain the afterpulsing probability as a function of pulse width. The plot shows the experimental results for the 25 μm (*blue*) and 100 μm (*red*) diameter SPAD. In the 25 μm SPAD, the afterpulsing probability remains as low as 0.1% for a pulse width of about 5 ns. With the current architecture, the minimum achievable SPAD dead time is 3 ns. On the other hand, the 100 μm SPAD shows a higher afterpulsing probability which does not exceed 2%. This higher value of afterpulsing probability for the 100 μm SPAD can find an explanation considering the higher probability of having traps located inside the larger SPAD active area.

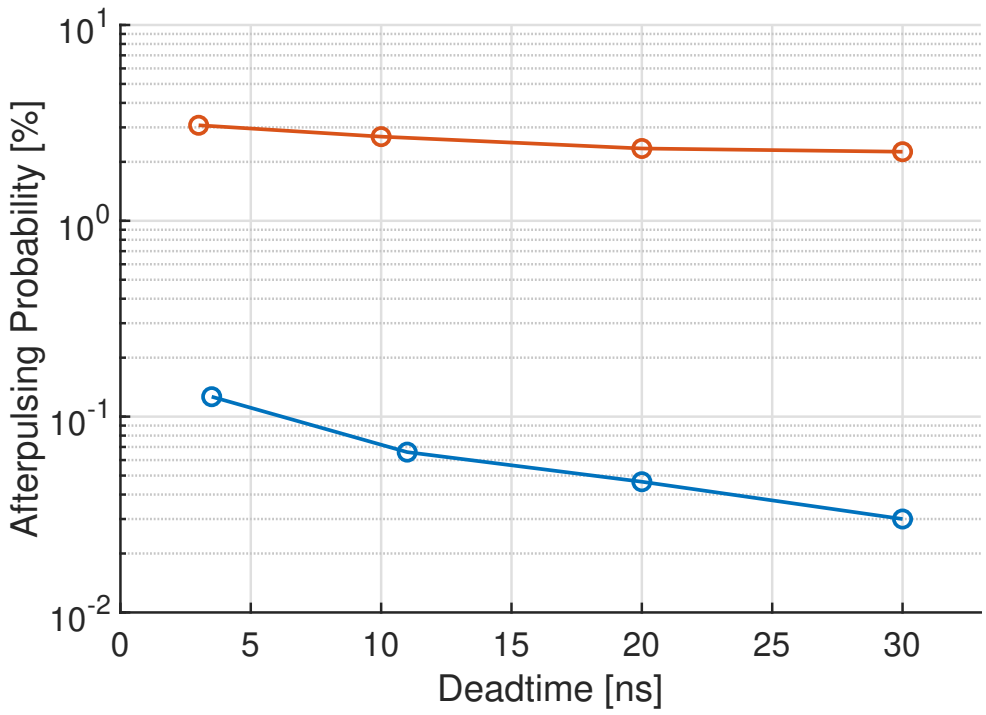


Figure 3.11: Afterpulsing probability as a function of the pulse width on the 25 μm diameter (*blue line*) and 100 μm diameter (*orange line*) SPADs at 6 V excess bias. All the measurements are taken at a temperature of 25°C [232].

3.1.3 Sensitivity

Section A.6 discussed two possible measurement protocols and their respective setups, typically used for such a characterization. For the device presented in this chapter, the continuous light technique was the methodology of choice. The measurement was performed for several excess bias voltages, from 1 V to 6 V, with a step of 1 V. In addition, the acquisition of an accurate sensitivity profile was enabled by a measurement step of 10 nm over a complete wavelength range that spans from 320 nm to 960 nm. Fig. 3.12 shows the results of this characterization. The PDP profile over wavelength (Fig. 3.12 *top*) highlights a relatively large sensitivity spectrum peaking at 480 nm with a value of 55%. Recalling what was discussed in Section A.6, we expect to see a saturation behavior in the PDP when increasing the excess bias voltage. This behavior is visible in Fig. 3.12 *bottom*. Indeed, here one can note how above 3 V_{ex} only marginal changes are observable in the PDP value. This behavior can be advantageous when building arrays with this device where this compression suppresses the impact on the variability of breakdown voltage due to local process variations.
Further, not only the variation of the breakdown voltage can create a nonuniform bias of the SPADs, but also, when the SPADs share the same high-voltage power line, an irregular and unbalanced power grid can cause voltage drops, with consequent reduction of the actual bias voltage applied to SPADs in some regions of the array. Therefore, a saturation effect of the PDP

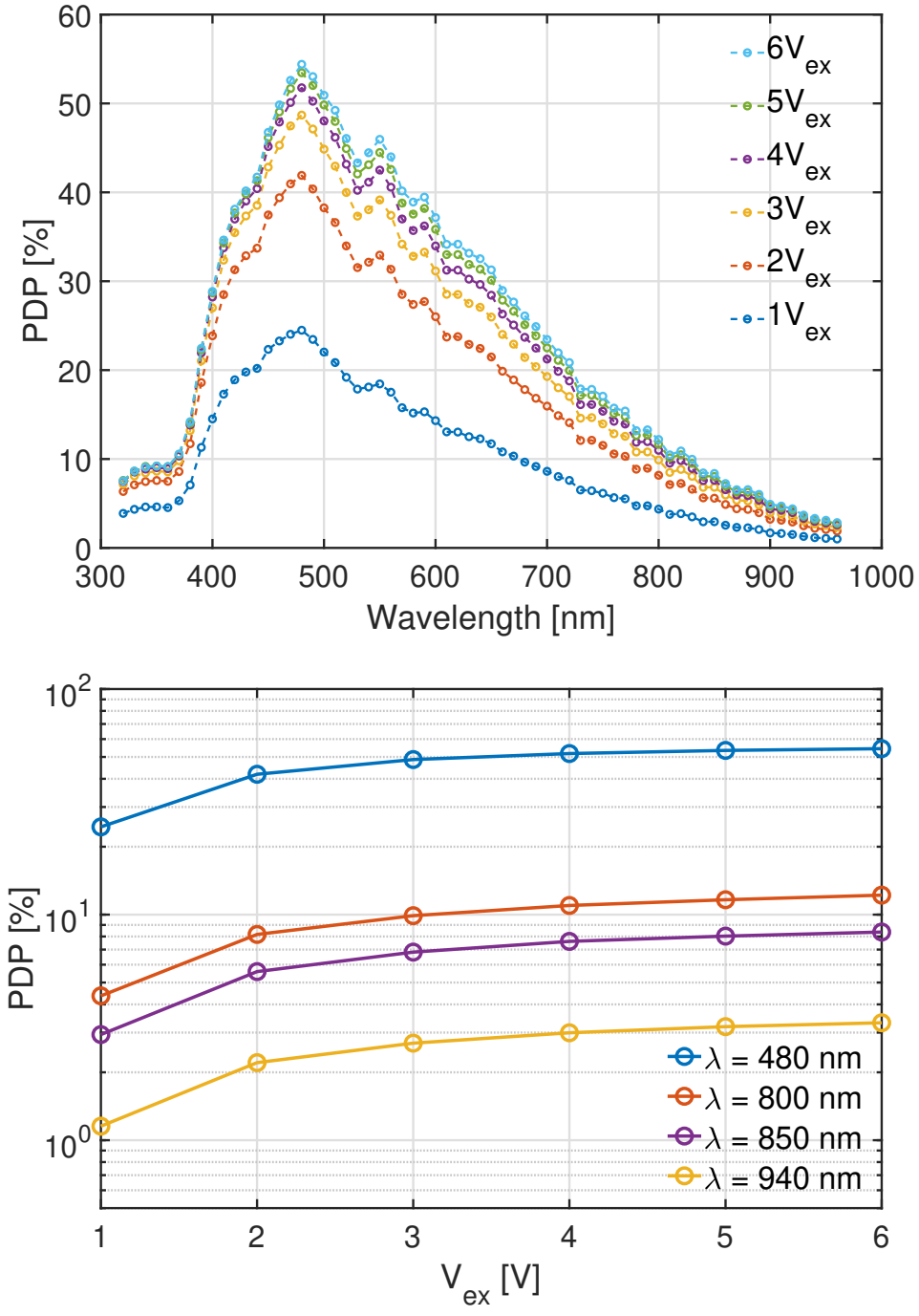


Figure 3.12: *Top*: PDP measurements for several excess bias voltages (from 1 V to 6 V excess bias). The measurement is performed with a wavelength step of 10 nm in a wavelength range between 320 nm and 960 nm. *Bottom*: saturation effect of PDP over bias voltage for four selected wavelengths [232].

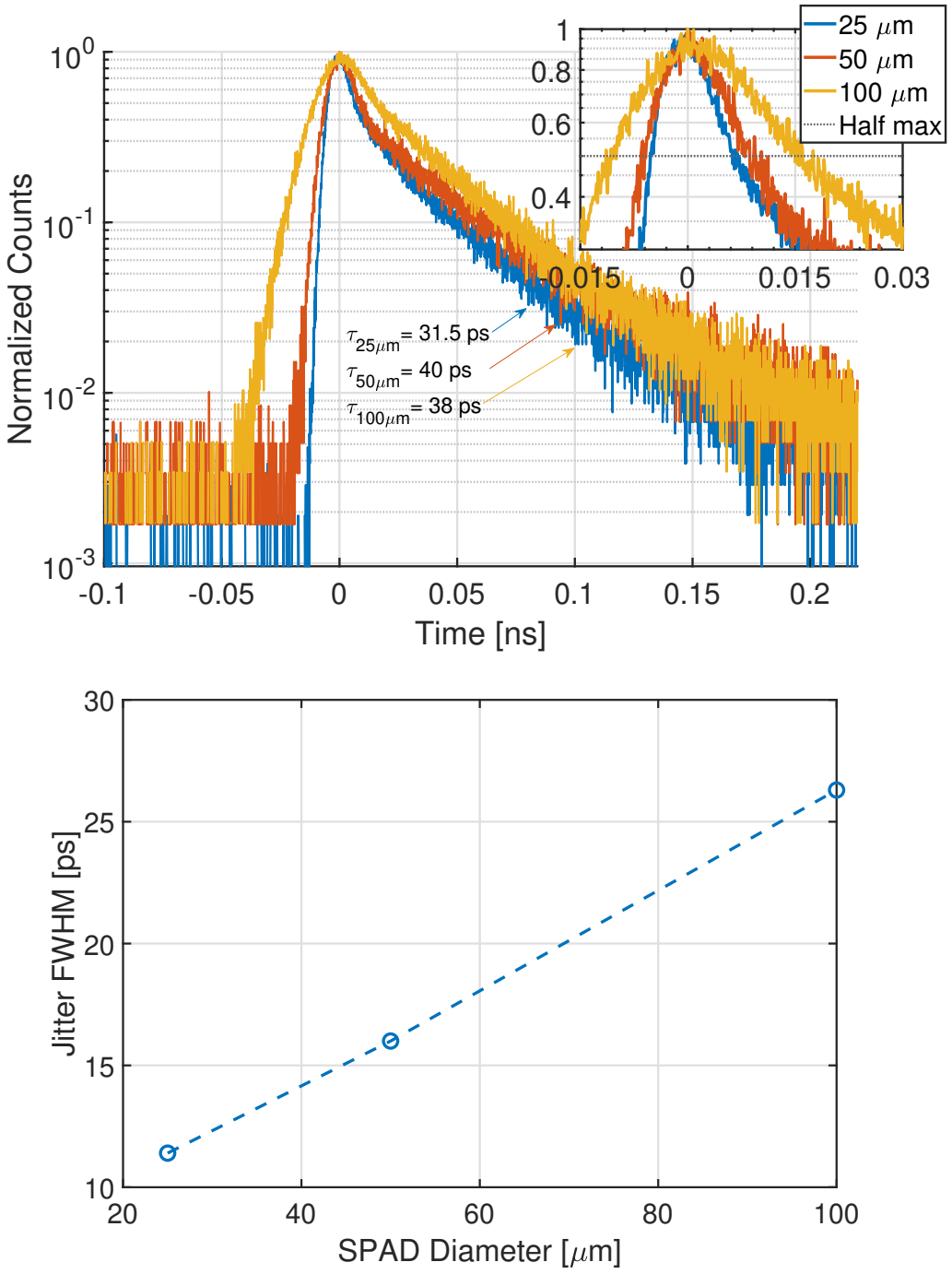


Figure 3.13: Jitter measurements performed with a 515 nm femtosecond pulsed laser. Histogram results (*top*) and extracted timing jitter as a function of the area (*bottom*) are seen in the graphs. The timing jitter is around 12.1 ps, 16.0 ps and 27.2 ps FWHM for the 25 μm, 50 μm, and 100 μm diameter SPADs, respectively. The measurement is performed at V_{ex} = 6V excess bias for the three samples [232].

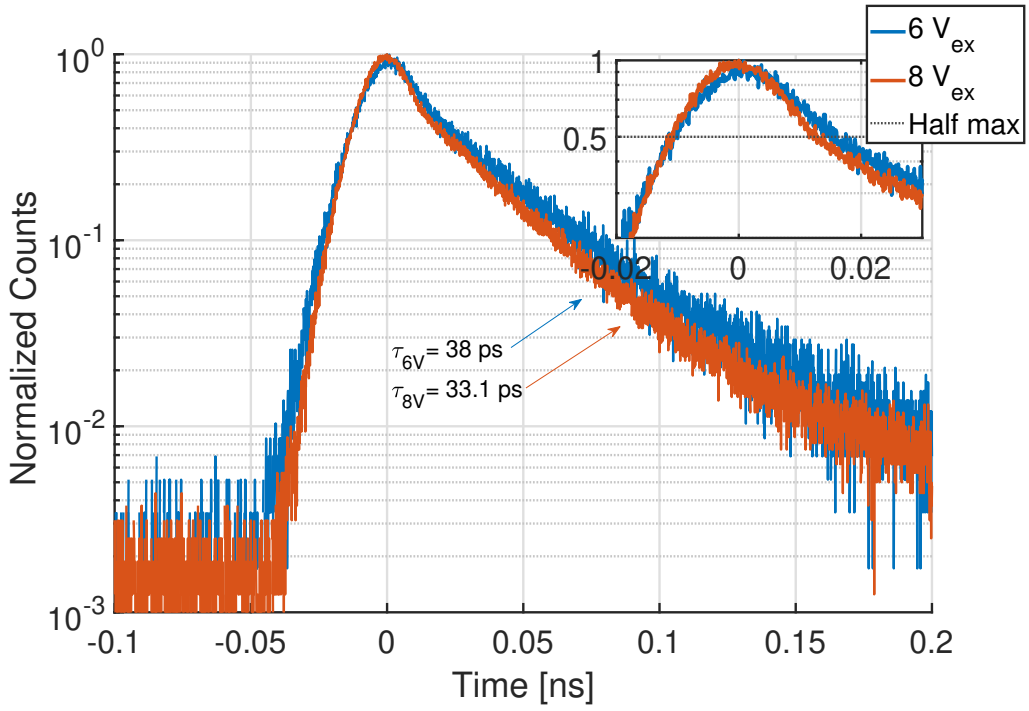


Figure 3.14: Timing jitter measurement comparison at two excess bias voltages (6 V and 8 V) performed on the 100 μm SPAD. The jitter is 27.2 ps and 23.5 ps FWHM, respectively [232].

above a specific excess bias voltage is desirable to overcome (or mitigate) this issue.

3.1.4 Timing Performance

The time resolution of a SPAD, often named timing jitter (or SPTR), is another essential parameter to determine device performance. Several factors are influencing the timing jitter of a device. During the characterization, the typical variable considered is the excess bias voltage applied to the device. As mentioned in Section 2.1.7, the voltage applied across the SPAD will influence the electric field distribution and the dynamics of the avalanche, ultimately affecting the timing jitter of the device. Another critical parameter that influences the overall IRF is the wavelength of the light source used during this experiment. Indeed, different wavelengths can determine quite different results when measuring the timing jitter.
As discussed in Section A.7, the experimental protocol is based on time-correlated measurements, comparing the response of the SPAD with the one of a reference signal synchronous to a pulsed laser. In the experimental characterization of the presented devices, the setup used comprises a femtosecond laser (Amplitude Systèmes SA, S-Pulse HR SP), capable of generating 150 fs pulses at a wavelength of 1030 nm and 515 nm after second-harmonic generation (SHG). A fast photodiode (Newport InGaAs Photodetector, 45 GHz bandwidth) is used as a timing reference, while neutral density filters (NDFs) attenuate the upconverted beam to achieve a detection regime as close as possible to single-photon. The DUT has a high-impedance

output, which is captured by an active probe. In addition, an oscilloscope (LeCroy WaveMaster 813 Zi-B) is used to record both the waveform from the DUT and the reference PD [232].
Fig. 3.13 *top* shows the IRFs resulting from the timing jitter measurements for the three device sizes. The plot shows the histograms of the response of the SPADs when biased at an excess bias voltage of 6 V and room temperature. The oscilloscope trigger threshold was set at 400 mV for the SPAD pulse and 300 mV for the PD. The laser repetition rate is 100 MHz, and the light was reduced to detect less than a laser pulse every 100, on average. The response distribution's jitter value (FWHM) was measured at 12.1 ps for a diameter of 25 μm, 16.0 ps for 50 μm, and 27.2 ps for 100 μm. The full width at a tenth of maximum (FWTM) was also extracted to capture the diffusion tails; it results in 55.7 ps for a diameter of 25 μm, 66.8 ps for 50 μm, and 91.7 ps for 100 μm. The exponential time constant for the diffusion tails was also extracted from the plot to be 31.5 ps, 40 ps, and 38 ps for the 25 μm, 50 μm, and 100 μm SPAD, respectively. The bottom part of Fig. 3.13 shows the trend of jitter performance over the device size, indicating an almost linear behavior. The plots in Fig. 3.13 show a 100 μm SPAD response with two excess bias voltages of 6 V and 8 V, improving the jitter from 27.2 to 23.5 ps FWHM. In this case as well, the exponential time constant of the diffusion tail was extracted, and it is 38 ps for 6 V excess bias and 33.1 ps for 8 V excess bias.
The results presented here were obtained with the pixel circuit described in Fig. 3.16, integrated with the SPAD. This circuit allowed a deadtime on the order of 3 ns, compatible with the laser's high repetition rate. Moreover, it should be noted that the circuit does not embed low threshold comparators or other components that would cause a high static power consumption. Thus, these results demonstrate how a simplified circuit can also achieve excellent timing performance and be used as the SPAD front-end, thereby ensuring scalability in large arrays of pixels.

3.1.5 System Performance Optimization

Although the results obtained with the presented detector were already outstanding, we decided to optimize the readout PCB system to show the influence of noise and external readout on the performance. For this purpose, we implemented the PCB system shown in Fig 3.15. This system comprises a motherboard, where all needed voltage levels are derived from a single 5 V power supply. In addition, a power management unit was designed to filter most of the electronic noise and provide low-noise and stable voltage levels to the detector and its front-end circuits. Also, this system was fully integrated into a single system-on-board to reduce the noise picked up in cables and power cords that can act as antennas. Indeed, reducing the noise in the system is essential when the target timing precision approaches 10 ps timing resolution [287]. The system also embeds a serial interface that allows for remote control using a host computer. This feature turns out to be very useful in an experimental environment like the one that will be described in Chapter 6.
The output of the chip is connected to fast SiGe comparators (Analog Device ADCMP572) that drive 50 Ω lines (Fig. 3.16). This solution reduces the capacitive load at the chip's output

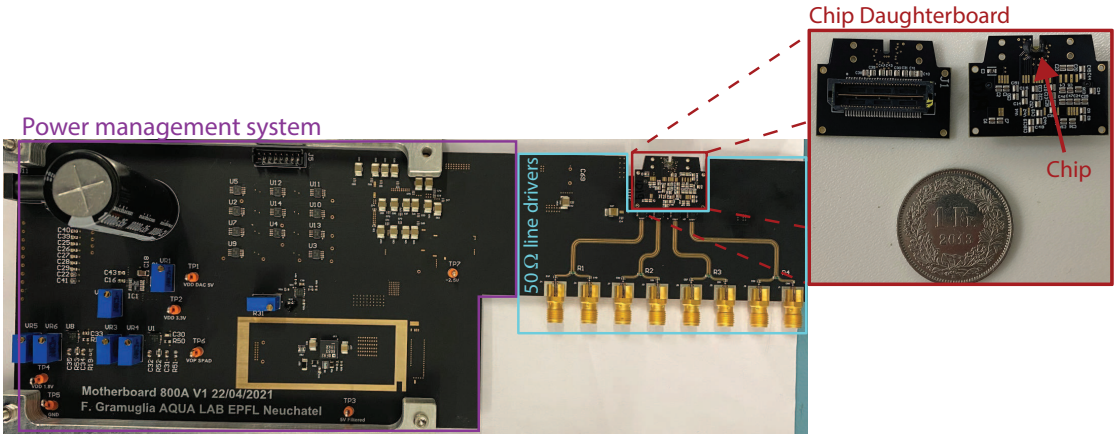


Figure 3.15: Optimized system-on-board setup. This system was designed to maximize the timing performance of the 180 nm CMOS SPAD-based sensor designed in this thesis. The system is divided in several parts: a power connection provides the 5 V supply to the system; a power management system derives all the required voltages from the 5 V supply, and can be controlled with a SPI bus interface from a host computer; fast SiGe comparators are used to read the chip signal and propagate it into differential 50 Ω lines; The SPAD detector is located on a separate chip daughterboard to allow an easy device change.

(high impedance node) and helps propagate the signal through a high-frequency cable to the timestamping electronics. In addition, the use of these comparators makes it possible to achieve a high signal slew rate (≥ 1.6 V/ns).

System Characterization

In order to analyze the performance of our system, we performed an optical characterization using a femtosecond pulsed laser setup as discussed in Section 2.1.7. The measurement was performed using two different wavelengths: 515 nm and 780 nm. We decided to use two wavelengths to compare the device's performance changing the mean free path of the optical photons in the device. For this test we used the 25 μm diameter SPAD-based device.

The experiment has been repeated for several excess bias voltages and two wavelengths, and the results are shown in Fig. 3.17. The timing results are expressed as FWHM of the IRF. The timing precision shows a dependency on the bias point, and it improves when increasing the applied voltage, as expected. With this improved system, we reached a timing jitter of 7.5 ps at a reverse bias voltage of 28 V, corresponding to an excess bias (V_{ex}) of about 6.5 V. Moreover, in Fig. 3.17 (*b*), we report the decay time constant of the exponential tail again as a function of reverse bias voltage. As we can see from the results reported in Fig. 3.17, the performance observed with the two wavelengths are very comparable. This similarity can be explained by the large high-field region present in the device (see the beginning of this chapter for more details). These results show an improvement of almost 40% with respect to what

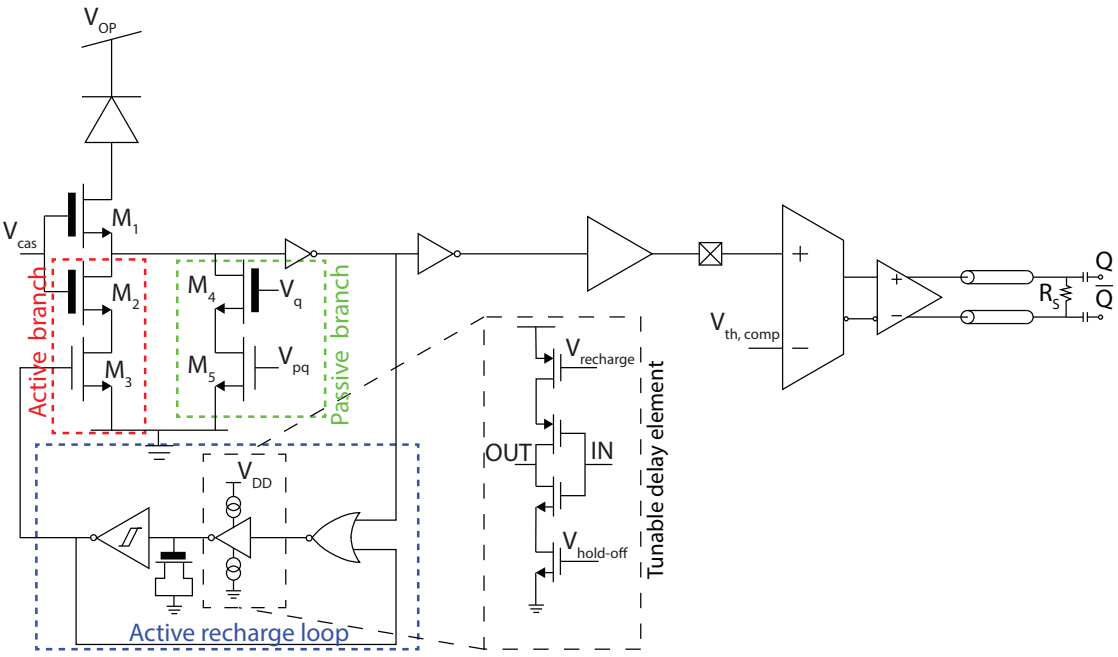
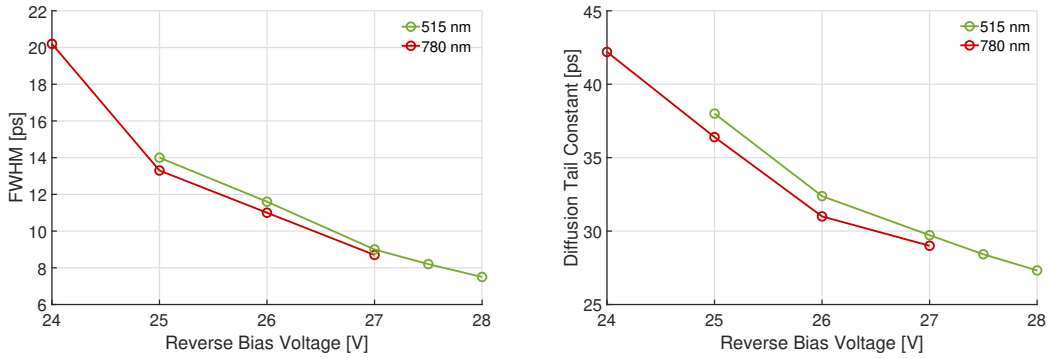


Figure 3.16: Device schematic. The chip output is connected to a fast SiGe comparator to drive a 50 Ω differential output line [278].

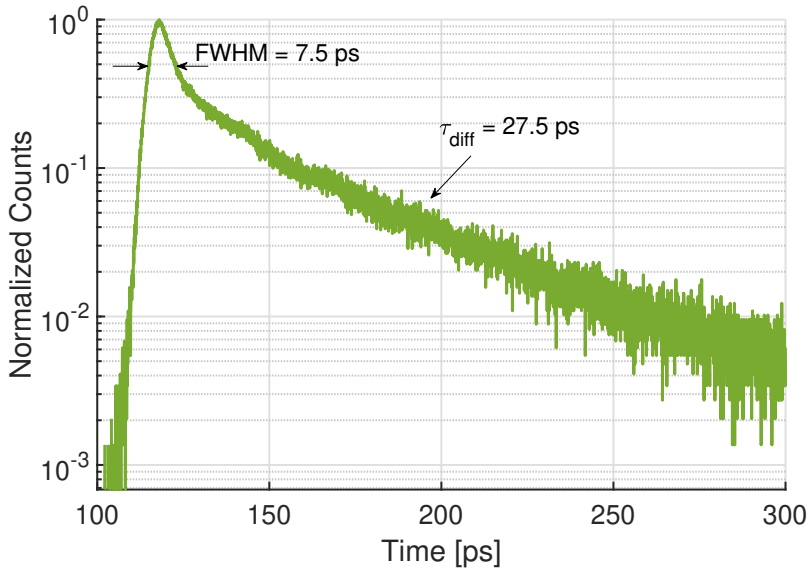
reported in the previous section [232], where the output was directly taken from the packaged die with high impedance 4 GHz active probes. This results shows how important is the proper implementation of a readout system for SPAD-based devices. Indeed, the reduction of noise and parasitic load can directly improve the timing jitter performance of the device.

3.2 SPADs in 55nm BCD Technology

In this section, we focus on discussing SPAD implementations in 55 nm BCDLite technology. Several reasons motivated this work. For example, the increased need for performance of SPAD-based detectors pushes the designer toward more advanced technology nodes. In addition, technology scaling enables higher complexity and more functionalities. However, deep sub-micron technologies do usually feature higher relative doping levels and reduced annealing phases that can compromise the performance on the detector side. Indeed, these features can typically cause a higher-order defect concentration in the SPAD, which usually leads to higher noise and modest sensitivity. Although, as discussed in the Section 2.3.4, 3D strategies exist which can allow heterogeneous integration and the decoupling of the technology used for the bottom tier logic from the one used for the top tier SPAD design, this approach is not wide spread yet, and it adds non-negligible costs, risks, and lead time to the sensor implementation.
Moreover, Bipolar-CMOS-DMOS (BCD) technology, introduced for the first time in 1985 by STMicroelectronics to fulfill the application demand for high voltage on-chip together with

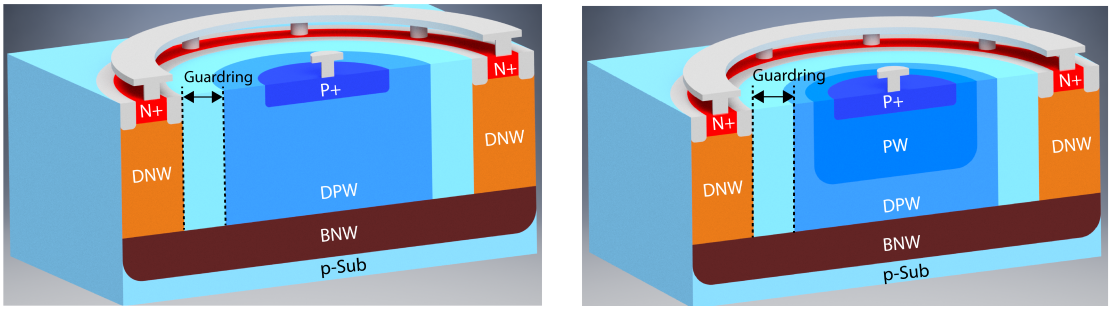


(a) (b)



(c)

Figure 3.17: (*a*) FWHM single-photon timing resolution obtained with an optimized system. (*b*) Exponential tail time constant measured at various excess bias voltages. For these experiments two wavelengths were used: 515 nm and 780 nm. The results show a jitter of ~7.5 ps FWHM for green and ~8.5 ps FWHM for red light at ~6.5 V and ~5.5 V of excess bias voltage, respectively. (*c*): IRF obtained using a 515 nm femtosecond laser and biasing the device at 6.5 V excess bias [278].



(a) (b)

Figure 3.18: Cross-sections of proposed 55nm BCD SPADs. A single implant (PW) highlights the difference between structures, named SPAD1 (*a*) and SPAD2 (*b*) [279].

fast switching speeds, implements additional deep and lower-doped layers, unlike standard CMOS nodes. The use of this kind of technology for SPAD implementations has been demonstrated in the 160 nm ST BCD technology [144], showing excellent device performance in terms of timing, noise, and sensitivity. Nevertheless, in this work, we wanted to push the limits of the SPAD design by scaling to a much more advanced node such as the 55 nm BCDLite.
For all these reasons, there is a particular interest in developing SPAD devices in deep submicron technologies, or even in sub-65 μm technology nodes, that show performance comparable to the state-of-the-art.

The present section shows the devices developed in the 55 nm BCDLite technology node. First, we focus on substrate-isolated SPADs based on deep junction structures as well as shallow-junction structures, focusing on their characterization.

3.2.1 Deep Junction Substrate-Isolated SPADs

In this section, we discuss the deep junction substrate-isolated devices named SPAD1 and SPAD2 (Fig. 3.18).
As discussed in Section 2.1.3, substrate isolation is desirable, especially when SPADs are integrated using the 2D approach, side-by-side with the electronics. On the other hand, the definition of deep junction comes from the actual position of the primary SPAD junction along with the cross-section depth. Considering the light absorption length of the silicon, the thickness of the n-well region contributes to the diffusion tail. At the same time, the overall resistivity of the path to the cathode contact and avalanche spreading dynamics determine the FWHM of the timing jitter [194], [288] (see Section 2.1.7 for a more detailed description). Moreover, the buried n-well (BNW), designed with retrograde doping, helps design the multiplication and space-charge regions critical for low-noise operation [289]. In addition, the width of the space charge region also influences the sensitivity spectrum of the device.
Fig. 3.18 shows the cross-sections of the two structures we presented here. The two structures

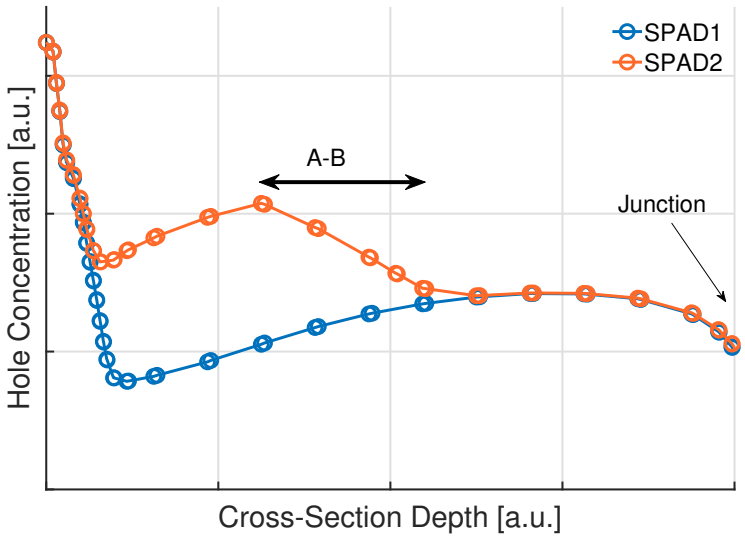


Figure 3.19: Comparison showing a decrease of hole concentration for SPAD2 in contrast with an increase for SPAD1 inside the A-B interval, corresponding to the PW region. The simulation results are provided at 5 V_{EX} [279].

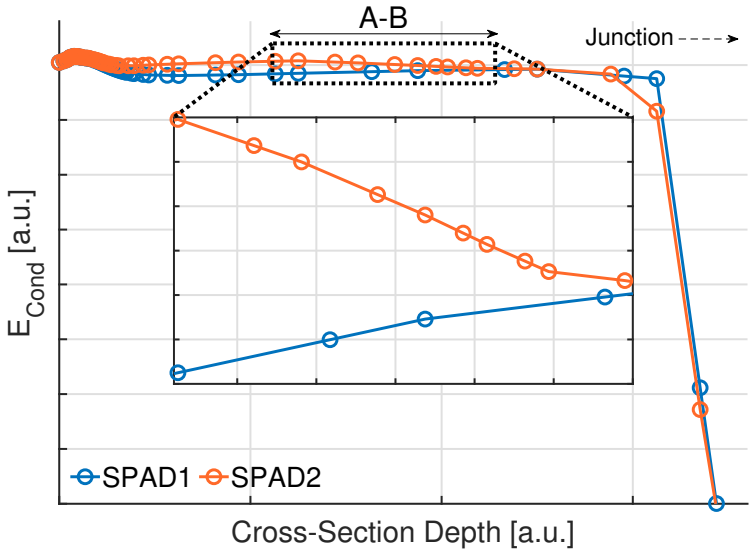


Figure 3.20: Simulated conduction band diagram of SPADs. Inset illustrates the drift barrier resulting from the net concentration difference between the proposed SPAD1 and SPAD2 [279].

are composed of a *p+* anode contact at the center of the SPAD and a deep N layer (*BNW*) that is contacted to the surface through a *DNW* and an *n+* to form the cathode of the diode. The main difference between the two devices is in the structure of the quasi-neutral (photo-collector)

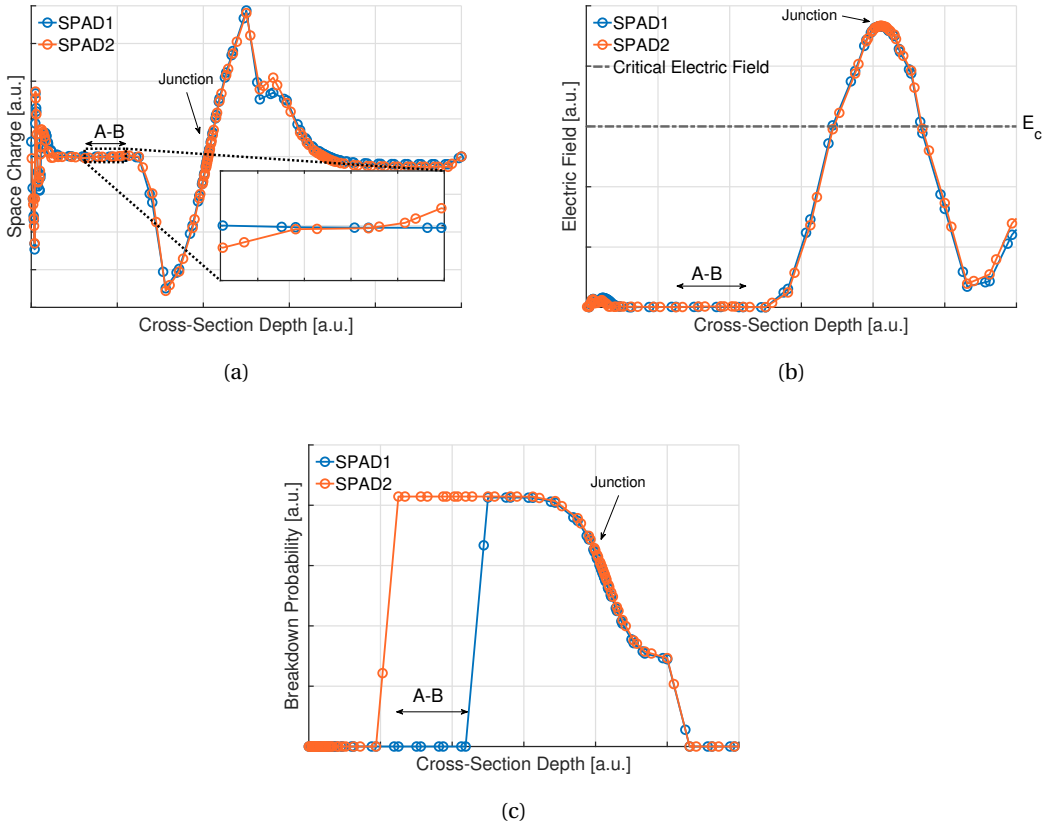
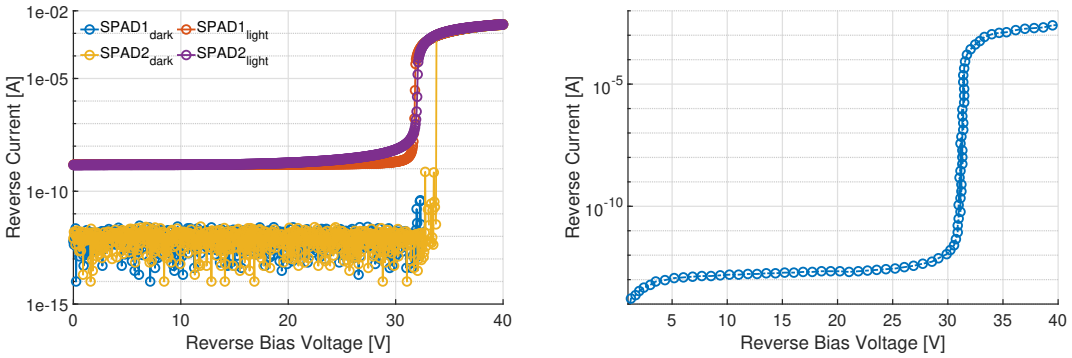


Figure 3.21: Comparison between the simulated junction parameters of SPAD1 and SPAD2 at breakdown. Space charge and E-field highlight nearly identical junction parameters. Normalized values are shown at 5 V_{EX}. *(a)*: Space charge illustrating nearly identical multiplication regions. *(b)*: Electric field magnitude, $|E|$, showing that the depth range of interest (A-B) is outside the multiplication region. *(c)*: Simulated breakdown probability for each SPAD, which results in a substantial difference in performance between the two devices [279].

region. SPAD2 presents an additional *PW* implant in the photo-collector region that, as we will see later, increases the structure's sensitivity.

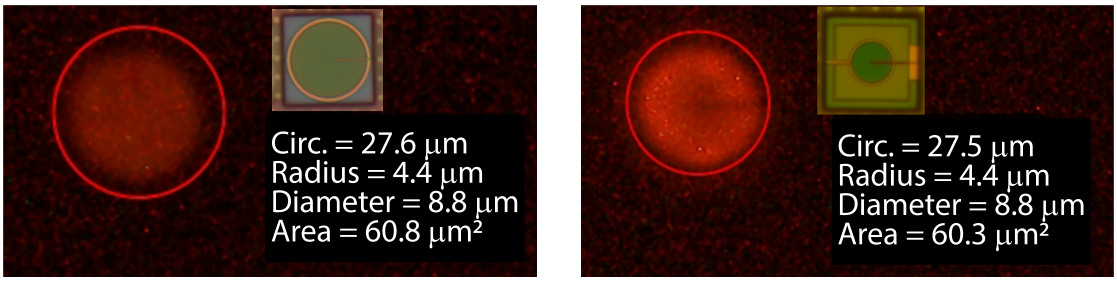
Device Structure Analysis

Simulation analysis can be very useful in understanding the behavior of the devices which we design. Here we performed a TCAD analysis on the two different cross-sections. Let us start by looking at the majority carrier (holes) concentration in a region close to the surface, corresponding to the transition between p+ and the photo-collector region for the two SPADs (Fig. 3.19). Here, we can note how the hole concentration monotonically decreases towards the junction for a certain depth in SPAD2. Despite this, in SPAD1, the hole concentration increases. Interestingly, this depth range (A-B) is outside the high-field multiplication region (Fig. 3.21, (*a*), (*b*), (*c*)). Consequently, as visible in Fig. 3.19, an energy barrier is created, inhibiting the



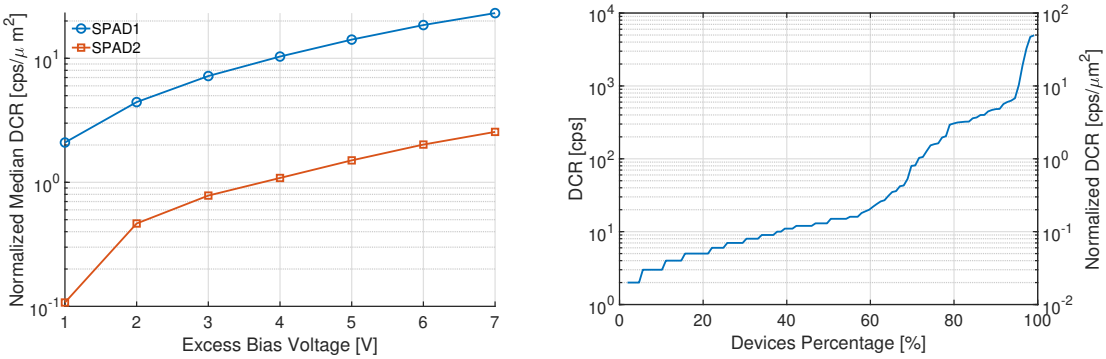
(a) (b)

Figure 3.22: SPAD1 and SPAD2 I-V curves measured in light and dark conditions. The breakdown voltage of both structures is of about 32 V (*a*). The I-V curve was also simulated, showing good agreement with the experimental results; (*b*) shows the simulated I-V curve for SPAD2 [279].



(a) (b)

Figure 3.23: SPAD1 and SPAD2 light emission test (LET) [279].



(a) (b)

Figure 3.24: Normalized median DCR for SPAD1 and SPAD2 (*a*), over ten devices. (*b*) shows the noise population distribution obtained with 108 samples of SPAD2 at 3 V excess bias [279].

photoelectrons generated close to the surface from diffusing toward the multiplication region in SPAD1. On the contrary, in SPAD2, the photo-collector region is much broader. Indeed, in this case, it allows the photoelectrons generated in the quasi-neutral region to transit to the multiplication region. This behavior is also explained by a significant difference in breakdown probability between the two SPAD structures. Fig. 3.21, (*c*) shows the combined breakdown probability for electrons and holes. As we can see from the plot, the probability that a carrier triggers an avalanche in SPAD2 is significantly higher and extends for a larger region. This vast region also relates to a broader spectral response of the device. Except for the highlighted difference in the energy diagram and the extension of the breakdown probability along the SPAD cross-section, the space charge and the electric field of SPAD1 and SPAD2 are nearly identical. This fact is justified by the same doping profiles of the two structures at the junction location, which also leads to the same breakdown voltage. Thus, the difference in performance that we will address in the section dedicated to the experimental characterization of these two devices could be attributed to the difference highlighted in the photo-collector region.

I-V curves and Breakdown

The I-V characteristic for the proposed structures were measured as described in Section A.1. Fig. 3.22 (a) shows the results of these measurements for SPAD1 and SPAD2. The two devices were tested in both dark and light conditions. Both devices show a breakdown of about 32 V. These results are also in good agreement with what was simulated using TCAD (Fig. 3.22 (b)).

Light Emission Test

The light emission test (LET) shows that both the SPAD structures present a uniform active area with no edge breakdown (see Section A.2 for more details on this test). Fig. 3.23 shows the results of this test. In addition, in the inset, we can see the micrographs of the SPADs used.

Noise Performance

Fig. 3.24 (*a*) shows the median noise normalized on the active area for SPAD1 and SPAD2, obtained on a set of 10 devices per family. The results show a pretty low noise level for the two devices, especially for SPAD2. Nevertheless, the low number of device tested makes this noise estimation not very accurate. Further testing is needed to improve the result. On the right-hand side of Fig. 3.24 (*b*) we can see the noise population distribution for 108 SPAD2 devices measured at 3 V excess bias. The results were obtained using the setup and measurement protocol described in Section A.4. In addition, a SPAD2 sample was used for noise characterization over temperature. Fig. 3.25 shows the results of these measurements. The temperature ranges from -60°C to 60°C. The device's noise improves by about three orders of magnitude within this range, going from about 7 kcps at 60°C and 7 V excess bias to 10 cps at -60°C and the same bias. From these data, we can extract the Arrhenius plots to understand

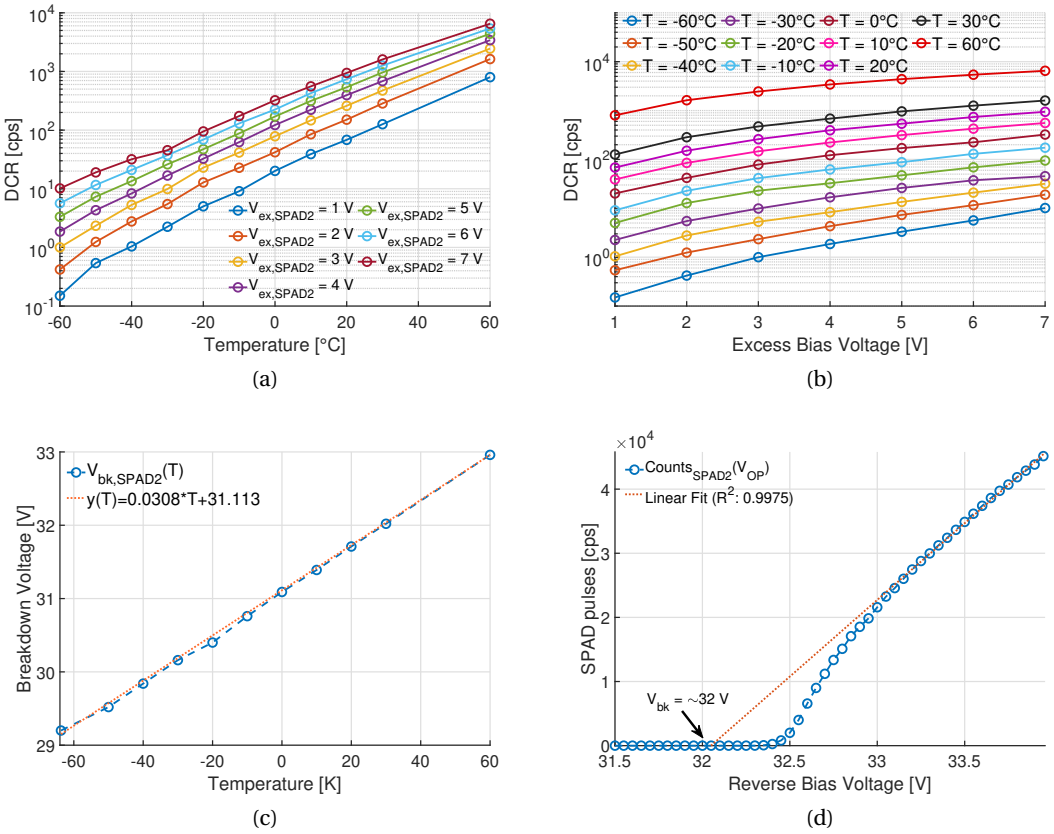


Figure 3.25: SPAD2 DCR characterization over temperature for several excess bias voltages (*a* and *b*). The breakdown shows a linear trend decreasing with temperature with a rate of 30.8 mV/K (*c*). (*d*) shows the technique used to extract the breakdown voltage when the SPAD output is not directly accessible and we observe the digital output of the pixel circuit. The plot shown refers to room temperature operation [279].

the main noise contributions. Fig. 3.26 shows the Arrhenius plot obtained with the SPAD2 data. The trend of the curves is almost constant, not showing a significant change in activation energy. The latter are summarized in Table 3.2. Looking at the data in Table 3.2, we can see how the activation energy turns out to be significantly lower than the energy bandgap of silicon. This result could indicate the presence of a contamination element that creates an intermediate energy level (see Section 2.1.5). In particular, referring to the data reported in [175], a possible guess could be the presence of Phosphorus vacancy defects, as also reported in [283].

Fig. 3.27 shows the afterpulsing of SPAD2 measured for several deadtimes and excess bias voltages. The measurements protocol follows what we discussed in Section A.5, using the inter-arrival time (IAT) technique. The SPAD integrates a passive quenching and active recharge (PQAR) circuit used to fine-tune the deadtime. This circuit is analogous to the one which we discussed in Section 3.1.1 (Fig. 3.16). The use of a more advanced technology node for its

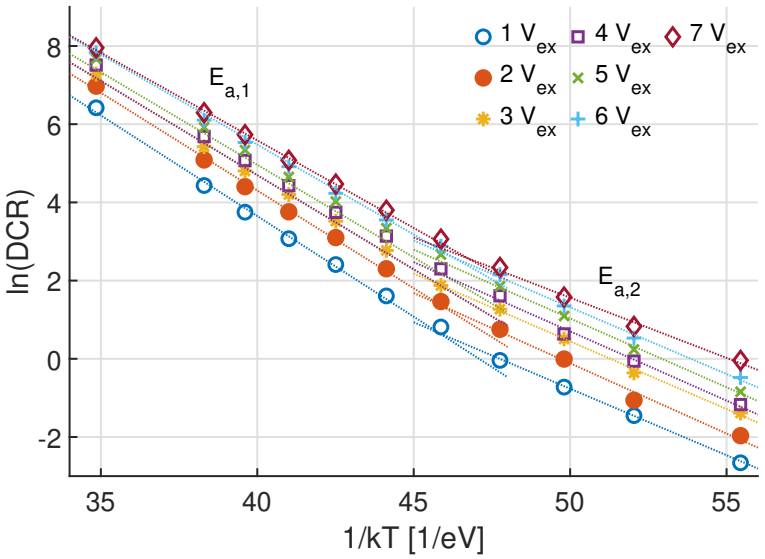


Figure 3.26: The Arrhenius plot for SPAD2, used to extract the activation energy values [279].

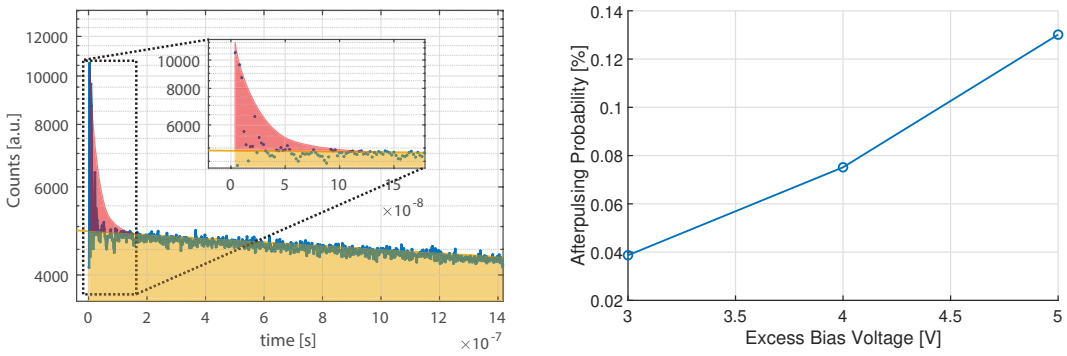
Activation Energies SPAD2		
Excess Bias Voltage (V_{ex})	Activation Energy SHR ($E_{a,1}$)	Activation Energy Tunneling ($E_{a,2}$)
1	0.5148	0.3401
2	0.4999	0.3595
3	0.4819	0.3465
4	0.4819	0.3543
5	0.4735	0.352
6	0.4599	0.3413
7	0.4479	0.308

Table 3.2: The activation energies extracted from the Arrhenius plot for SPAD1, in the SHR limited region and in the tunneling limited region. The data are taken for several excess bias voltage.

implementation helped in improving the circuit performance. Indeed, with this device, we were able to achieve a very low deadtime of about 1.5 ns. In addition, the output pulse shows a slew rate of 1.6 V/ns.
As we can see in the graph, the afterpulsing probability for this device never exceeds 1%.

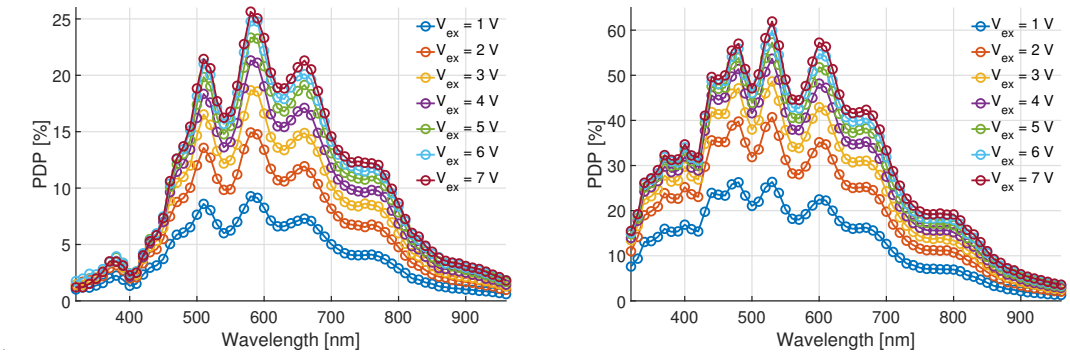
Sensitivity Performance

As for the previous device the sensitivity of the sensor, expressed in terms of PDP, is measured using the continuous light technique explained in Section A.6. Thus, we can now compare the PDP results of SPAD1 and SPAD2 shown in Fig. 3.28, respectively. The measurement is taken at room temperature and the steps are of 10 nm wavelength and 1 V excess bias voltage. From the graphs, we can observe a significant difference in the PDP profiles. Indeed, SPAD1 reaches a PDP peak of about 26% at 580 nm showing a cut in the sensitivity towards short wavelengths.



(a) (b)

Figure 3.27: (*a*): inter-arrival time histogram built for SPAD2 integrated with a PQAR circuit and a deadtime set to about 1.5 ns. (*b*): afterpulsing probability at different excess bias voltages.



(a) (b)

Figure 3.28: SPAD1 and SPAD2 PDP as a function of wavelength for several excess bias voltages [279].

On the other hand, SPAD2 reaches a PDP peak above 60% and a significant sensitivity below 400 nm. Fig. 3.29 and Fig. 3.30 show a comparison of the two SPAD sensitivity profiles, highlighting a significant difference between the two devices. In addition, in Fig. 3.29 a saturation effect is observable in the SPAD above 5 V excess bias for both devices. Moreover, we can notice that the two SPADs feature a similar behavior at longer wavelengths. This can be explained by considering that lower energy optical photons have a longer absorption length in silicon (Section 1.1),thus resulting in absorption at lower depths. Analyzing the two SPAD cross-sections in Fig. 3.18, it is clear that they show the same structure when going deeper in the substrate, while their main difference is the presence of a shallower PW implantation. This behavior was also predicted by the breakdown probability simulations reported in Fig. 3.21, showing a more extended high probability profile for SPAD2 at short wavelengths, while at larger depth, the probability of triggering an avalanche is similar for both SPADs. Finally, the standing wave pattern across the visible spectrum (Fig. 3.28) is a result of a non-optimized

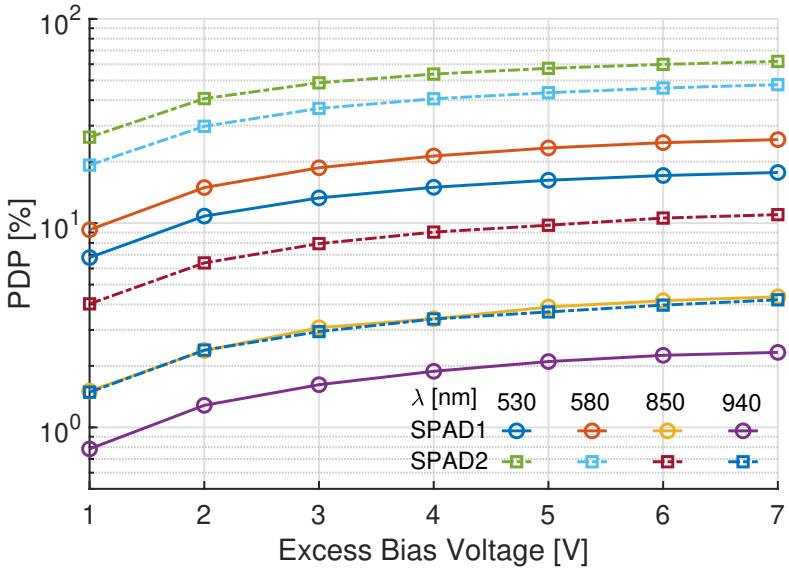


Figure 3.29: SPAD1 and SPAD2 PDP as a function of the excess bias voltage for different wavelengths at room temperature. It is possible to see the saturation trend typical of these devices as from about 4 V V_{ex} [279].

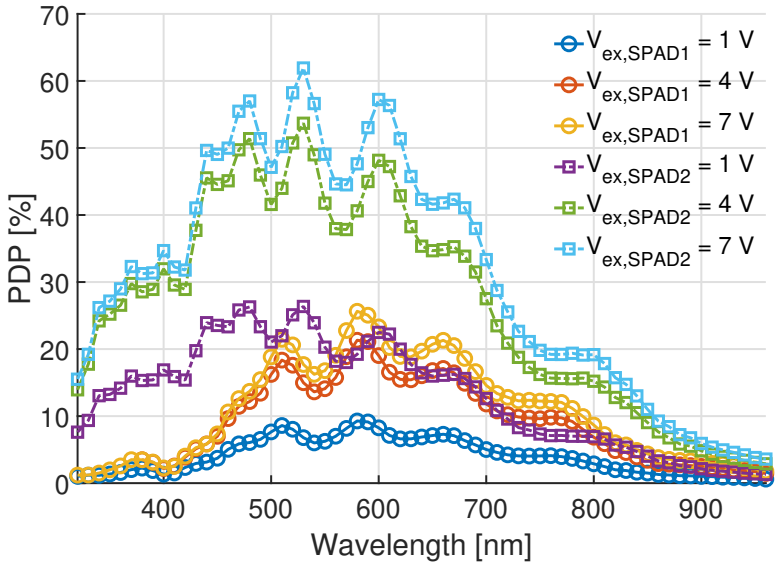
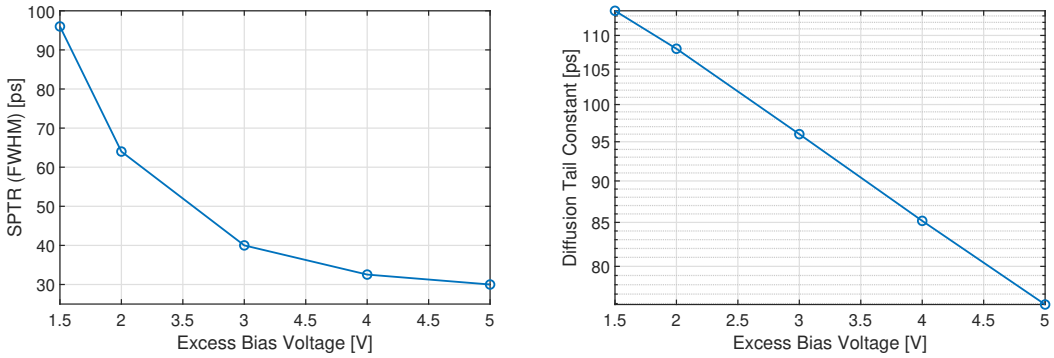


Figure 3.30: PDP comparison of SPAD1 and SPAD2 at room temperature [279].

optical stack on top of the silicon. An optimization of the optical stack is expected to enhance the PDP performance.



(a) (b)

Figure 3.31: In (*a*) we can see the trend of the SPAD2 SPTR as a function of the excess bias voltage applied. In this case the maximum excess bias voltage applied is 5 V, corresponding to an SPTR of about 30 ps. (*b*) shows, in log-lin scale, the trend of the diffusion tail constant increasing the excess bias voltage.

Timing Performance

Time-correlated single-photon counting (TCSPC) was utilized to obtain the IRF (see Section A.7). The setup used for characterization is analogous to the one reported in [232], and used in Section 3.1.4. A 1560 nm laser (NKT Onefive ORIGAMI-08) with second harmonic generation was used to output 150 fs pulses at a wavelength of 780 nm and a repetition rate of 50 MHz. A 45 GHz optical receiver was used as a trigger on the 1560 nm branch. Moreover, neutral density filters were placed to attenuate the light intensity to ensure operation in the single-photon detection regime, thus avoiding pile-up. Utilizing the PQAR circuitry at several excess bias voltages (V_{EX}), we achieved a minimum full width at half-maximum of 30 ps at 5 V V_{ex}, which is comparable to other CMOS SPADs in the literature that performed this measurement using integrated pixel circuits. The trend of the exponential time constant of the diffusion tail is presented in Fig. 3.31, (*b*). As we can see from the results reported in Fig. 3.31 the timing jitter improves by a factor of three over a range of 3.5 V of excess bias. Similarly, the diffusion tail as well becomes shorter at higher excess bias voltages, reaching a value of ~76 ps.

Geometrical Parameter Influence

During the design of SPAD2, various structure variations have been implemented to analyze the influence of several geometrical parameters on the device performance.
As mentioned before, this device relies on a virtual guard ring for edge breakdown prevention. Thus, an analysis of the guard ring size impact on the noise performance is mandatory to optimize the SPAD structure. Fig. 3.32 shows the influence of the active area (*a*) and guard ring size (*b*),(*c*) on the noise of the SPADs. As we can see, a virtual guard ring size smaller than 0.7 μm limits the operation of the SPAD to about 4 V. After this limit, the noise increases drastically. The reason behind this behavior can be explained by considering that the guard

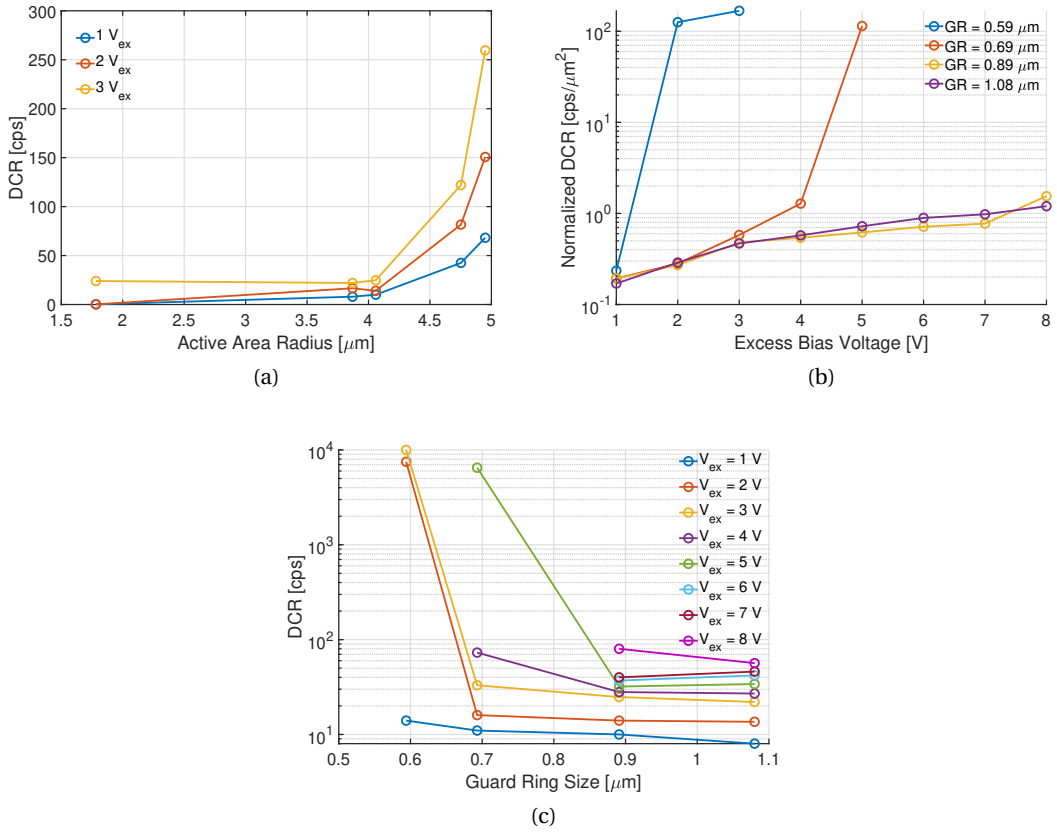


Figure 3.32: (*a*): the noise of SPAD2 devices with different active area for three excess bias value. (*b*), (*c*): guard ring size influence on the noise.

ring is a lightly doped region between the anode and cathode contact and that it can be fully depleted under high bias. Therefore, the smaller the size of the guard ring, the smaller the voltage needed to fully deplete this region. Nevertheless, it is crucial to optimize the size of this layer in order not to limit the overall fill factor of the structure and ensure good performance during operation.

3.2.2 Shallow Junction Substrate Isolated SPAD

To further explore our 55 nm BCD technology, we now focus our attention to another kind of substrate-isolated SPADs. These SPADs feature a shallow junction, in contrast to the structures discussed in the previous section, and they aim to enhance the NUV sensitivity and lower the breakdown voltage. In this context, we will analyze two distinct designs called SJ1 and SJ2.

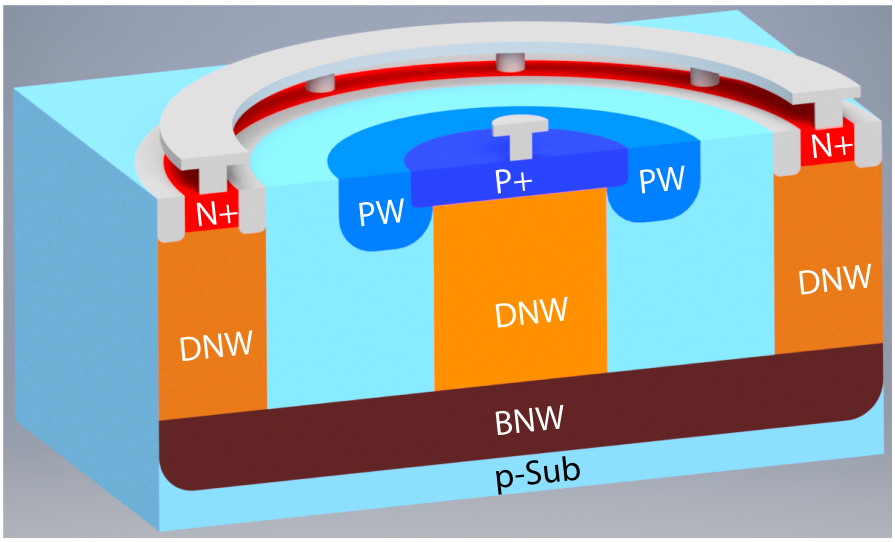


Figure 3.33: Cross-section of the first shallow junction SPAD implemented in 55 nm BCDLite, named SJ1.

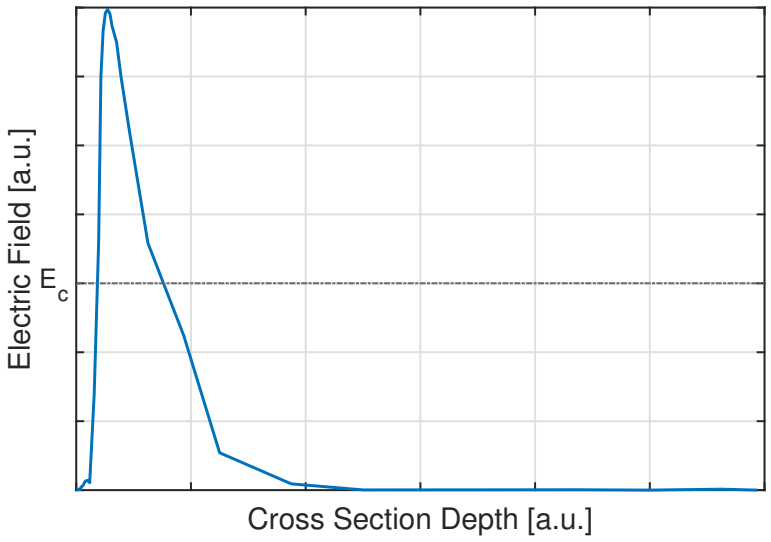
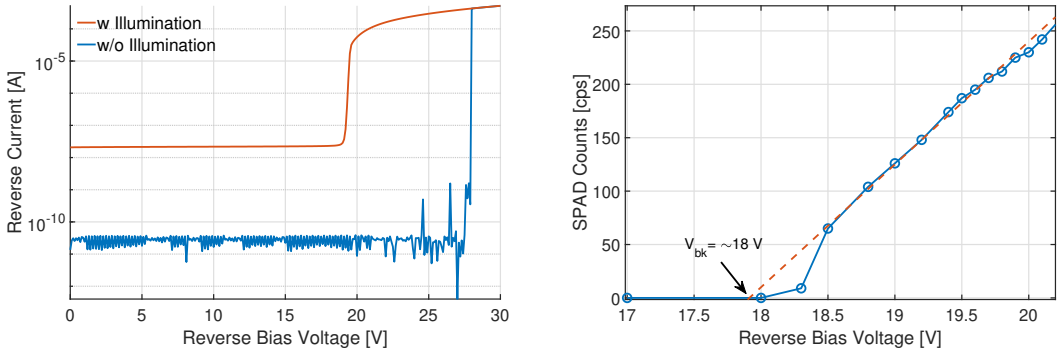


Figure 3.34: Electric field simulation of the SJ1 SPAD structure along the cross section depth. The electric field peak is quite narrow and it does not penetrate inside the DNW region of the junction.

SJ1 SPAD Structure

Fig. 3.33 shows the SJ1 device cross-section. This device cross section features a center anode formed by a *p+* implantation region, and it uses a physical guard ring structure based on a PW ring around it. The latter is used to reduce the net doping at the edge of the *p+* increasing the breakdown of this region and preventing premature edge breakdown, which can



(a) (b)

Figure 3.35: SJ1 SPAD I-V curve (*a*) and breakdown estimation in a pixel structure (*b*). In both cases, the breakdown estimation is about 18 V. The estimation performed on the pixel structure is based on the count rate variation at different excess bias voltages, under illumination.

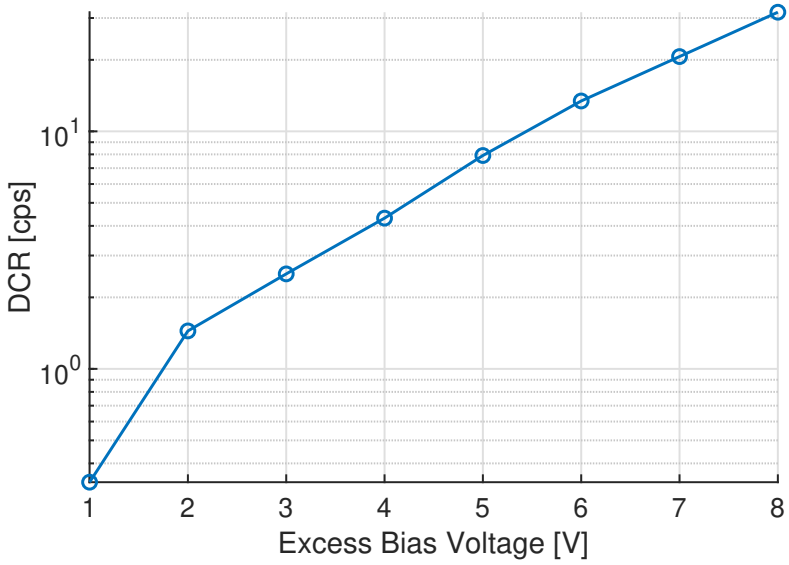


Figure 3.36: SJ1 SPAD DCR measurement as a function of the excess bias voltage.

compromise proper Geiger-mode operation. The Junction is formed with a DNW connected to the peripheral cathode contacts through a BNW layer. These two layers have relatively high doping, thus a relatively low breakdown voltage for this structure is expected.
The simulation result presented in Fig. 3.34 shows the electric field profile of the SJ1 SPAD structure along the cross section depth. We can notice how the electric field peak is confined in the proximity of the shallow junction location, formed between the p+ and DNW layers, and it does not penetrate the DNW limiting the collection efficiency for this device. This result suggests that the sensitivity of the device will be mostly in the blue and NUV region of the wavelength spectrum.

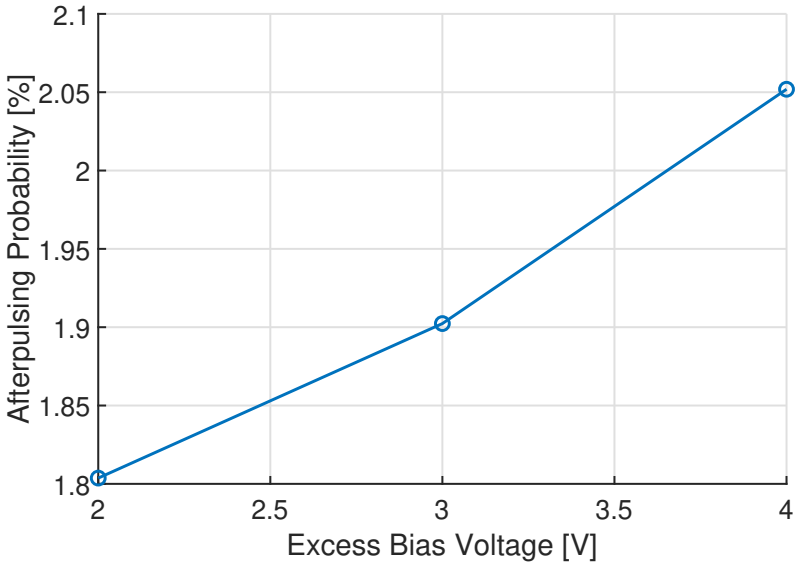


Figure 3.37: SJ1 SPAD afterpulsing measurement as a function of the excess bias voltage. All the values correspond to a deadtime of 1.5 ns.

SJ1 Measurement results

We will now go through the characterization of SJ1 main performance parameters. To get good performance of the device, we integrated the SJ1 SPAD with a PQAR circuit to reduce the deadtime of the device as well as the capacitive load at the SPAD output node.
Fig. 3.35 (*a*) shows I-V curve of the device both in the dark and under illumination. From the graph we can see how the breakdown of the device is about 18 V. This result was confirmed also by the extraction of the breakdown voltage for an SJ1 device integrated with the PQAR pixel circuit on-chip (Fig. 3.35 (*b*)) showing the same result.
The device's DCR has been measured, as explained in Section A.4, for several excess bias voltages, and Fig. 3.36 reports the results obtained. The device shows a correct operation up to 8 V excess bias, with a DCR level that does not pass the 30 cps. Above this bias value, the guard ring seems not to be effective anymore and the noise level is dramatically increases.
The inter-arrival time technique was used to characterize the afterpulsing of this structure. This measurement was performed as a function of the excess bias voltage and the results are shown in Fig. 3.37. As we can notice, the SPAD shows a maximum afterpulsing level at 4 V excess bias of about 2% (wiht a dead time of 1.5 ns).
Fig. 3.38 shows the PDP of the SJ1 device measured as discussed in Section A.6 with the continuous light technique (see Section A.6). The experiment was performed for several excess bias voltage levels (up to 7 V excess bias), and 10 nm wavelength step from 320 nm to 960 nm. As we can see, the PDP features a peak at 380 nm of about 28%. In addition, we can

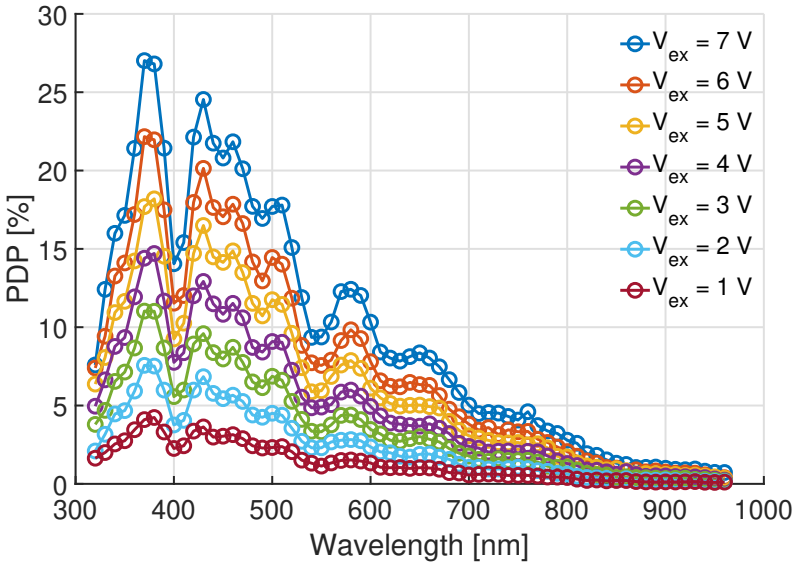


Figure 3.38: PDP measurement results for the SJ1 SPAD structure at several excess bias voltages.

observe a standing wave pattern caused by a non-optimized optical stack.
The timing performance of this device was measured using the same setup introduced in Section A.7, based on a 780 nm femtosecond laser. The device used for this experiment integrates the aforementioned PQAR circuit that allows a reduction of the dead time to 1.5 ns, overcoming the saturation issue given by the 50 MHz repetition rate of the laser system utilized, and it reduces the capacitive load at the output, improving the signal slew rate. In Fig. 3.39, we can see the results of two SJ1 structures with different active area size. The two structures present an active area radius of 4.5 μm and 1.9 μm, respectively. As expected, the larger structure gives a higher jitter at a low excess bias voltage, while the results of the two structures converge to a similar value for higher excess bias voltages. The minimum value obtained by these structures is about 50 ps FWHM for a bias of 4 V. Increasing the bias above this value provides a minimal performance improvement. However, 5 V excess bias is the limitation set by the pixel circuit used.

SJ2 SPAD Structure

The second structure presented in this section is the SJ2, whose cross section is shown in Fig. 3.40. As we can see, this structure consists of a shallow junction, implemented between a shallow p-well (SPW) and a DNW regions. Also this device features a substrate isolated structure, in which a BNW surrounds the device sensitive region. The cathode contacts are placed at the periphery of the device and implemented using the DNW layer. We should notice in Fig. 3.41 (*a*) that the electric field presents a peak close to the surface of the device at the junction location (multiplication region), and that the field intensity quickly drops going deeper into the cross-section. However, it remains high enough to extend the drift region up

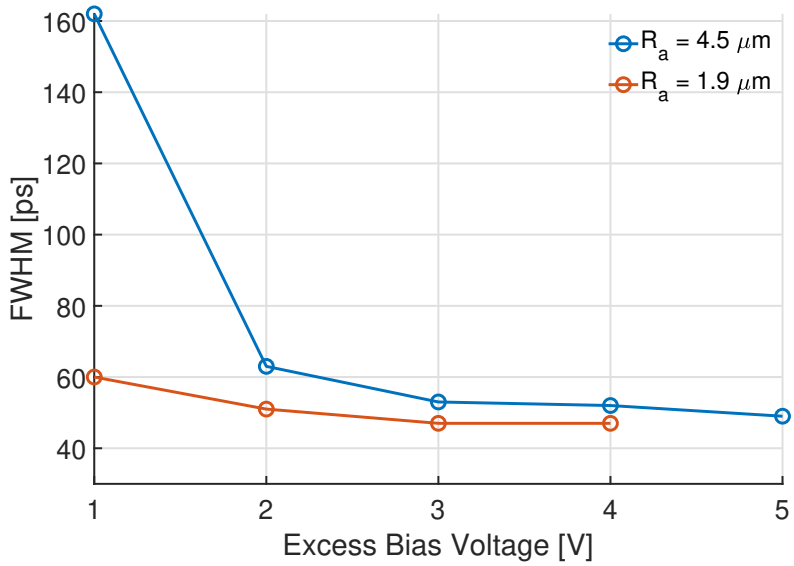


Figure 3.39: Jitter timing performance over several excess bias voltage for two SPAD sizes of the SJ1 structure.

to the BNW layer, indicating that the NW layer is fully depleted. This result is confirmed also by the breakdown probability (Fig. 3.41 (*b*)) that shows a nonzero value for the entire depth of the cross-section. With these simulation results, we can expect a peak of sensitivity centered towards the NUV and a non-negligible sensitivity also at longer wavelengths.

SJ2 Measurement Results

As a first step in the characterization of our device, we measured the I-V curve of the device (see Section A.1 for more details on this kind of measurement). Fig. 3.42 shows the result of this measurement both under illumination and in the dark. From the graph we can estimate that the value of the breakdown voltage for this device is about 20 V. The next step for the characterization of this device is the the DCR measurement. The results obtained following the procedure discussed in Section A.4 are shown in Fig. 3.43. The SPAD has a very stable behavior and does not show a abrupt increment of DCR raising the excess bias voltage up to V_{ex} = 18 V. The DCR at V_{ex} = 18 V is about 350 cps, and the device has active area diameter of 3.62 μm (see Fig. 3.44). Thus, the noise level at the highest bias point is about 34 cps/μm^2. This value is still relatively low, and it allows for an increase in the bias point of the device in order to improve its sensitivity performance. Moreover, to verify the size of the active area of the device we performed a laser scan as described in Section A.3. The setup used for this experiment is based on a confocal microscope and a 405 nm pulsed laser. The scanning was performed with a step of 100 nm, integrating the output pulses from the SPAD for 30 s. Fig. 3.44 shows the results of the scan, performed at different excess bias voltages. This experiment shows an insignificant increment of the active area size, which remains reasonably close to

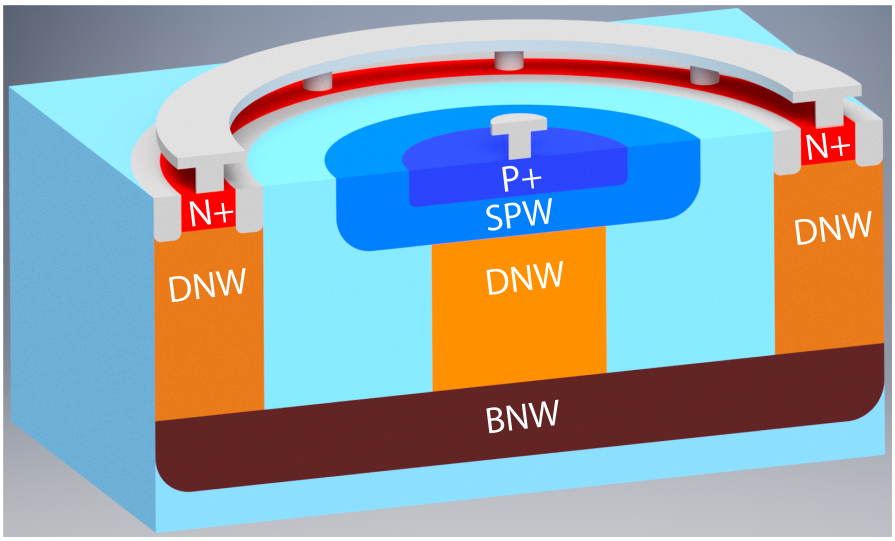
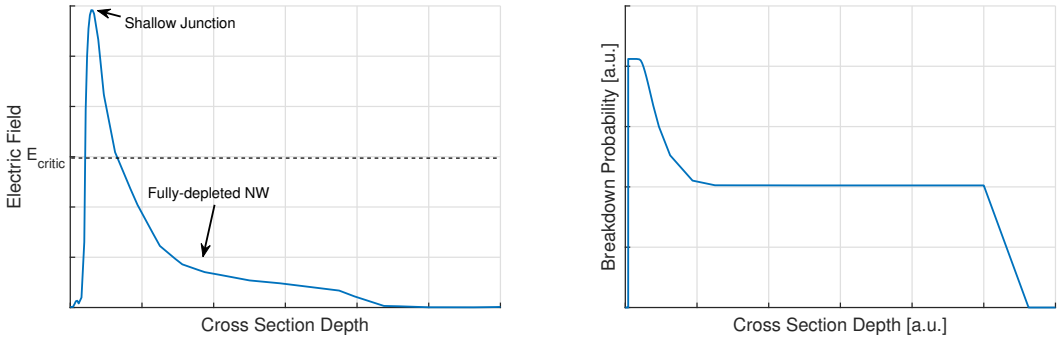


Figure 3.40: Cross-section of the SJ2 shallow junction SPAD implemented in 55 nm BCD technology.



(a) (b)

Figure 3.41: SJ2 simulation results. (*a*) shows the electric field along the cross section depth, while (*b*) presents the the breakdown probability profile.

the one drawn in the layout, during the design phase. Moreover, we can see how the SPAD shape is more defined at higher excess bias voltage also because of an increment of the SNR in the measurement. The wavelength used (λ = 405 nm) allowed us to decrease the achievable spot size, remaining in the specification of the optical element used. This reduced spot size allows to increase the final image contrast, specially when the SPAD under test has a very small active area. However, this wavelength corresponds to one of the deeps observed in the PDP spectrum of the device, reducing its sensitivity. To characterize the device sensitivity, we performed a PDP measurement using a continuous light technique [290] (see Section A.6) using the same setup presented in [232], [279]. The measurements were performed for various V_{ex} with a wavelength step of 10 nm. The results are presented in Fig. 3.45, where the PDP is shown as a function of the wavelength for some excess bias voltages (*a*) and as a function of

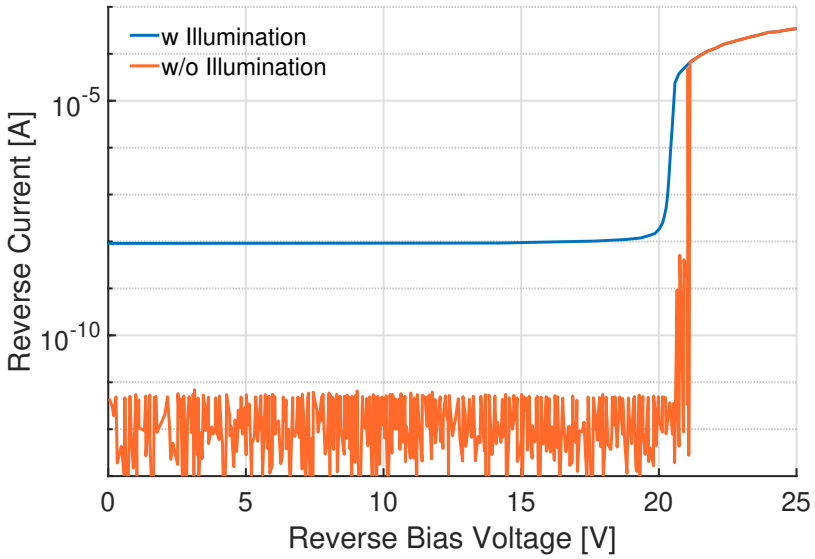


Figure 3.42: I-V characteristics of the SJ2 structure under illumination and in the dark.

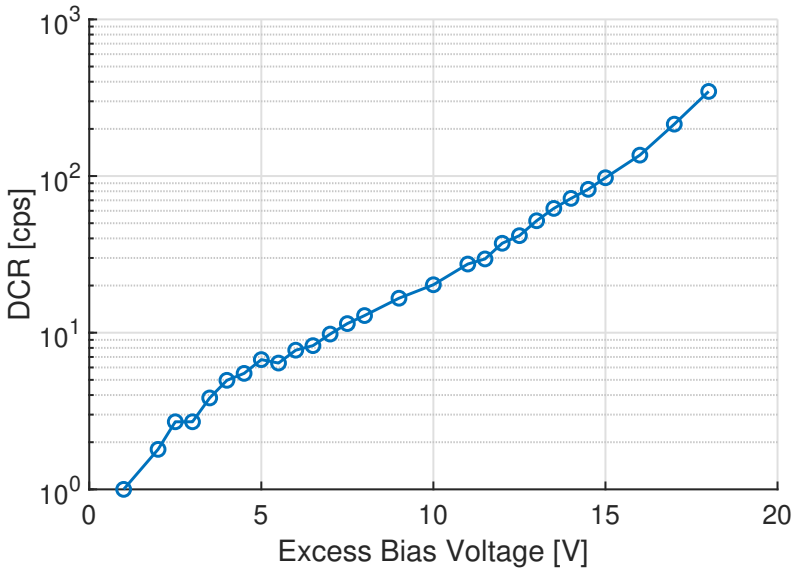


Figure 3.43: DCR measurement of the SJ2 SPAD structure.

V_{ex} (with finer steps) for some selected wavelengths (*b*). We can notice how the peak PDP is centered at about 440 nm reaching a peak of about 60% at 18 V excess bias. Also, The device has a PDP of about 45% at 360 nm. We can also note how the PDP tends to saturate increasing the excess bias voltage (see Fig. 3.45 (*b*)) because of a saturation of the ionization coefficients in silicon [190].
Moreover, we can see an interference pattern present on the PDP envelope, showing dips at 400 nm, 500 nm, and 650 nm. This pattern is the result of a unoptimized optical stack that

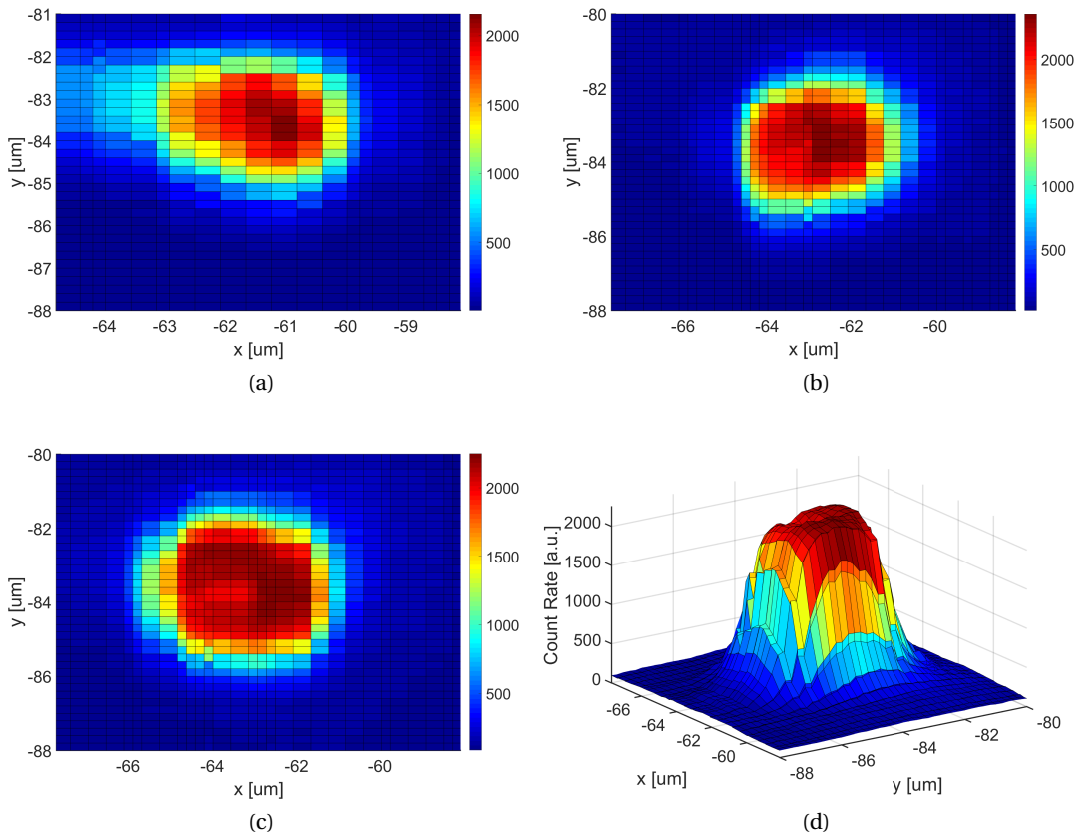
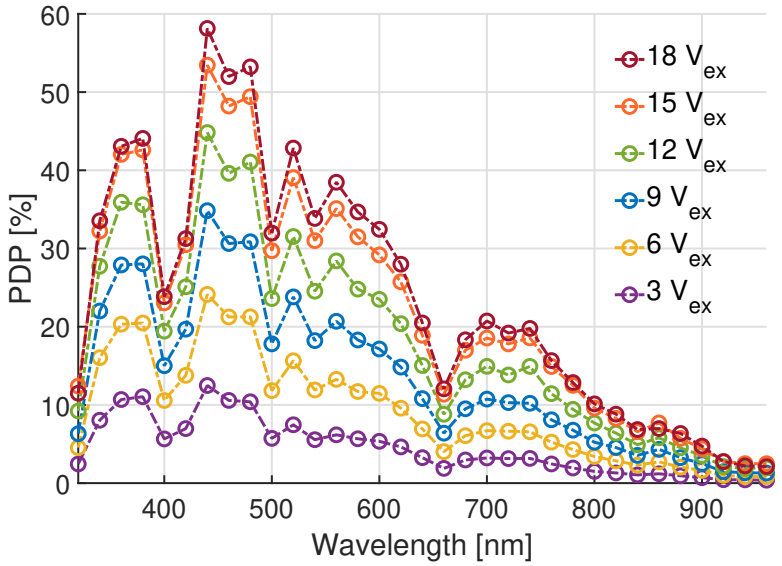
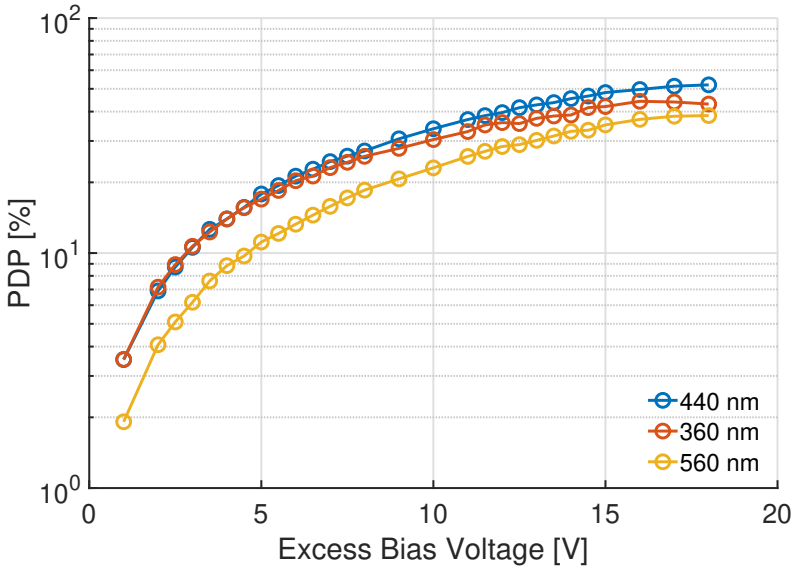


Figure 3.44: Laser scanning of the SJ2 SPAD active area at several excess bias voltage: (*a*), 5 V excess bias; (*b*), 10 V excess bias; (*c*), 15 V excess bias. A higher excess bias helps in improving the SNR of the measurement, and better define the active area of the device. In addition, we do not observe a significant increment of the active area with V_{ex}. For this measurement, we used a 405 nm wavelength pulsed laser, and a scanning step of 100 nm.

limits the number of optical photons reaching the sensitive region of the detector.
Next, the timing jitter was measured. In this case the SPAD does not have an integrated front-end circuit. For this reason, the timing jitter was measured by sampling the voltage waveform across an external load resistor. The result of this measurement at V_{ex}=10 V is shown in Fig. 3.46. The FWHM of the distribution is about 230 ps. This sub-optimal timing jitter can be explained considering the measurement conditions. The SPAD is quite small and it has a relatively low saturation current, as shown in Fig. 3.42. Consequently, the high excess bias voltage used for this measurement requires the use of a high resistive load to properly quench the avalanche. The use of external discrete components significantly increases the parasitic load at the SJ1 SPAD anode, which degrades the signal quality in terms of noise and slew rate. In the next implementation, we plan to add a front end circuitry to this device to allow a proper timing jitter characterization.



(a)



(b)

Figure 3.45: PDP measurement results at several excess bias voltages for the SJ2 SPAD structure. (*a*) shows the PDP as a function of the wavelength for some excess bias voltages, while (*b*) presents the PDP as a function of the excess bias voltage for some selected wavelengths.

3.3 Discussion

In this chapter we analyzed and discussed the SPAD structures implemented during this thesis, together with their pixel circuits. These structures were implemented in two different standard technology nodes: 180 nm CMOS and 55 nm BCD. Table 3.3 provides a comparison of the

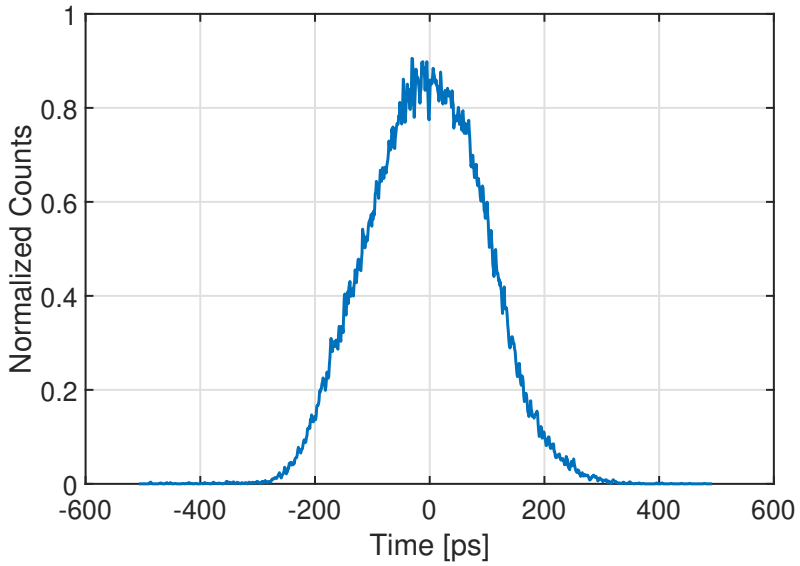


Figure 3.46: SJ2 SPAD Jitter Measurement. The result shows a jitter of about 230 ps FWHM.

key performance parameters of many SPADs shown in the literature, in comparison with what presented in this thesis. In this chapter we demonstrated unprecedented performance for SPAD implemented in standard technology, using a 180 nm CMOS node. Moreover, we have shown how it is possible to implement SPADs with state-of-the-art performance also in advanced sub-65 μm nodes. The use of a properly designed pixel circuit shows to have an important impact on the performance of the implemented device. The front-end is indeed related to the parasitic load at the the output node of the SPAD. This load is responsible of both afterpulsing and timing jitter performance.

Table 3.3: Performance Summary and Comparison to FSI SPAD

Reference	Year	Tech. [nm]	Junc.	Guard Ring	Active Diam. [μm]	V_{EX} / V_{BD} [V]	Peak PDP [%] @λ [nm]	PDP [%] @940 [nm]	DCR/A [cps/μm^2]	DCR @V_{EX}[f] [V]	AP [%]	Jitter [ps]	Jitter λ[nm]	Jitter V_{EX} [V]
Niclass et al. [123]	2007	130	p+/n	PW	10	2.7 /10	41 @450	1	1300[a,f]	100k @1.7	N.A.	144	637	1.7
Ghioni et al. [111]	2007	custom	n+/p+	p	50-200	5-10 /30-35	52-68 @550	9	0.4 - 1.6[f]	1k @5[f]	2	35	820	N.A.
Gersbach et al. [291]	2008	130	p+/NW	STI[m]	8.6	2 /9	30 @480	2	11.5k[a,f]	670k @2	<1[g]	125	637	1
Richardson et al. [292]	2009	130	PW/ DNW	N.A.[c]	8	1.4 /14	28 @500	2	0.6[a,f]	30 @1.4	0.02[h]	200	815	N.A
Richardson et al.[y] [142]	2011	130	PW/ DNW[v] p+/NW	N.A.[c]	8[x]	0.2-1.2 /12-18	18-33 @450- 560	2-3	0.5-0.97[a,f]	47 @0.8	0.02[i]	183 - 237	470	1.2
Gulinatti et al. [115]	2012	custom	n+/p+	p-epi	50	20 /45-55	58 @650	33	0.3[j]	580 @20[j]	N/A	93	820	20
Webster et al. [293]	2012	90	DNW/ p-epi[k]	N.A.[c]	6.4	2.46 /14.9	44 @690	12	4.6[a,f]	~150 @1	0.38[l]	51	470	2.46
Webster et al. [294]	2012	130	p-epi[k]	N.A.[c]	8	12 /20	72 @560	12	0.36[a,f]	18.0 @2	<4	60	654	12
Leitner et al. [295]	2013	180	p+/NW	N.A.[c]	10	3.3 /21	47 @450	N.A.	0.3-1.8[a,f]	180 @4	N.A.	N.A.	N.A.	N.A.
Charbon et al. [296]	2013	65	n+/PW	NW	8	0.4 /9	5.5 @420	0.2	340- 15.6k[a,f]	105 @.05	1[n]	235	637	0.4
Villa et al. [286]	2014	350	p+/NW	PW	10-500	6 /25	55 @420	2	0.05[a,f]	1 @4	1[o]	56- 4470	780	6
Veerappan et al. [139]	2014	180	p+/NW	pw	12	10 /23.5	48 @480	N.A.	12.8[a,f]	20 @2	0.3[p]	86	637	10
Lee et al. [297]	2015	140[d]	p+/NW	PW	12	3 /11	25 @500	3	244[a,f]	30k @3	1.7[p]	65	405	3
Veerappan et al. [298]	2015	180	p+/NW	PW	12	4 /14	47 @480	N.A.	16[a,f]	2k @4	0.2[q]	95	405	4
Veerappan et al. [280]	2016	180	p-epi/ DNW	N.A.[c]	12	12 /25	47 @520	N.A.	1.5[a,f]	200 @11	7.2[q]	97	637	11
Takai et al. [299]	2016	180	n/p[r]	N.A.[c]	16	5 /20.5	62 @210	8	0.5[a,f]	100 @5	0.35	161	635	5
Pellegrini et al. [300]	2017	130	PW/ DNW	N.A.[c]	8	0.5 /14.2	43 @480	1.4	1.6[a,f]	80 @0.5	0.08	100	N.A.	2.4

Continued on next page

Table 3.3 – continued from previous page

Xu et al. [301]	2017	150	p+/NW	N.A.[c]	10	5 / 18.0	31 @450	2	0.4[a,f]	200 @5	0.85	42	831	4
Pellegrini et al. [119]	2017	40	PW/ DNW	N.A.[c]	N.A.	1 /15.5	45[s] @500	3	N.A.	50 @1	0.1	140	850	1
Sanzaro et al.[y] [144]	2017	160[e]	p+/n p/ DNW	N.A.[c]	10-80	3-9/ 25-36	58-71 @450-490	3	0.1-0.2[a,f]	100 @5	0.02-1.59	28-41	820	5
Moreno-Garcia [302]	2018	110	p+/LDNW	N.A.[c]	10	8/20	58@455	NA	1.16	450@6	NA	71	831	4
Vornicu et al. [303], [304]	2020	110	p+/DNW	polysilicon	5	5/18	67.4 @500	NA	2.61	204.5 @5	0.08	92	850	3
This Work (SPAD1)	2021	55[e]	DPW/ BNW	N.A.[c]	8.8	7/ 31.5	27 @530	4.2	NA	NA @1-7	NA	NA	NA	NA
This Work (SPAD2)	2021	55[e]	DPW/ BNW	N.A.[c]	8.8	7/ 31.5	62 @530	4.2	2.6[a,f]	156 @7	~0.13[t]	30	780	5
This Work (SJ1)	2021	55[e]	p+/ DNW	PW	9-3.8	7/18	27 @370	0.8	0.47[a,f]	30 @8	~2[β]	52	780	5
This Work (SJ2)	2021	55[e]	SPW/ DNW	N.A.[c]	3.6	18/ 20	60 @440	2.6	34[a,f]	350 @18	NA	230[γ]	780	10
This Work PIN	2021	180	p-epi/ BNW	N.A.[c]	~18	6/22	55 @480	~3	0.23[a,f]	100 @6	0.12[z]	7.5[α]	515	6

[a]Taken at max excess bias if not range of excess bias values not specified. [c]Virtual guard ring structure. [d]Silicon-on-insulator. [e]BCD. [f]At 20°C. [g]180 ns dead time. [h]200 ns dead time. [i]50 ns dead time. [k]Not substrate isolated. [l]15 ns dead time. [m]Shallow trench isolation with passivation implants to create p-type glove structure. [n]At 1 μs dead time. [o]30 μm diameter. [p]200 ns dead time. [q]300 ns dead time. [r]Surface-isolated n-spad/p-spad junction. [s]With microlens. [t]1.5 ns dead time with 50 % level @ 3 V_{EX}. [v]Two different deep structures presented one with an epi layer and one with a pw implant. [x]Multiple diameters demonstrated. [y]3 SPAD structures proposed. [z] 3 ns deadtime. [α] at 6 V excess bias.[β]at 1.5 ns dead time and 4 V excess bias.[γ]at 10 V excess bias, measured with external load.

4 3D-Stacked FSI Multichannel Digital SiPM

TOF-PET requires the design of large area photodetectors on the order of a few mm^2 to detect gamma rays in scintillator-based modules. To achieve this goal, we chose to implement a dSiPM that could, in principle, improve timing performance with respect to aSiPMs [69]. Typically, a dSiPM integrates complex CMOS circuitry to enable extra on-chip functionalities, such as active quenching and recharge, masking, memory, timestamping, and processing. To enable the implementation of significant functionality on chip and maintain high fill factor, we opted for a 3D-stacked FSI approach, as shown in Section 2.3.4. Moreover, we decided to implement our sensor relying on an MD-SiPM architecture to enable the acquisition of multiple timestamps per event. Having multiple timestamps can be helpful, as it allows one to take advantage of photon statistics. Over the years, a theoretical background was established on the statistics of photons generated by a gamma through scintillation process [79], [305]–[311]. These studies showed that it was possible to estimate gamma-ray's timemark from multiple timestamps with predictable optimal timing resolution. The lowest variance is indeed not necessarily obtained with the first optical photon generated by the scintillator [306]. Rather, it is a combination of multiple photons whose timestamp is properly weighted that enables one to improve the timing precision of the gamma ray. Such timing precision is essential to improve SNR and the overall image quality [312].

The specifications of the detector presented in this chapter are the result of a compromise between application requirements and the performance enabled by 3D technology process. The technology process used to implement our 3D-stacked FSI sensor set a limit on the minimum size of the final chip. These limitations are linked to the capability of the manufacturing process.For this reason, the chip cannot be smaller than about 6 mm × 3.5 mm. This constraint made us select the large size of 7.5 mm × 4.2 mm for our detector. To decrease the complexity of such a large chip, we decided to segment the sensor into two parts. The SPADs used here were designed to reach a higher than ~60% FF and to achieve high PDE. The two chip segments rely on an MD-SiPM architecture, and they are sub-divided in clusters. The clusters'size was selected to achieve a good compromise between granularity and system complexity and also to allow the coupling with a scintillator needle matrix with sub-millimeter pitch. In addition, each cluster includes timestamping and photon counting circuitry.

In this chapter, we focus on our detector implemented in a FSI 3D-stacked CMOS process. This process was selected for the implementation of a MD-SiPM. The chapter starts with a description of the process flow employed for the fabrication of our detector. Next, the FSI architecture of the sensor is presented and discussed. Finally, the results of the sensor characterization are presented.

The work presented in this chapter is based on [313]–[315].

4.1 3D Stacking Process Flow

3D integration started to attract the researchers' interest as it opens the door to an extension of Moore's law. Indeed, the goal of 3D integration is to provide multiple stacked semiconducting

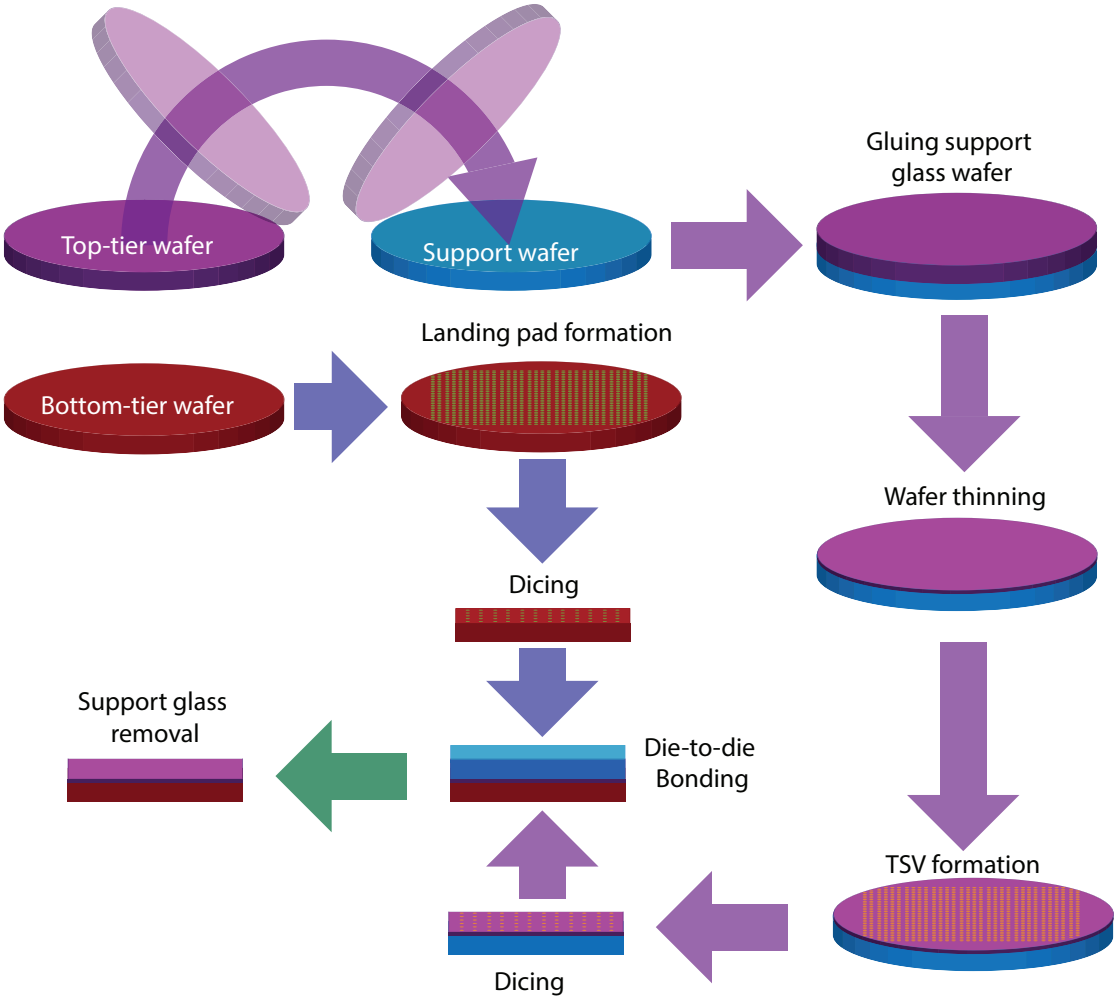


Figure 4.1: 3D stacking process flow of our 3D-stacked FSI MD-SiPM. Two separate wafer (one for the top- and one for the bottom-tier) are processed in parallel. The top-tier wafer is initially bonded on a support glass wafer for mechanical stability. Then, the wafer is thinned to a specific thickness, and the TSVs are implemented from the back-side. Next, the wafer is diced and the top tier die is extracted. At the same time, the landing pad and under bump layer metalization (UBM) is implemented on the front-side of the top wafer before dicing it to extract the bottom tier die. The two tiers are now bonded together in a face-to-back fashion at die-to-die level and the support glass is removed.

materials for devices. Two approaches are available for the 3D integration of SPAD-based devices: face-to-face (F2F) and face-to-back (F2B) [316], [317]. The F2F approach implies that the front sides of two wafers are bonded together, exposing the back-side of the top-tier to light. For this reason, this approach is the one used when implementing 3D-BSI devices (see Section 2.3.4). Conversely, the F2B approach relies on the use of TSVs to connect the two tiers. Indeed, they are assembled with the back-side of the top-tier facing the front-side of the bottom-tier. The F2B approach leaves the front side of the top-tier wafer exposed to light, and

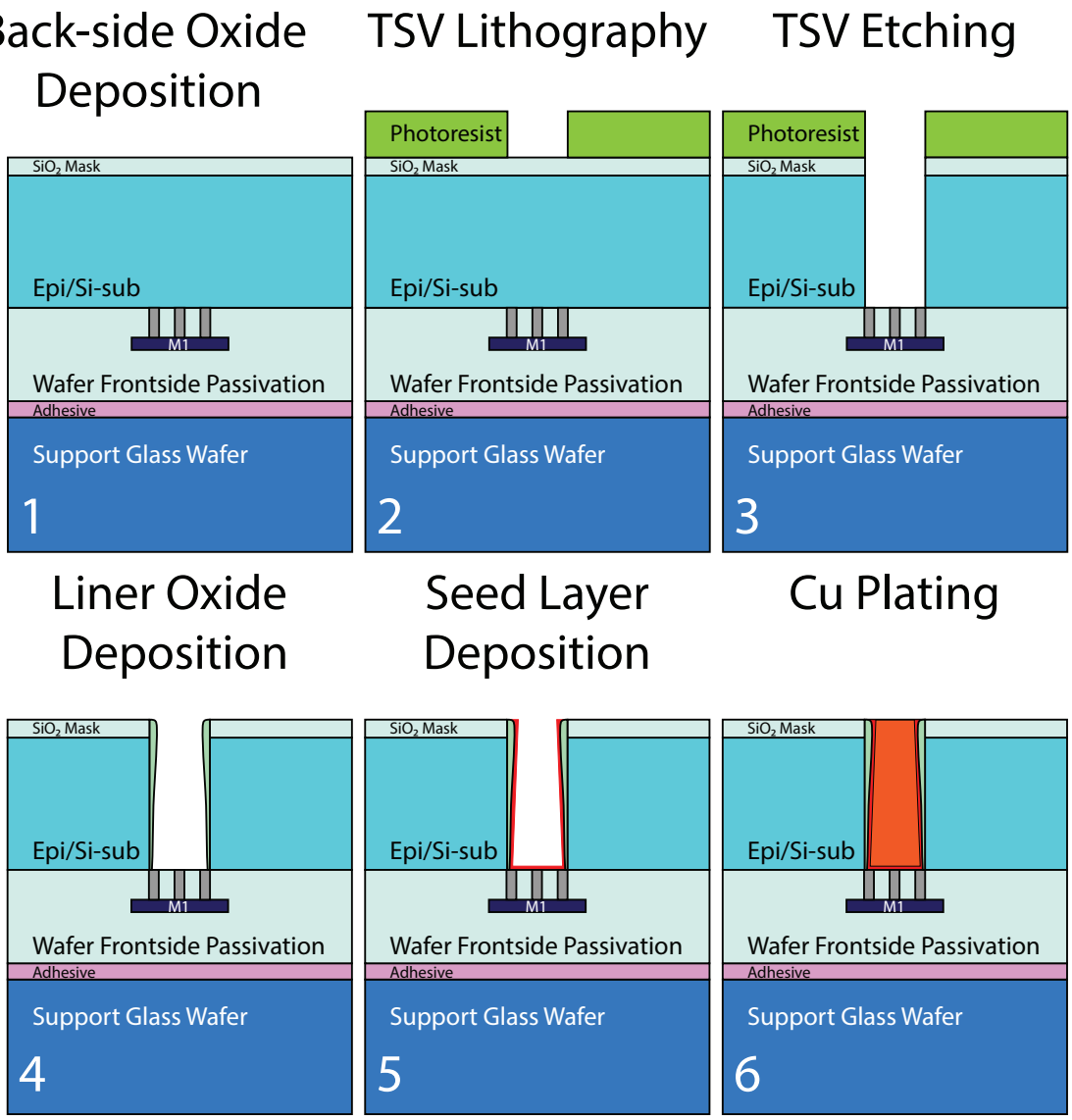


Figure 4.2: TSV process flow.

for this reason, it is suitable for the implementation of 3D-FSI devices (see Section 2.3.4).

Although the 3D integration approach is pretty attractive, it is not universally standardized for SPAD-based devices. Therefore, the cost and the risk associated with this kind of implementation are usually higher than those of standard 2D approach. Fig. 4.1 shows the process flow used for implementing the detector presented in this chapter. Two wafers containing the design of the top- and bottom-tier chips are implemented in standard technology nodes. In principle, the two wafers can employ different technologies, non-standard as well. However, in this work, we opted for the use of a 180 nm CMOS technology for both tiers.
Proper alignment markers are placed during the design of the two tiers to allow lithography

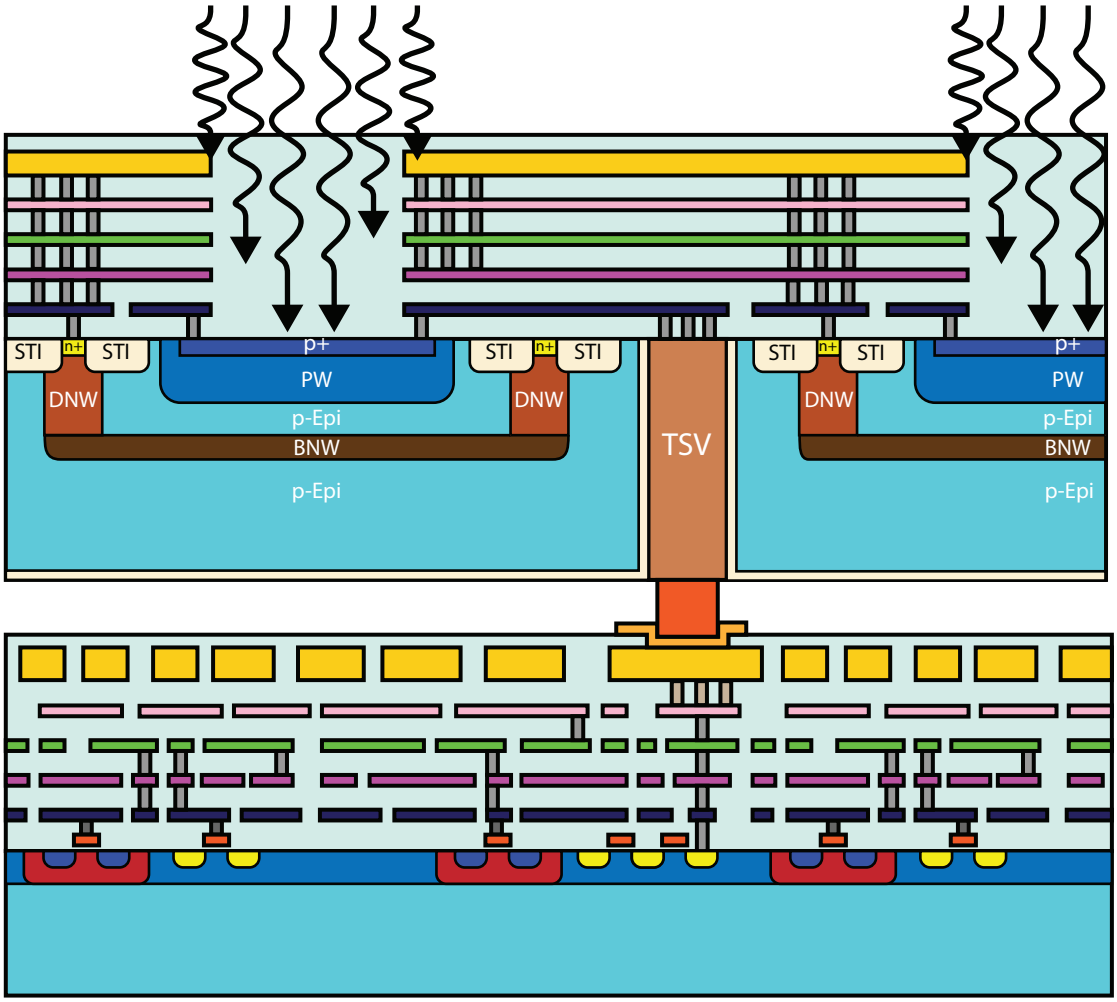


Figure 4.3: Architectural scheme of the 3D-stacked FSI MD-SiPM.

techniques in the post-processing phases and 3D bonding. After the CMOS fabrication process is complete and the wafers are ready, they undergo a step dedicated to the 3D. In this phase, the two wafers are post-processed in parallel: one wafer will be the top-tier die and the other the bottom-tier die.

Let us start by following the steps the top-tier wafer undergoes during the post-processing. First, the top-tier wafer is glued on a glass support wafer for mechanical stability. The use of a glass support wafer allows thinning the top-tier up to a suitable level for the following process steps (less than 100 μm), avoiding risks of cracks and making it easier to handle the thinned wafer. Next, the TSV implementation takes place on the backside of the top-tier, as shown schematically in Fig. 4.2. During this process, a SiO_2 layer is deposited on the top-tier backside, and lithography is used to create a mask for selective etching of the TSV region. A liner oxide is then deposited to avoid electrical connections between the TSVs and the silicon substrate. Next, a seed layer deposition, followed by Cu plating, is performed to fill the TSV. Fig. 4.4 shows a cross section of the TSVs and microbumps. Once the TSVs are complete,

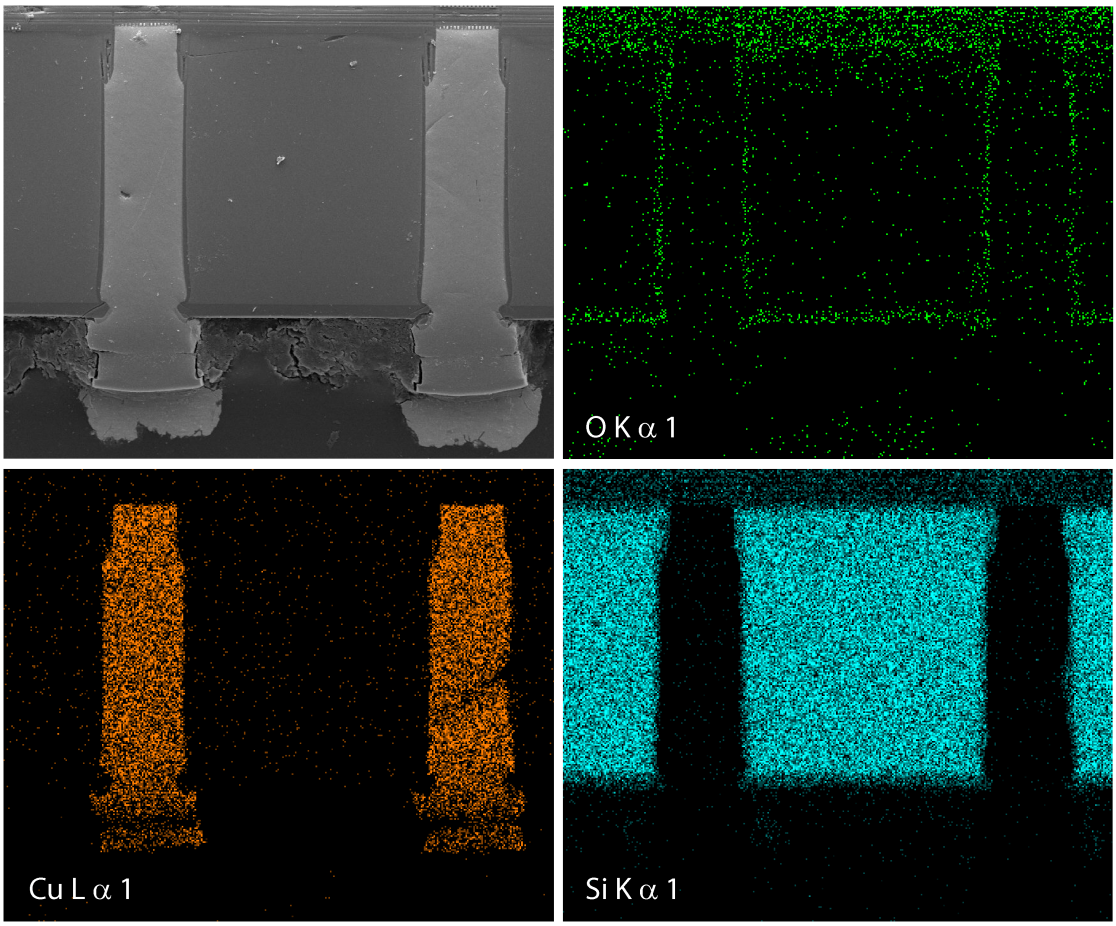


Figure 4.4: SEM image of TSV cross section (*top-left*). An energy dispersive X-ray Analysis (EDX) is performed to check the cross section. The oxygen distribution is used to highlight the presence of oxides (*top-right*) around the TSV. The copper distribution highlights the correct filling of the TSV (*bottom-left*). Finally, the distribution of silicon highlights the substrate area (*bottom-right*).

microbumps are formed, again through a lithography process, over the TSV connections on the top-tier backside. Finally, dicing the wafer allows to extract the post-processed top-tier chips.
The bottom-tier wafer follows parallel post-processing steps. First, the landing pad area is prepared, forming the under bump metallization (UBM) regions. Then the wafer is diced to extract the bottom-tier dies. Once the two tiers are ready, the 3D bonding process at the chip-to-chip level is performed. The glass support, glued on the top-tier die, can now be removed.

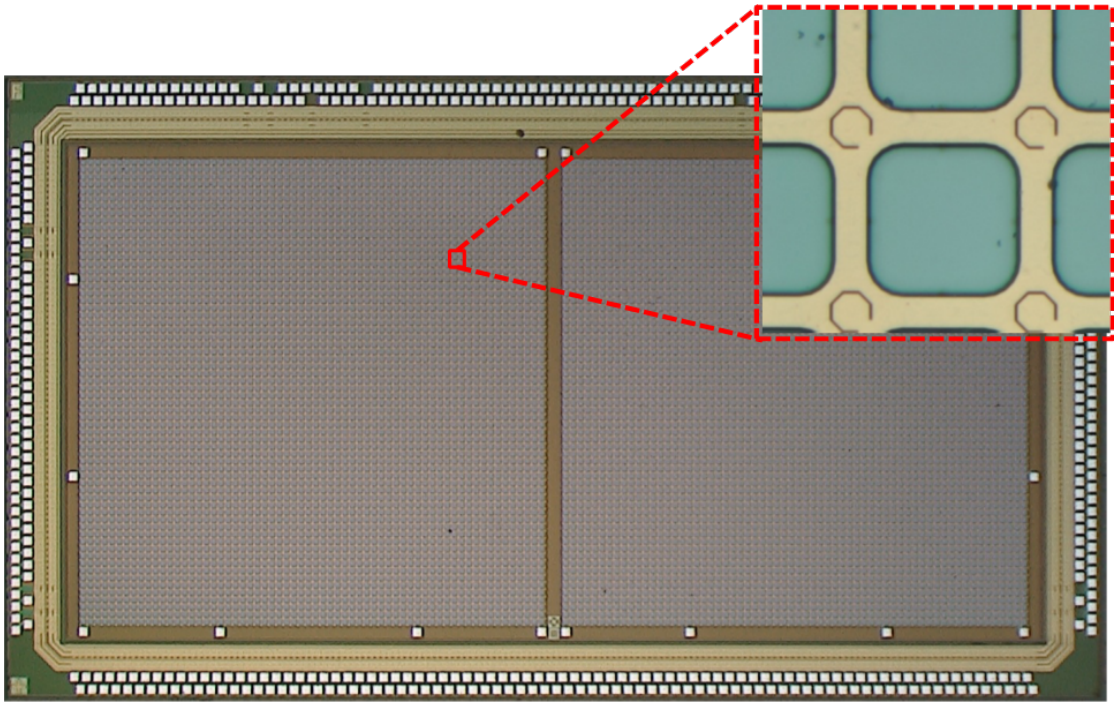


Figure 4.5: 3D-Stacked Multi-Digital SiPM micrograph.

4.2 3D-Stacked FSI Architecture Implementation

At the architectural level, the detector presented here follows the 3D per pixel connection approach discussed in Section 2.3.4. Fig. 4.3 shows the 3D-stacked FSI cross-section of the implemented chip. The top tier houses the SPAD, whose cross-section is the same as the device presented in Section 3.1, based on a *p-i-n* structure. This SPAD is a substrate-isolated device, with the anode contact placed at the center of the device (implemented using a *p+* implantation region). The SPAD has a square shape with round corners and a pitch of 50 μm (Fig. 4.5), reaching a FF of about 67%. The 3D integration technique that we selected relies on the F2B architecture. Thus, the anode contact of the SPAD connects directly to the bottom tier through the TSVs implemented in post-processing. Fig. 4.4 *top-left* shows a cross-section of the implemented TSVs. In addition, an energy dispersive X-ray analysis (EDX) was performed to check the composition of the cross-section and verify that the structure of the implemented TSV was correct, looking for the presence of some specific elements in the various chip regions. First, the oxygen highlights the oxide regions present around the TSV and on the backside of the top-tier (Fig. 4.4 *top-right*). This oxide layer is created to avoid short circuits between the TSV conductor and the silicon substrate. Next, the copper shows if the TSV conductor is correctly filling the entire TSV region (Fig. 4.4 *bottom-left*). Finally, silicon highlights the remaining region of the substrate (Fig. 4.4 *bottom-right*). Although this exam can be very detailed, it is destructive, and it can not be performed over too large an area. Fig. 4.5 shows the micrograph of the implemented device. The whole chip measures about 7.5×4.2 mm^2, and it comprises two arrays of 64×64 SPADs each. Each SPAD has its own TSV anode connection to

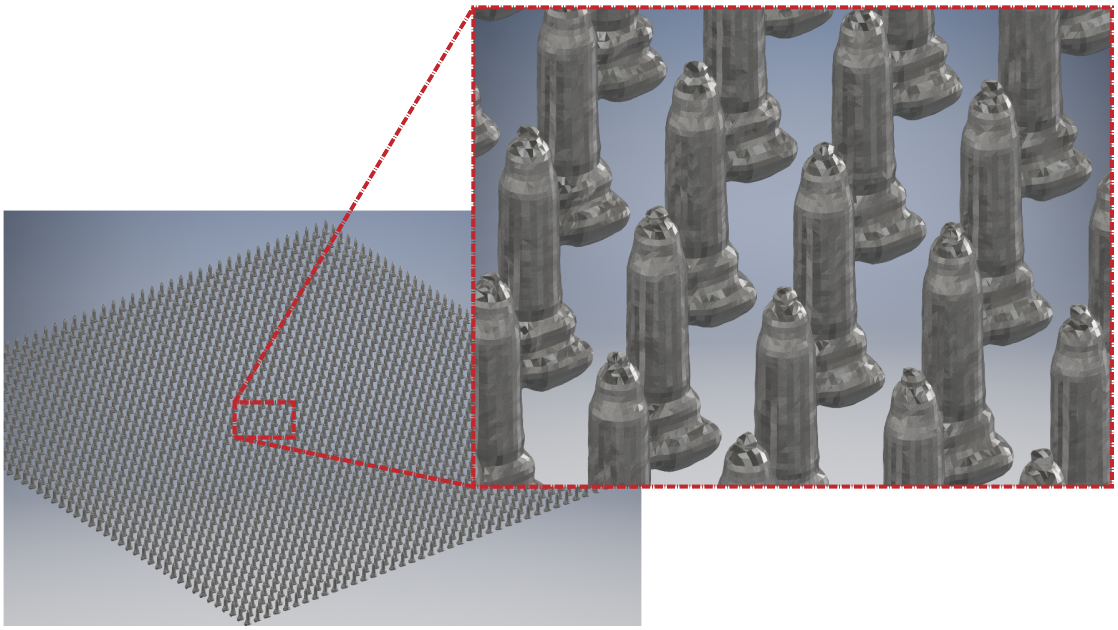


Figure 4.6: Top-tier X-ray tomography. The TSVs are reconstructed with a voxel of 1.42 μm size.

the bottom-tier, for a total of 8192 3D connections distributed with a 50 μm pitch along both the x and y-axis.
Because the detector presents a large size, X-ray tomography was also performed on the entire array to check the uniformity of the TSV structures. Fig. 4.6 shows the results of the tomographic analysis. In this test, we reconstructed a 3D model of the detector, separating the TSV conductors from the silicon substrate with a voxel size of 1.42 μm. Although the resolution is way lower than the one achieved with SEM imaging, this test is non-destructive and can cover a large area. Furthermore, from the analysis of the reconstructed image, we could confirm a good uniformity of the TSV structures, their size, and pitch. Nevertheless, this does not allow us to provide an absolute value of the yield. However, it allows us to highlight the presence of any severe problem in the structure.
From Fig. 4.5, we can note that some bonding pads are located on the top-tier die. These pads are the substrate and high-voltage connections of the SPADs in the top-tier. The reason behind this choice relies on the properties of the implemented TSVs. As mentioned, the TSV conductor region is surrounded by an oxide layer to prevent electrical contact with the silicon substrate. The thickness of this oxide is related to the maximum voltage which the TSV can sustain. Indeed, a too high voltage across the oxide layer can induce dielectric breakdown and form percolation paths that can create electrical connections between the TSVs' metallic region and the silicon substrate. This voltage limitation implies that, ideally, different TSV structures should be implemented depending on their function. Indeed, the ones dedicated to high-voltage connections should feature a thicker oxide layer and a larger conductor area to sustain the large current which the SPAD array could sink. To reduce the complexity and the cost of the process, we decided to implement only one type of TSV structure, namely

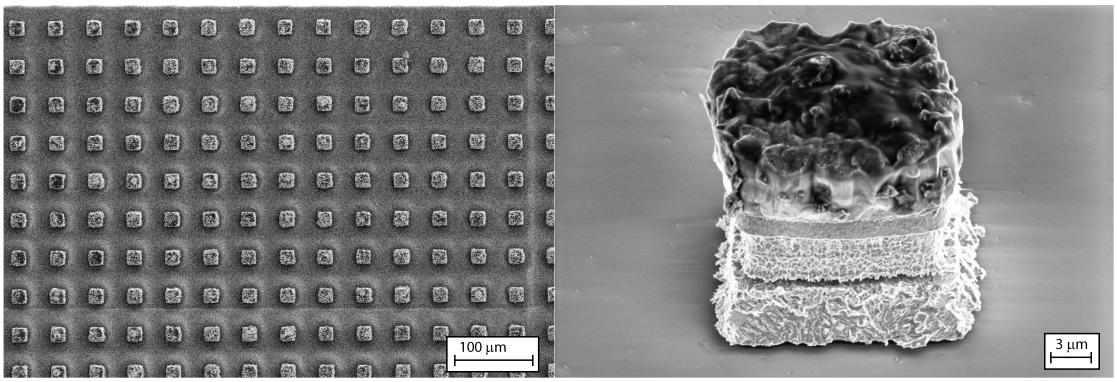


Figure 4.7: SEM image of the top-tier back-side. The micropillars implemented by lithography are uniformly distributed with a pitch of 50 μm along both x and y direction (*left*). A detail of a micropillar is shown on the right-hand side of the image.

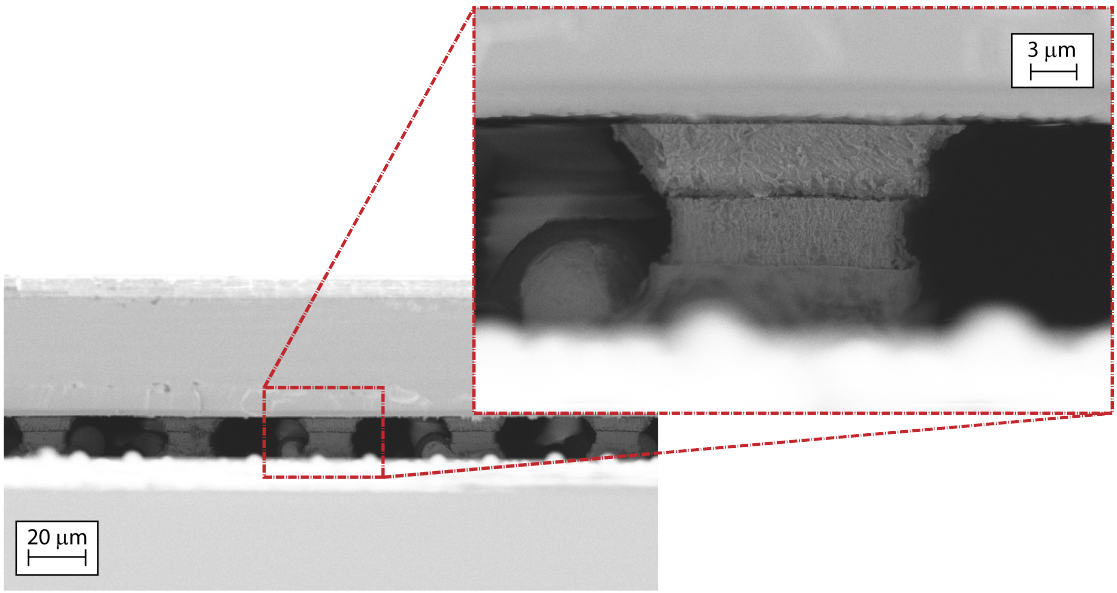
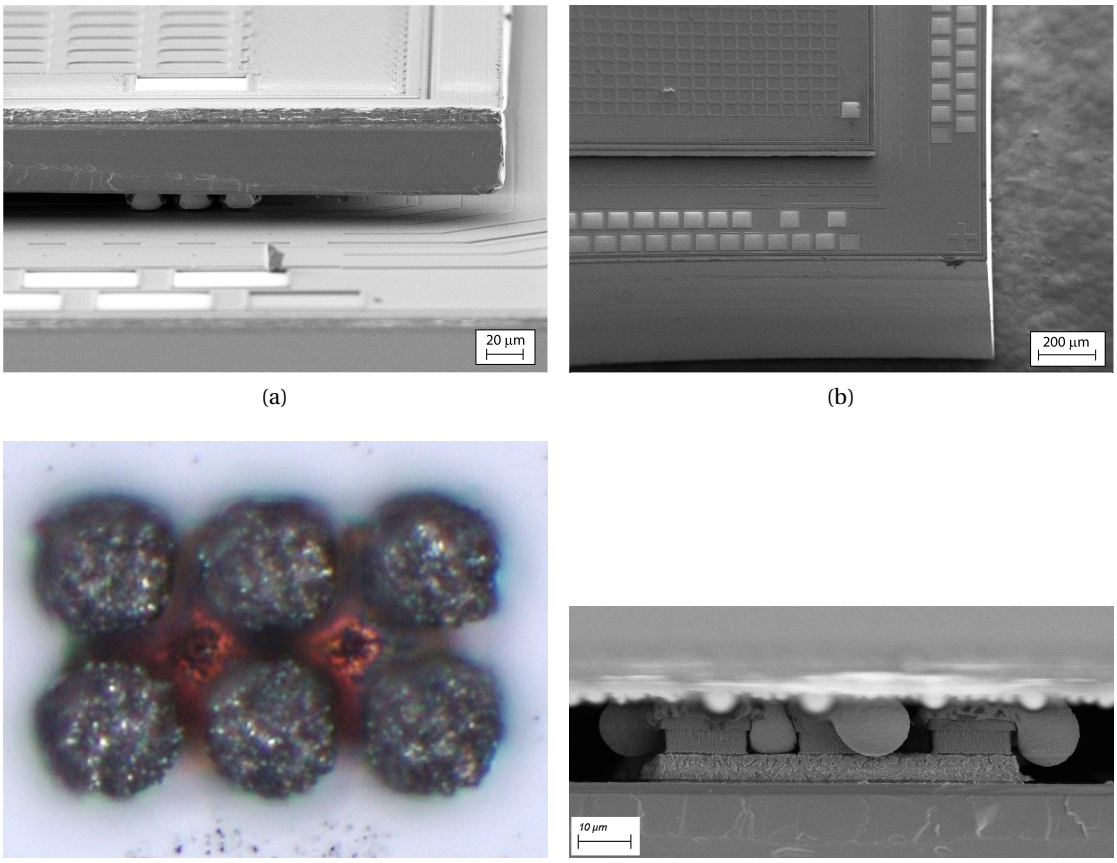


Figure 4.8: SEM image of micropillars connected to the TSVs after 3D bonding.

the one dedicated to the SPAD anode connection. The voltage that the TSV core can reach corresponds, in this case, to the excess bias voltage applied to the SPAD. Therefore, we set the limit on the maximum excess bias voltage to 6 V. The choice of this value comes from the performance observed in SPADs based on the same cross-section presented in Section 3.1. For this design, we opted for a conductor diameter of about 10 μm and a TSV pitch of 50 μm, matching the SPAD size.

As mentioned in Section 4.1, microbumps are fabricated on top of the TSV on the top-tier backside to allow 3D bonding. Fig. 4.8 shows an SEM imaging of the top-tier backside (left) and the detail of a single microbump (right) before the 3D bonding. The bumps are made of a Sn-Ag alloy region on top of a Cu pillar and a redistribution layer (RDL). Fig. 4.8 shows



(a) (b)

(c) (d)

Figure 4.9: SEM images of the corner of the chip at different magnification levels. (*a*) Gap between the two tiers with the micropillars placed below a bonding pad for mechanical support. (*b*) Top view of the 3D-stacked system. (*c*) Top view of the pillars. (*d*) Side view of the microbump cluster placed below the bonding pad, after 3D bonding.

the microbumps after the 3D bonding process. It is possible to see how the microbumps present a small extrusion caused by the pressure applied during the stacking process. This extrusion was considered during the detector design phase and provided some indications for selecting the pixel pitch. The wire bonding process implies the application of a certain pressure. Considering the reduced thickness of the top-tier wire (see Fig. 4.9), thinned to fabricated the TSV connections for the SPADs, a pressure applied close to the edge of the device can crack the die, irreparably damaging the detector. For this reason, as shown in Fig 4.9, we placed a cluster of 6 microbumps beneath the pad region (*c* and *d*) to guarantee proper mechanical support during the bonding process. Nevertheless, the portion of top-tier comprised between the bump supports and the edge of the chip remains suspended, forming a cantilever profile that can easily break. The injection of organic material below the corner seems to be a solution to prevent this issue.

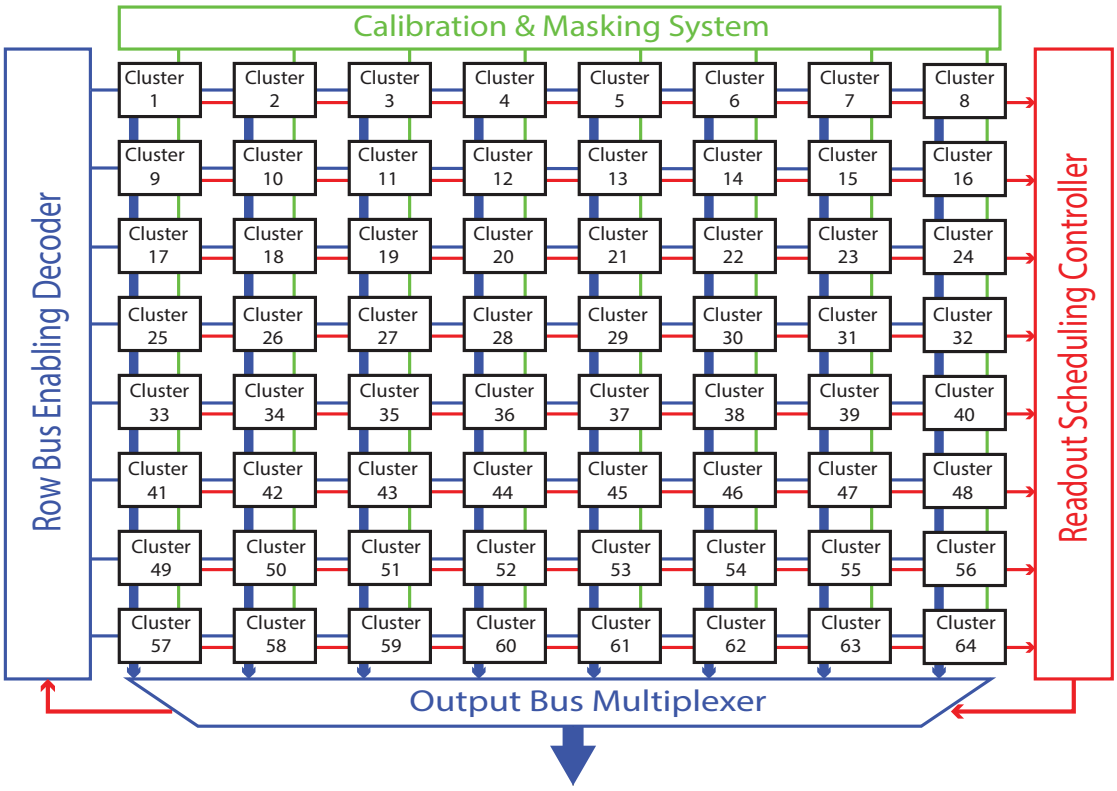


Figure 4.10: MD-SiPM top level core architecture. Each of the two cores composing the chip is independent and is divided in 64 clusters. A random-access scheme readout is used to access the cluster data (*blue*). A calibration and masking system is added to allow the masking of the noisy SPADs and the calibration of the TDCs (*green*). A scheduling system is added to the chip to allow an automatic event-driven readout.

4.3 Multi-Digital SiPM Architecture

In this section, we will discuss the architectural details of this MD-SiPM following a topdown approach. Starting from the complete detector level, we will go down step by step to discuss all the details of the structure.

4.3.1 Core Architecture

The proposed sensor comprises two arrays, or cores, of 64×64 SPADs each. The two cores are identical and independent. Indeed, the SPAD bias, control voltages, and TDC supplies are provided separately for the two arrays. Further, each SPAD has its own TSV 3D connection to an independent pixel circuit located in the bottom tier.
Fig. 4.10 shows a simplified architecture scheme of one of the two cores. The core is segmented into 64 parts, called clusters, organized as an 8×8 matrix. A random-access architecture-based readout system is implemented using a row encoder and a column multiplexer; it allows reading out the data contained in each one of the clusters. The clusters share data (high-

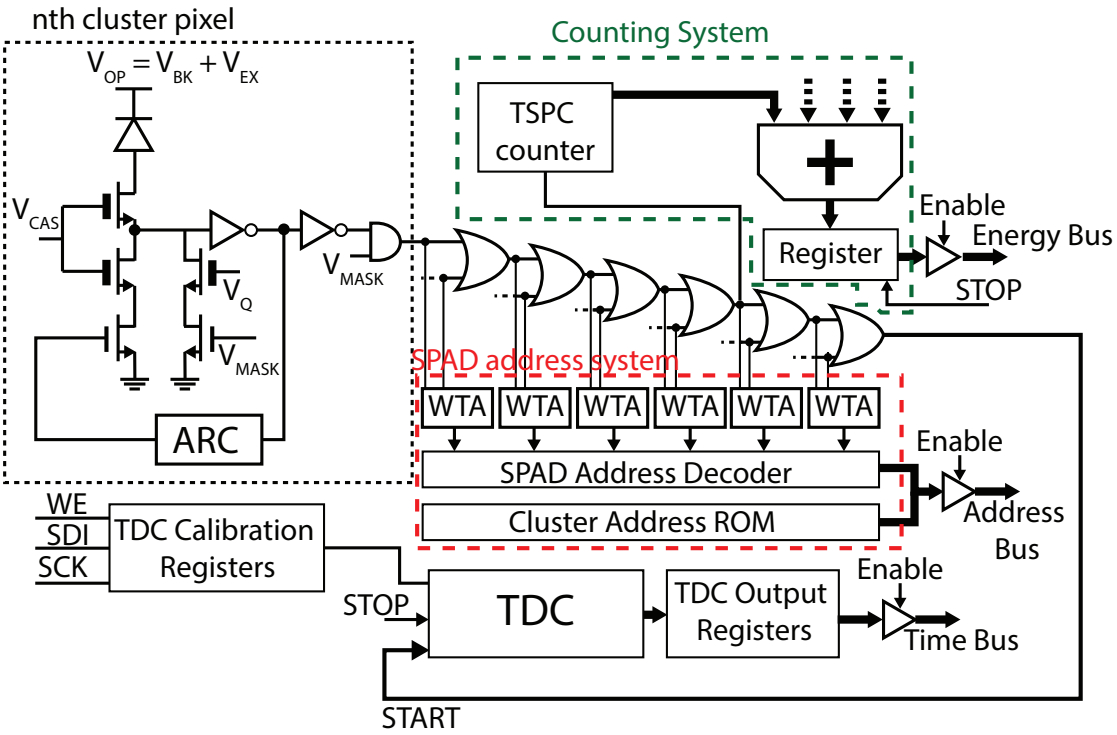


Figure 4.11: Each MD-SiPM cluster is composed of 64 pixels spatially compressed using an OR-tree. The output of the OR-tree is used as a start for a TDC. Along the propagation in the OR-tree, a winner-take-all (WTA) circuit returns the address of the first pixel that fired within a frame. This allows to increase the spatial resolution to the single SPAD level. The number of pulses are counted by a counting system located at the 4th level of the OR-tree. The TDC, photon counting, and address data are sent to a column shared bus through high-impedance buffers.

impedance) buses column-wise at array level, while the data write enable of each cluster is granted by the readout decoder row-wise.
The system allows two separate ways to read the data contained in the chip. First, a user-driven readout allows the cluster address selection through digital inputs. In this way, the user can choose to read out a specific cluster or a sequence of clusters. Second, an event-driven readout can also be used with a scheduling system based on a priority ceiling protocol. A calibration and masking techniques are also integrated on-chip to allow the masking of noisy pixels in the array and to calibrate the TDC in each cluster.

4.3.2 Cluster Architecture

Each cluster contains 64 SPADs arranged as an 8×8 matrix (Fig.4.11). All the SPADs are equipped with a passive quenching and active recharge circuit that allows fine-tuning of the SPAD dead time. The output of each pixel is input to an OR-tree for spatial compression. At the 4th level of the tree, a photon counting system based on true single-phase clock (TSPC)

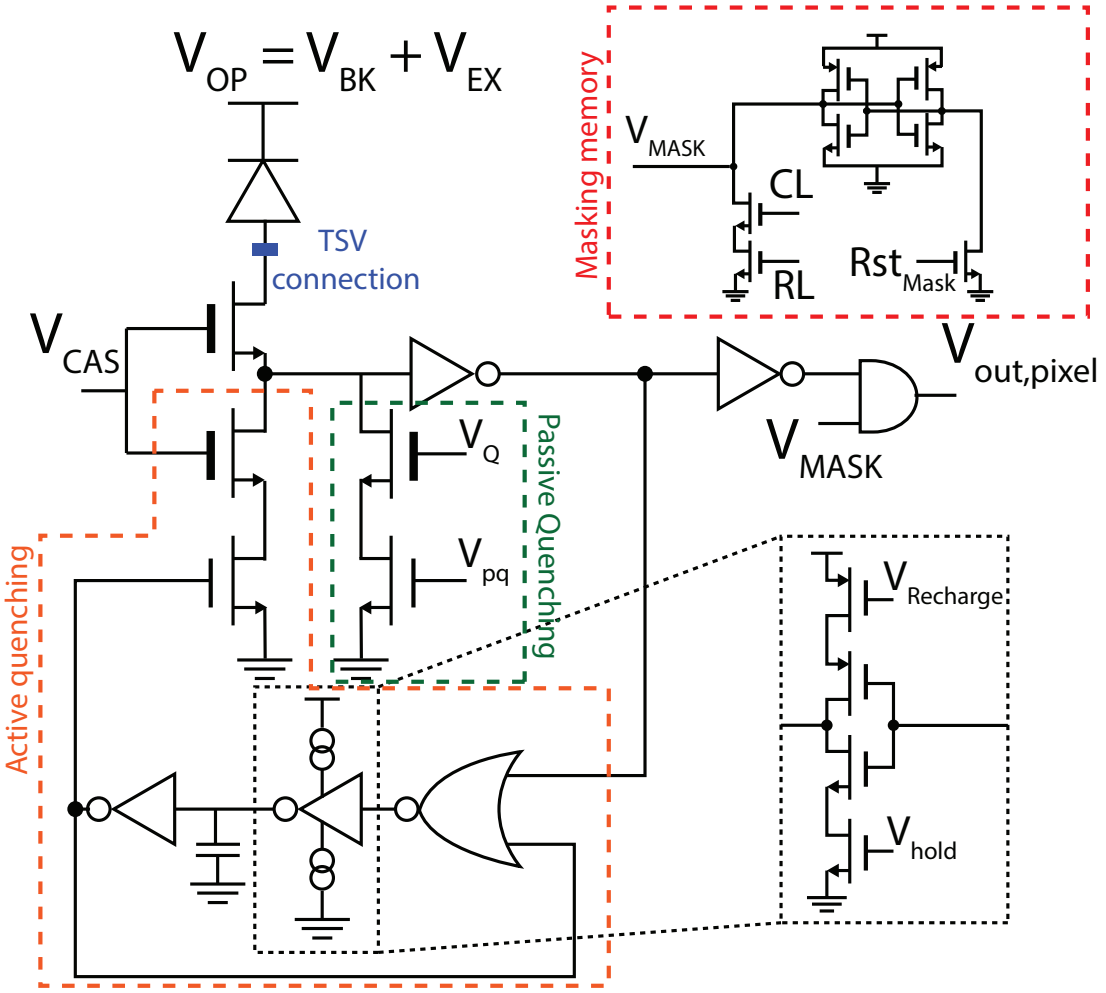


Figure 4.12: The MD-SiPM pixel circuit is based on a passive quenching with active recharge architecture. A cascode structure is used to apply a higher excess bias voltage to the SPAD. The active reset consists of a tunable pulse generator to allow a fine regulation of the dead-time. An SRAM is used to provide a logical masking to the pixel.

counters records the number of pulses generated by the pixels. Each counter takes one of the four branches of the tree as input, and its output is sent to an adder to compute the final sum, providing the number of detected pulses in a frame (see Fig. 4.13). Moreover, a winner-take-all system is implemented along the tree and used to identify the first SPAD that fired in the cluster. Finally, the output of the OR-tree triggers the TDC based on a start and stop architecture. A global stop signal is used to sample the data of the counting, timing, and address systems of the clusters and save them in a memory buffer. As mentioned, the clusters share data buses among the columns, and the data writing process is regulated by the readout system activating the high impedance buffers at the output of the clusters.

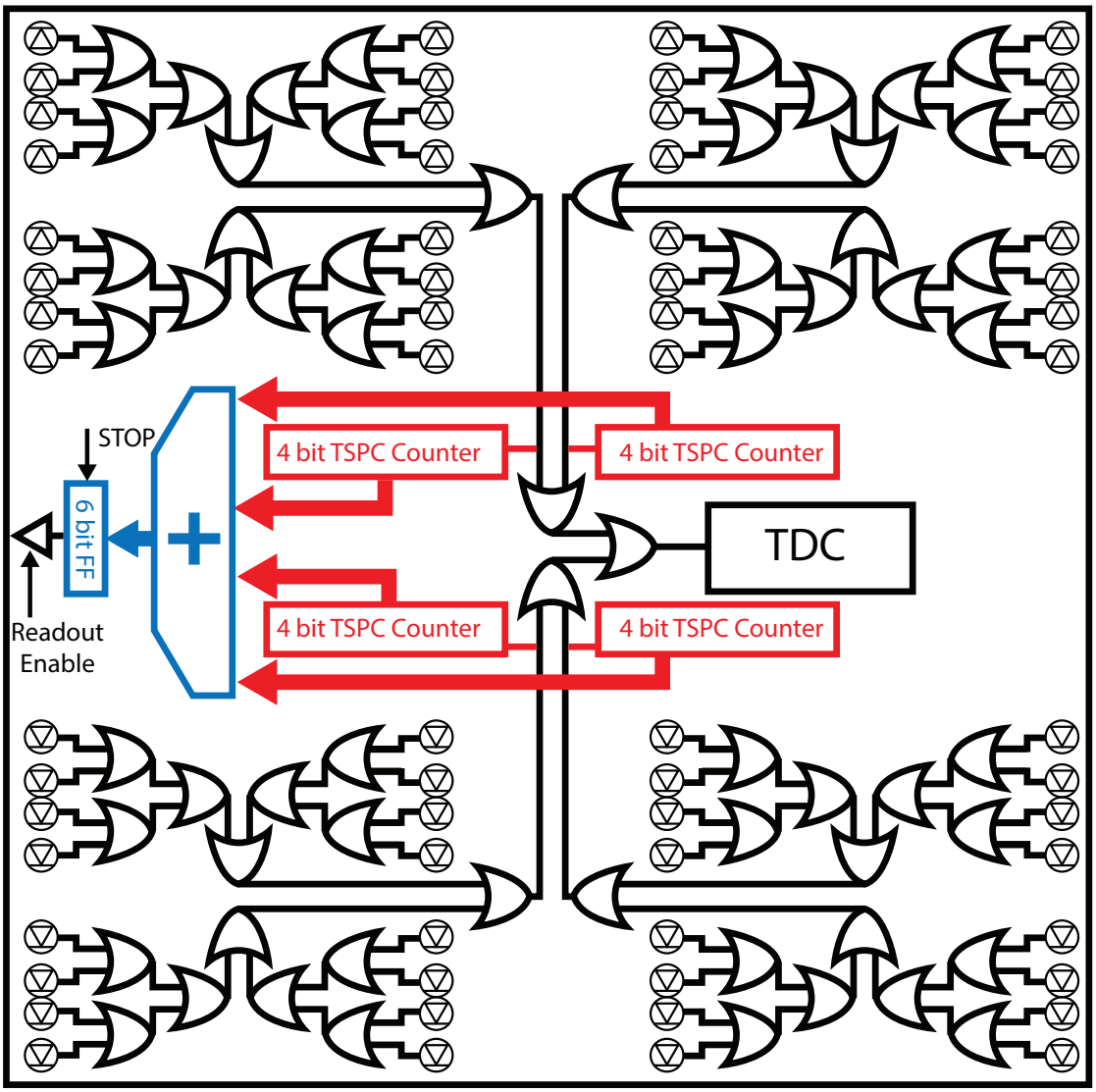


Figure 4.13: MD-SiPM cluster counting system scheme. Four TSPC counters are connected to the 4th level of the propagation OR-tree. The output of the counters are input to an adder for the calculation of the final sum. The counting data are sampled by the STOP signal and stored in a memory, ready to be read out.

4.3.3 Pixel Architecture

The pixel follows the same architecture as described in Section 3.1.1. It comprises a cascode structure to implement passive quenching and to enable a higher excess bias voltage. A feedback loop, which employs a tunable pulse generator, acts as an active reset circuit and allows fine-tuning of the SPAD dead time. Fig. 4.12 shows the schematic of the pixel. In this implementation, an SRAM cell stores the digital masking state of the pixel. If CL and RL are asserted, the SRAM's output becomes 0, blocking the pulse propagation thanks to the AND output gating. The RST_{mask} is a global signal used to reactivate all the pixels in the array.

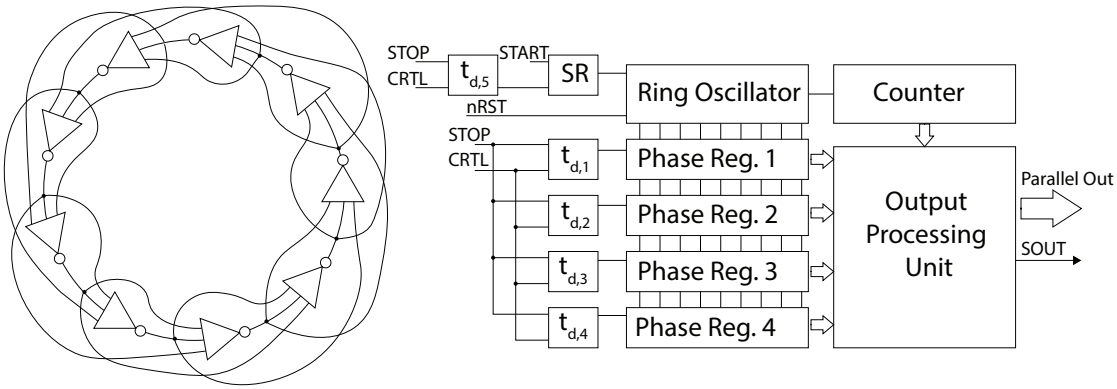


Figure 4.14: The MD-SiPM TDC architecture is based to a multipath gated ring oscillator (*left*) [318]–[320]. The TDC follows a start and stop architecture, where the output of the OR-tree triggers the oscillator and a global STOP signal is used to stop it and sample the phases. Four sets of phase registers are triggered by the STOP signal with progressive delays ($t_{d,1}$<$t_{d,2}$<$t_{d,3}$<$t_{d,4}$<$t_{d,5}$). The four samples allow for linear interpolation and improvement of the resolution.

In contrast to what we have seen in Section 3.1.1, the pixel circuit is here connected do the SPAD anode through a TSV connection. This connection can be seen as a source of additional parasitic capacitance. Indeed, the TSV can be considered, as a first order approximation, equivalent to a cylindrical capacitor where the liner oxide later works as the capacitor's dielectric, while the two plates are represented by the TSV copper core and the silicon substrate, connected to the ground. This additional parasitic capacitance can add several hundreds of femtofarads to the anode node, which needs to be considered during the front-end design.

4.3.4 Counting System

The counting system comprises four 4-bit TSPC counters, connected to the 4th level branches of the propagation OR-tree, that count the number of pulses arriving from the four quadrants of the cluster. The outputs of the four counters are then summed up in an adder to form a final 6-bit word, whose output is sampled by the global STOP signal and saved in a memory buffer. The choice of using four counters placed in close proximity to the tree input was motivated by allowing higher counting capability, reducing the dead time, which is intrinsic to the OR-tree implementation [270]. Moreover, the choice of the 4th level was found to be a good compromise between improvement of count rate and system complexity (linked to the number of counters inserted in each cluster). Fig. 4.13 shows the schematic of the proposed solution.

4.3.5 TDC Architecture

The TDC is based on a multipath gated ring oscillators (MGRO) architecture [318] (Fig. 4.14). Similarly to what was shown in [319], [320], each inverter in the chain has three inputs.

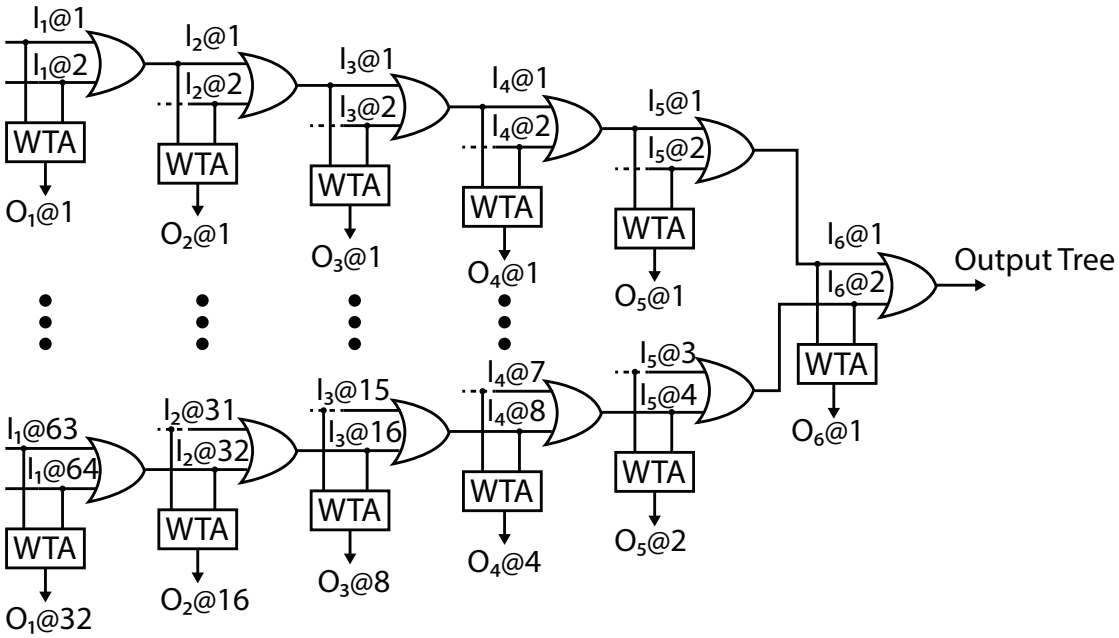


Figure 4.15: The MD-SiPM cluster SPAD address system is composed of WTA cells located at each level of the OR-tree. Defining N as the number of pixels in the cluster, a total of $log_{10}(N)$ WTA cells are present.

So, each one of them is connected to the previous three delay elements in the chain. This feedforward path is used to increase the oscillation frequency of the ring and reduce the LSB. The oscillator is sampled by four sets of registers triggered by the global stop signal. Each register is triggered with a specific delay (sub gate level) to allow a linear interpolation of the four samples and improve the timing resolution. In addition, the delays of each phase register can be independently calibrated through the calibration circuit mentioned before (Fig. 4.11).

4.3.6 SPAD Address System

A system based on winner-take-all cells (WTA) identifies which SPAD fired first within one cluster in a frame and triggered the TDC. Fig. 4.15 shows a scheme of the OR-tree used to propagate the pixel signals. A WTA is inserted in parallel to each tree's OR-gate to identify which of the two OR inputs switched first. Thus, the WTA cell acts as a time-domain comparator, comparing the switching time of its two inputs. This approach is similar to what is implemented in [38] in the so-called decision tree. That approach also uses a cell that acts as a decision-maker to understand which input switched first. However, the WTA cell used here is different. One main difference relies on the propagation of the signal along the tree. Indeed, in [38], the tree propagates only the *winning* signal while blocking the others. In the circuit presented in this section, the decision cell is parallel to the propagation tree. This choice allows propagating multiple pulses per frame along the tree, also allowing to count them. In Fig. 4.16, we can see the schematic of the implemented WTA cell (a). The circuit

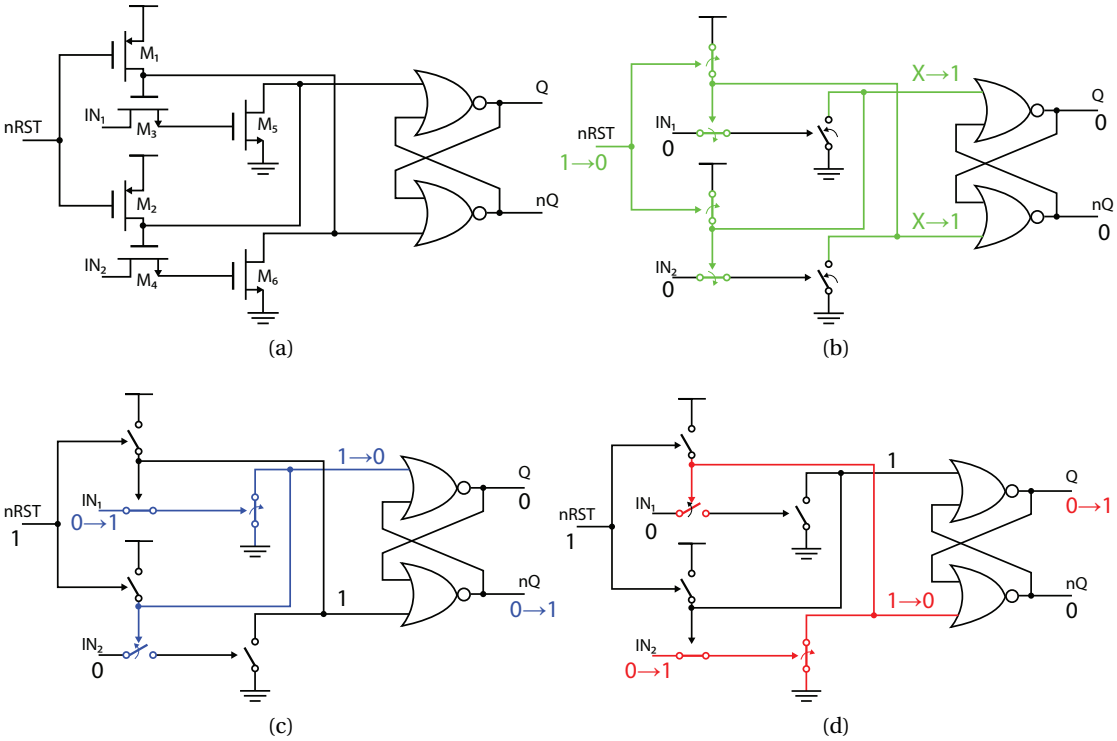


Figure 4.16: (*a*): WTA cell schematic. This cell is based on a precharged logic. When nRST is activated (logic zero), M_3 and M_4 in series with the inputs are activated by M_1 and M_2, which charge the line connected to their gate (*b*). When IN_1 switches before IN_2, M_5 activates, discharging the gate of M_4. This prevents IN_2 from propagating in the cell (*c*). When IN_2 switches first, the gate of M_3 is discharged, blocking the propagation of IN_1 (*d*).

relies on a precharge mechanism. Indeed, when the *nRST* signal goes to zero, M_1 and M_2 activate precharging the nodes at the M_3 and M_4 gates, turning them on. In this phase, the SR latch inputs are set to 1, and as a consequence, both its outputs go to 0 (*b*). When, for example, IN_1 switches to one before IN_2, it activates M_5 that discharges the gate of M_4, turning it off. This mechanism prevents the propagation of IN_2 and the subsequent activation of M_6. M_5 also discharges one of the SR latch inputs, inducing a change in its outputs (*c*). Fig. 4.16 (*d*) shows the opposite case in which IN_2 switches to one before IN_1. In this case, the codification at the output of the SR latch is the opposite.
The output of each WTA cell (O_i@j, i: tree level, and j: branch of the tree at the ith level) can be combined to extract the address of the SPAD that fired first in the tree. In the presented detector, we implemented a multiplexer-based encoder that inputs all the outputs of the WTA cells and provides the 6-bit address of the pixel that fired first. Fig. 4.17 shows the schematic of the multiplexer-based encoder. The logical expressions for each address bit are the shown in the following equations, starting from the most significant bit (MSB):

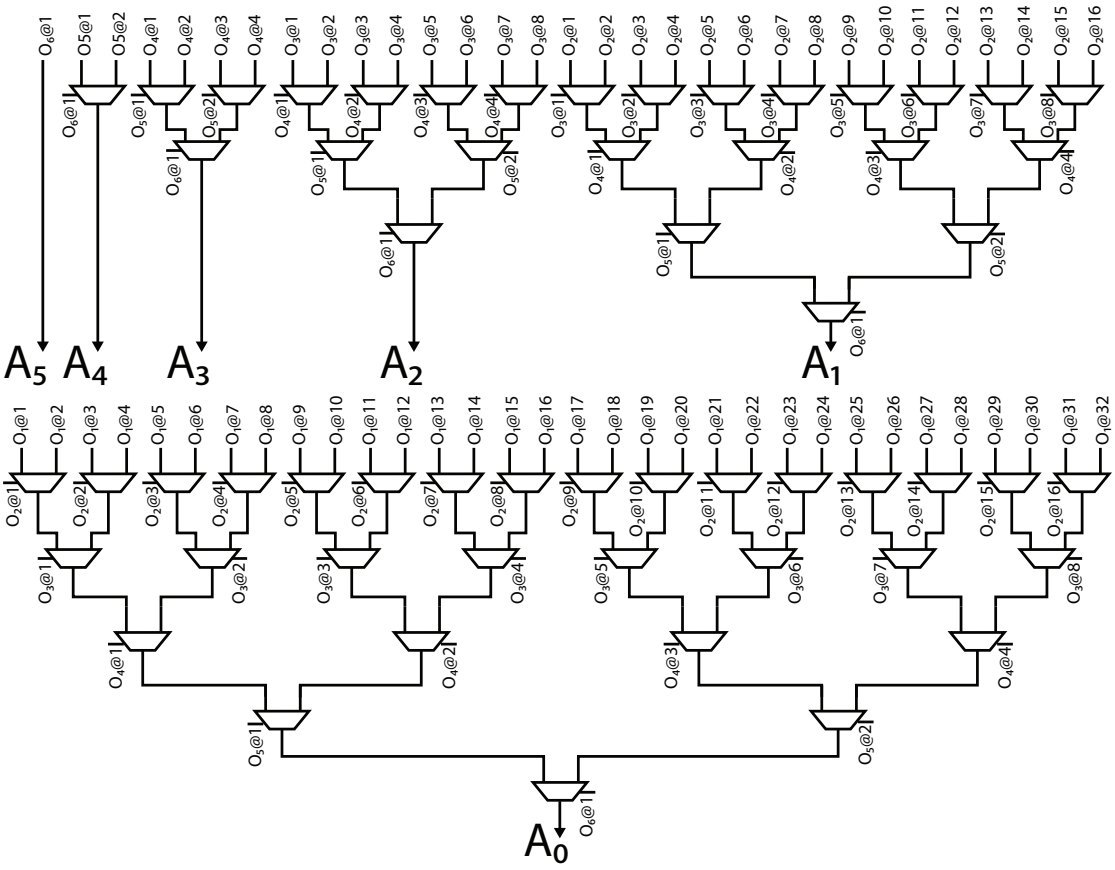


Figure 4.17: The address calculation of the SPAD address system is performed by a multiplexer-based encoder.

$$A_5 = O_6@1 \tag{4.1}$$

where $O_6@1$ corresponds to the output of the last WTA cell in the tree.

$$A_4 = O_6@1 \times (O_5@1 + O_5@2) \tag{4.2}$$

$$A_3 = O_6@1 \times (O_5@1 \times (O_4@1 + O_4@2) + O_5@2 \times (O_4@3 + O_4@4)) \tag{4.3}$$

$$\begin{aligned} A_2 = {} & O_6@1 \times (O_5@1 \times (O_4@1 \times (O_3@1 + O_3@2) + O_4@2 \times (O_3@3 + O_3@4)) + \\ & + O_5@2 \times (O_4@3 \times (O_3@5 + O_3@6) + O_4@4 \times (O_3@7 + O_3@8))) \end{aligned} \tag{4.4}$$

$$
\begin{aligned}
A_1 = {} & O_6@1 \times (O_5@1 \times (O_4@1 \times (O_3@1 \times (O_2@1 + O_2@2) + O_3@2 \times (O_2@3 + \\
& + O_2@4)) + O_4@2 \times (O_3@3 \times (O_2@5 + O_2@6) + O_3@4 \times (O_2@7 + O_2@8))) + O_5@2 \times \\
& \times (O_4@3 \times (O_3@5 \times (O_2@9 + O_2@10) + O_3@6 \times (O_2@11 + O_2@12)) + O_4@4 \times \\
& \times (O_3@7 \times (O_2@13 + O_2@14) + O_3@8 \times (O_2@15 + O_2@16))))
\end{aligned} \tag{4.5}
$$

$$
\begin{aligned}
A_0 = {} & O_6@1 \times (O_5@1 \times (O_4@1 \times (O_3@1 \times (O_2@1 \times (O_1@1 + O_1@2) + O_2@2 \times \\
& \times (O_1@3 + O_1@4)) + O_3@2 \times (O_2@3 \times (O_1@5 + O_1@6) + O_2@4 \times (O_1@7 + O_1@8))) + \\
& + O_4@2 \times (O_3@3 \times (O_2@5 \times (O_1@9 + O_1@10) + O_2@6 \times (O_1@11 + O_1@12)) + O_3@4 \times \\
& \times (O_2@7 \times (O_1@13 + O_1@14) + O_2@8 \times (O_1@15 + O_1@16)))) + O_5@2 \times (O_4@3 \times \\
& \times (O_3@5 \times (O_2@9 \times (O_1@17 + O_1@18) + O_2@10 \times (O_1@19 + O_1@20)) + O_3@6 \times \\
& \times (O_2@11 \times (O_1@21 + O_1@22) + O_2@12 \times (O_1@23 + O_1@24))) + O_4@4 \times (O_3@7 \times \\
& \times (O_2@13 \times (O_1@25 + O_1@26) + O_2@14 \times (O_1@27 + O_1@28)) + O_3@8 \times (O_2@15 \times \\
& \times (O_1@29 + O_1@30) + O_2@16 \times (O_1@31 + O_1@32)))))
\end{aligned} \tag{4.6}
$$

The six bits output of the multiplexer-based encoder is sampled by the global STOP signal and saved in a memory buffer, ready for readout.

4.3.7 Calibration System

The TDC calibration system is based on two shift registers used for column and row enable. If asserted simultaneously, the two signals activate the write enable of the TDC calibration registers. The shift registers have a serial interface composed of a serial clock (SCK - *green* and *violet*), a serial bus for the data (SDI - *green* and *violet*), and an output-enable (OE - *green* and *violet*) that activates the parallel output of the shift register, and a reset signal (RST - *green* and *violet*) to reset the shift register content. The calibration data are then written in the TDC calibration registers through two binary trees. One binary tree is used to send a serial clock (SCK - *red*), while the other one is used to send the serial data (SDI- *blue*). Fig. 4.18 shows a simplified scheme of the described architecture.

4.3.8 Masking System

Similarly to what we discussed in the previous section about the TDC calibration system, the masking system relies on shift registers for its operation. Two shift registers of 64 bits are connected to 64 column lines (CL) and 64 row lines (RL), respectively, forming a mesh. Fig. 4.19 shows a simplified scheme of this architecture. Each pixel embeds an SRAM cell dedicated to masking. The datum in the SRAM is written through two transistors whose gates

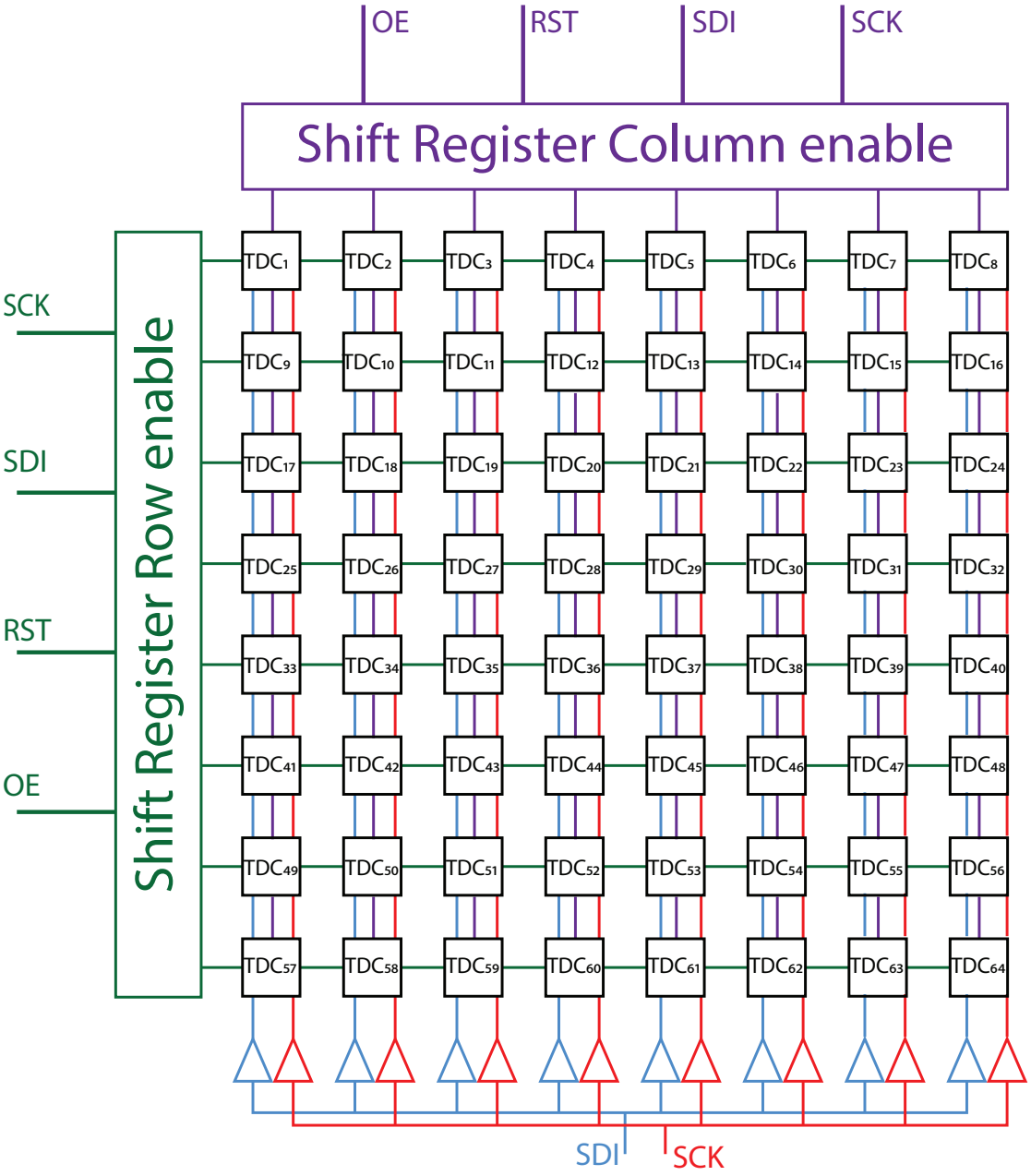


Figure 4.18: Schematic view of the MD-SiPM TDC calibration system. Two shift registers are used to assert the write enable (WE) of the calibration registers of each TDC in the array. The data are then serially sent using binary trees (*red* for the serial clock, and *blue* for the serial data).

are connected to a specific pair of CL and RL signals, respectively. Then, by simultaneously asserting the CL and RL of a pixel, V_{mask} is set to zero, masking the pixel.
Moreover, a global reset signal ($Reset_{Mask}$) is used to reset all SRAMs and reactivate all the pixels.

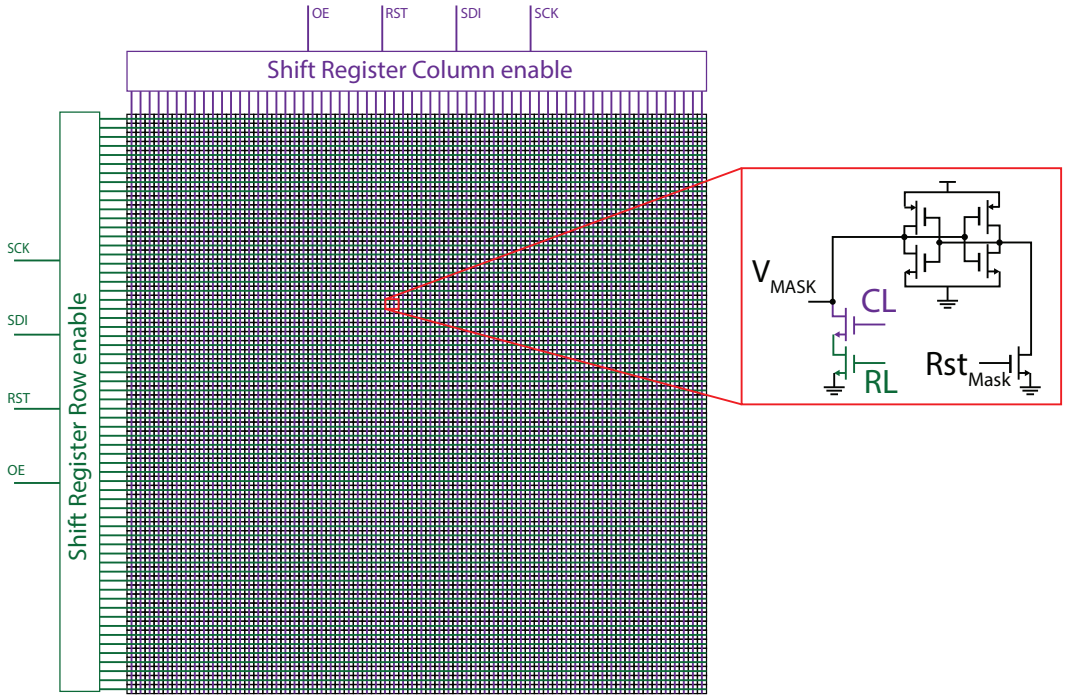
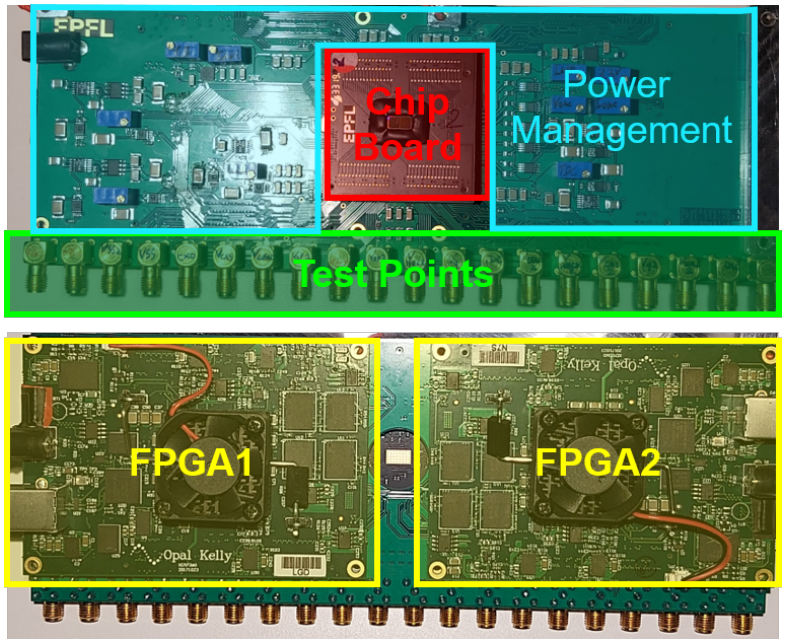


Figure 4.19: Schematic view of the masking system. Two shift registers, one for the columns and one for the rows, are employed to write the 1 bit SRAM embedded in each pixel circuit (*inset*).

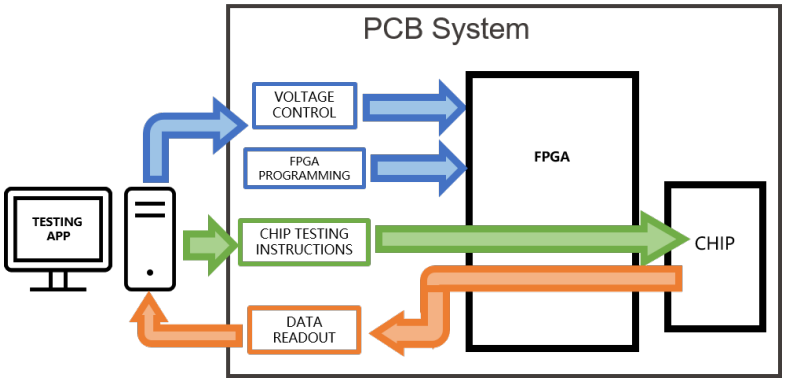
4.4 Testing Setup and Preliminary Results

This section will discuss some preliminary and qualitative results obtained from the characterization of the detector described in the previous section. In this context, a testing system was developed to allow the operation and readout of the 3D-Stacked Multi-Digital SiPM. This system is composed of a motherboard embedding a power management unit that derives all the voltages required by the system from a single 5 V supply. The motherboard connects two FPGA boards (Opal Kelly XEM7360-K160T) used for the readout of the two cores of the chip independently (Fig. 4.20 (*a*)). A graphical user interface (GUI) developed in C++ remotely controls the two FPGAs from a host computer through USB 3.0 communication, see Fig. 4.20 (*b*). The GUI allows setting the operating voltages to the chip and reading out the data contained in each chip cluster.

As a first step to verify the functionality of our system, we analyzed the chip response qualitatively. Fig. 4.21 shows the results of the first test performed, in which we left the sensor in the dark, reading out the counting system output of each cluster with various integration times. This kind of measurement highlights the uniformity of the chip activity and the presence of noisy pixels in the array. In addition, we can also indirectly estimate the process yield by evaluating the number of pixels that fired at least once during the measurement. The integration time was set to 100 ns (*a*), 500 ns (*b*), and 1 μs (*c*), and each measurement was repeated 2^{11}



(a)



(b)

Figure 4.20: (*a*): PCB system for MD-SiPM testing. The system is composed of a motherboard housing two FPGA boards and a chip daughterboard. A power management system and 20 test points complete the structure. (*b*): the testing system diagram. A custom user interface on a host computer controls the communication with the two FPGAs and enables the parameter setting and the data readout of the MD-SiPM.

times (set by the user through the GUI). We can see how the counting results increase with the integration time, as expected. Moreover, using the SPAD address system output, we can build a pixel map that shows which pixel fired for each acquisition frame. When increasing the integration window, the results show a good uniformity of the noise over the array. In the bottom-left part of the intensity maps (Fig. 4.21 $(a)-(c)$), we can clearly identify the presence of some clusters with high activity that can be attributed to a large pixel DCR. This result is confirmed by the pixel maps (Fig. 4.21 $(d)-(f)$) that show some highly noisy pixels in the

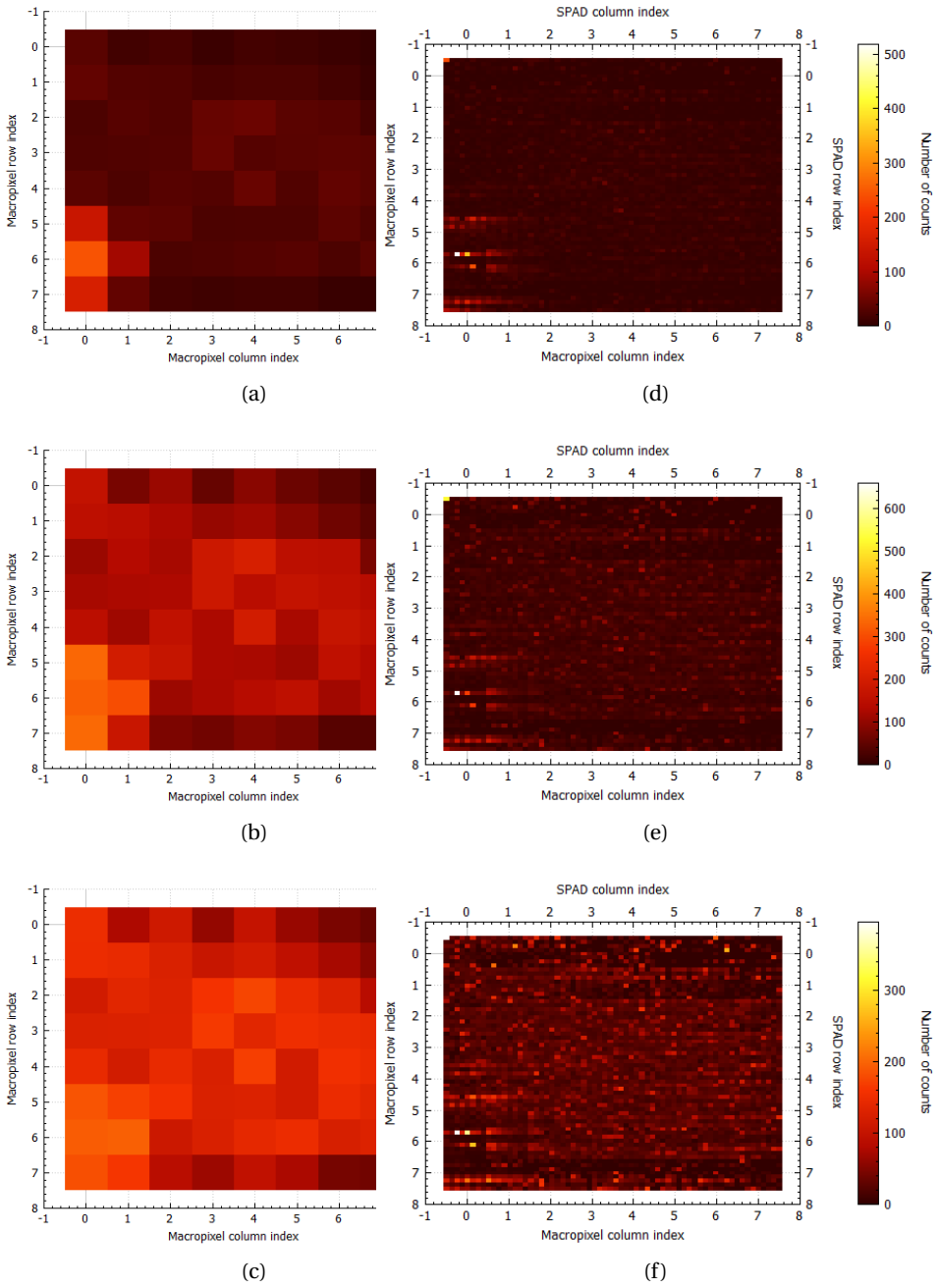


Figure 4.21: 3D-Stacked Multi-Digital SiPM DCR vs. integration time. The output data from the counting system of each cluster are used to build an intensity map, leaving the sensor in the dark. The DCR was integrated for various time lapses: 100 ns (*a*), 500 ns (*b*), and 1 μs (*c*). In addition, at the bottom we can see the pixel map built with the output of the SPAD address system of each cluster at core level. In (*d*), (*e*), and (*f*) the results of the pixel maps are reported for 100 ns (*a*), 500 ns (*b*), and 1 μs (*c*), respectively.

same region of the array. Finally, we estimated the yield of the working pixels to be about 90%. However, this value could be inaccurate because of the different DCR levels of the pixels in

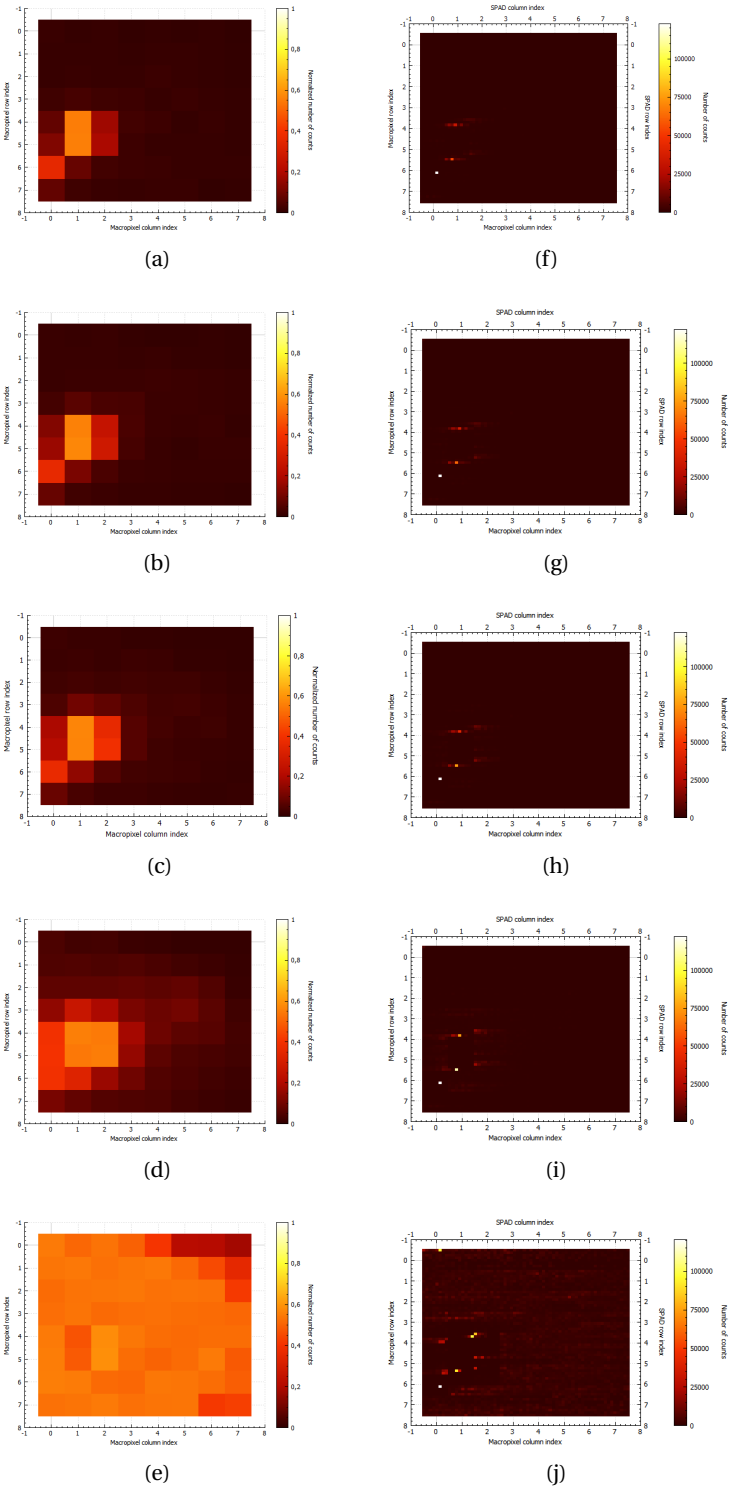


Figure 4.22: The intensity map with a pulsed laser on the array with increasing intensity $(a)-(e)$. The corresponting pixel maps are reported in $(f)-(j)$. We should note how the laser spot size increases with the laser intensity, eventually saturating the array.

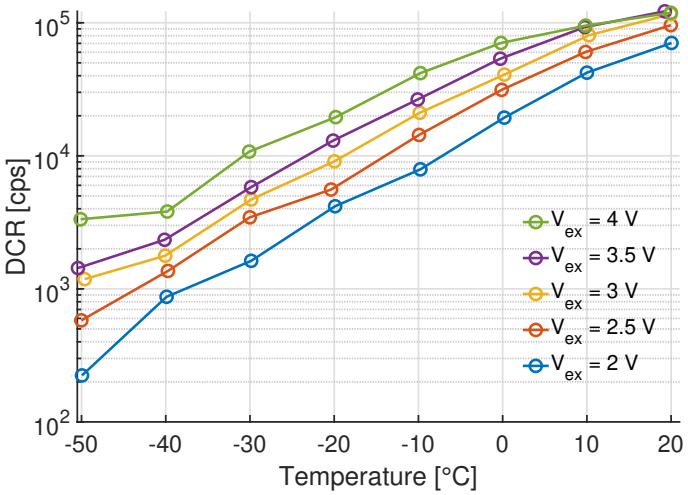


Figure 4.23: Preliminary DCR characterization over temperature of the MD-SiPM SPAD. The graph shows the average DCR of a single pixel of the array over a temperature range from -50°C to 20°C, and for 5 different excess bias voltages. The value of DCR is calculated as the mean of all the 4096 SPADs of one core.

a cluster and the operating principle of the SPAD address system described in the previous section. In addition, the difference in noise can reduce the probability of a quiet pixel to fire as first in an acquisition frame and obtain the address tag from the SPAD address system. For this reason, the measurement should be repeated for a considerable number of cycles to be accurate, becoming impractical and highly time-consuming.

A second test performed to qualitatively verify the functionality of the chip and to verify its response to a light stimulus consists of illuminating the array with a pulsed laser and controlling its power. This test was performed at controlled temperature using a temperature chamber. Since we observed a relatively high noise level in the previous test, we decreased the temperature of the system to -50°C in order to mitigate the thermal noise contribution and check if it was the main one in the SPAD used. Fig. 4.22 shows the intensity maps $(a)-(e)$ and the pixel maps $(f)-(j)$ obtained. Moving from the left to the right of Fig. 4.22 the laser intensity increases and we can see how the laser spot becomes more visible until it saturates the entire chip. Also in the pixel maps we can highlight the position of the laser spot and how it enlarges increasing the laser intensity. All these maps were built with 2^{11} acquisition frames and 100 μs frame integration time time.

Once we verified the basic functionality of the chip, we moved to a quantitative characterization of the DCR of the detector. This experiment was performed at several temperature points, by using the same temperature chamber, and several bias points. Figure 4.23 summarizes the results of the average DCR of a single SPAD in the array as a function of temperature, from -50°C to 20°C, and from 2 V to 4 V excess bias voltage. As it can be noted in the graph, the noise presents a strong temperature dependence decreasing of about three orders of magnitude

over a ΔT=70°. This confirms that the thermal generation is the dominant contribution of noise of these SPADs and that lowering the temperature can be particularly beneficial. In addition, we can note how the DCR does not increase abruptly with the excess bias voltage, confirming that the virtual guard ring of the device is operating correctly. Finally, observing the V_{ex} = 4 V curve, we can note a saturation effect on the counts. This effect is probably due to the count losses caused by high noise of the SPADs and the dead time of the OR-tree that propagate the pulses.

4.5 Discussion

In this chapter, we discussed the structure and architectural details of our 3D-Stacked Multi-Digital SiPM. This sensor is the first demonstrated large-scale SPAD array implemented using the 3D FSI approach. The FSI approach makes this sensor suitable for applications where the wavelength range of interest focuses on the blue and NUV (< 500 nm), such as bioimaging, medical, and high energy physics. This sensor was fully integrated in a 180 nm CMOS technology node. However, since the two tiers are decoupled, future implementations can rely on integration heterogeneity (i.e., different and optimized technology nodes for the two tiers).
We presented and briefly discussed the testing platform developed to semi-automate the test of this detector, based on custom system-on-board and GUI. We discussed the system's basic functionality, presenting the results of some qualitative tests to assess the presence of potential problems in the system operation. The estimation of connection yield was addressed indirectly by analyzing the pixel activation. This information is made available by the output of the SPAD address system of each cluster by acquiring multiple frames and building a pixel map. We also verified the response of the detector to a light stimulus, such as a laser pulse. These results highlighted the correct behavior of the detector, identifying the presence of the laser spot on the intensity map of the array.
Finally, we conducted a quantitative measurement of the DCR of the device, also showing its temperature dependency. The relatively high DCR value shown by the device at room temperature can suggest that cooling is beneficial. Moreover, it suggests that changes in the SPAD design could be needed in a future implementation. Further characterization steps are currently ongoing to provide more accurate quantitative results.

5 Case Study: Scintillator-Based Module Optimization

In this chapter, we will focus on the problem of light extraction from inorganic scintillators [68]. This problem is one of the limiting factors in terms of performance for the PET modules. For this reason, a PET module structure analysis and an optimization of its optical interfaces is crucial to improve the overall system performance. Here, we analyze, combine and compare several conventional and novel light extraction techniques applied to two conventional inorganic scintillators, namely LYSO and BGO. While LYSO is the scintillator of reference in Time-of-Flight PET (TOF-PET), BGO has seen renewed interest in the recent past, primarily due to the possibility of detecting the small but prompt Cherenkov emission component thanks to impressive progress in SiPM performance [321]. BGO features a higher attenuation coefficient and photoelectric fraction than LYSO and lower cost but suffers from lower light yield and a slower scintillation component. However, it is highly transparent down to 310 nm and has a high refractive index of about 2.15 at the peak emission wavelength (480 nm), which are expected to make it a good Cherenkov radiator.
The characterization of all the proposed solutions relied on the use of a PMT-based setup, planned as an intermediate step before moving to dSiPM-based measurements in the long term. The use of a PMT is justified by its good linearity with respect to the input light intensity, low cost, and ease of use. The availability of a system that was simple and easy to debug was extremely important to speed up the prototyping phase linked to the technology development of our novel light extraction solution.
The chapter starts with an introduction to the general concepts of using inorganic scintillators for gamma detection. Next, we will discuss conventional techniques to improve light extraction. Then, we will discuss novel approaches based on photonic crystals (PhC) and distributed Bragg reflectors (DBR). Finally, the experimental setup implemented to characterize the performance of our solutions is presented and discussed together with the testing results, at the end of the chapter.

The work presented in this chapter is based on [68].

5.1 Introduction

Scintillating materials are commonly used in high-energy physics and medical imaging applications because of their capability to downconvert high-energy radiation into optical photons. Scintillators are typically instrumented with a photosensor coupled to their sensitive surface. Originally, PMTs were the most used photodetectors to read out the scintillation light, but nowadays, silicon-based devices such as Silicon Photomultipliers (SiPMs) have moved into a prominent position. This trend is due to their more compact structure, lower supply voltage, and robustness to magnetic fields (see Section 1.1.1 and 2.3.3 for more details). The light sensitivity of both PMTs and SiPMs does usually not exceed 30% at short wavelengths (e.g., 300 nm) in commercial devices [67], and peaks at around 50-60% in the visible range for SiPMs [69]. The amount of light generated during the scintillation process in inorganic scintillators is in the range of 8k-60k photons/MeV (e.g., the LYSO light yield is around 30k photons/MeV) following a specific wavelength spectrum, typical for the material [65], [66].

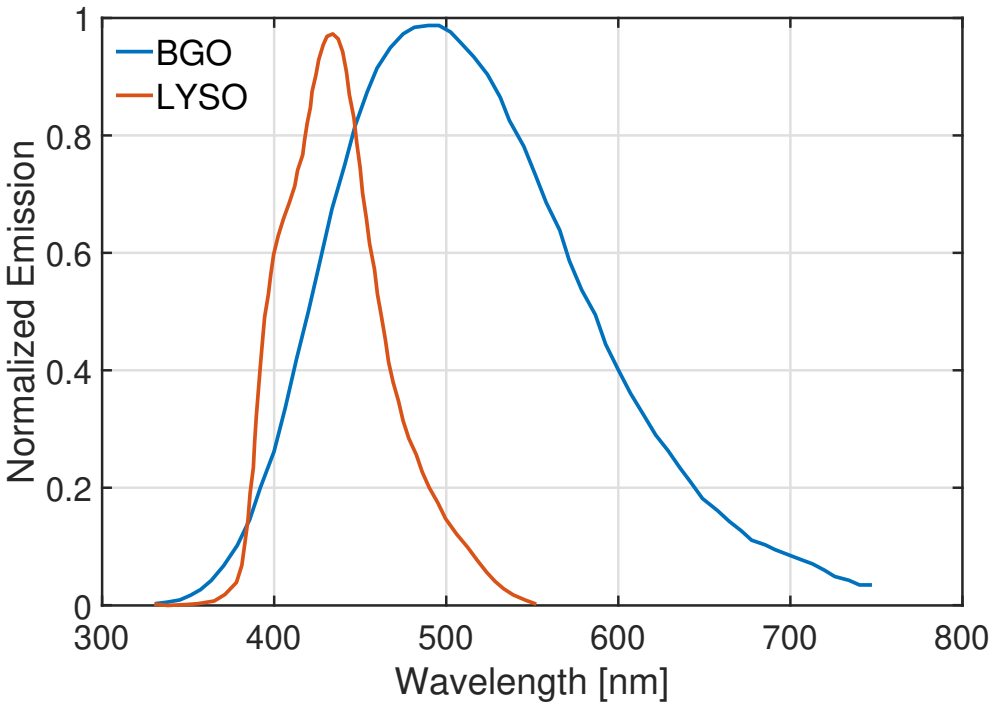


Figure 5.1: Emission spectra for LYSO and BGO crystals. The emission value are normalized to the peak. As it is possible to see, LYSO emission peaks at about 420 nm and the BGO spectrum peaks around 480 nm.

Fig. 5.1 shows the emission spectra for LYSO and BGO. However, in standard configurations, only a tiny percentage of this light reaches the photodetector. Extracting as much light as possible from the crystal becomes thus crucial, given that both energy and time resolution depend strongly on the amount of detected light. Indeed, extracting more light enables a more accurate estimation of the energy deposited in the crystal by the incoming radiation, allowing better discrimination of events that underwent Compton scattering. Moreover, the coincidence resolving time (CRT), which represents the FWHM of the distribution of the time difference between two events in coincidence, has been demonstrated to be proportional to the inverse square root of the number of detected photons [75], [76].

Several phenomena limit the amount of light extracted from the crystal and then detected by a photodetector. These causes include, at the scintillator-photodetector interface (Fig. 5.2), the actual light collection efficiency, which determines the amount of light that reaches the active area of the photodetector, as well as total internal reflection and Fresnel losses. In addition, light absorption in the scintillator itself affects the number of photons that are lost along their path to the output surface of the crystal, with the absorption length reducing at shorter wavelengths [65]. Moreover, if not covered by any material, the scintillator can let light escape through its lateral surfaces, thus losing a significant amount of optical photons (Fig. 5.3, *a*). A solution to overcome this problem is to confine the scintillation light inside the scintillator and

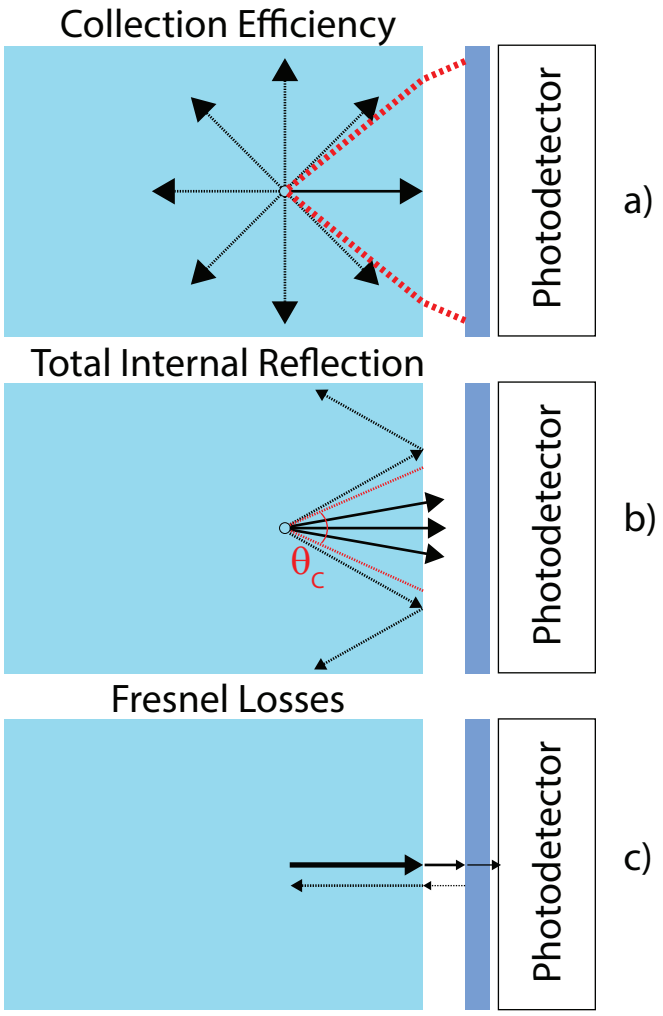
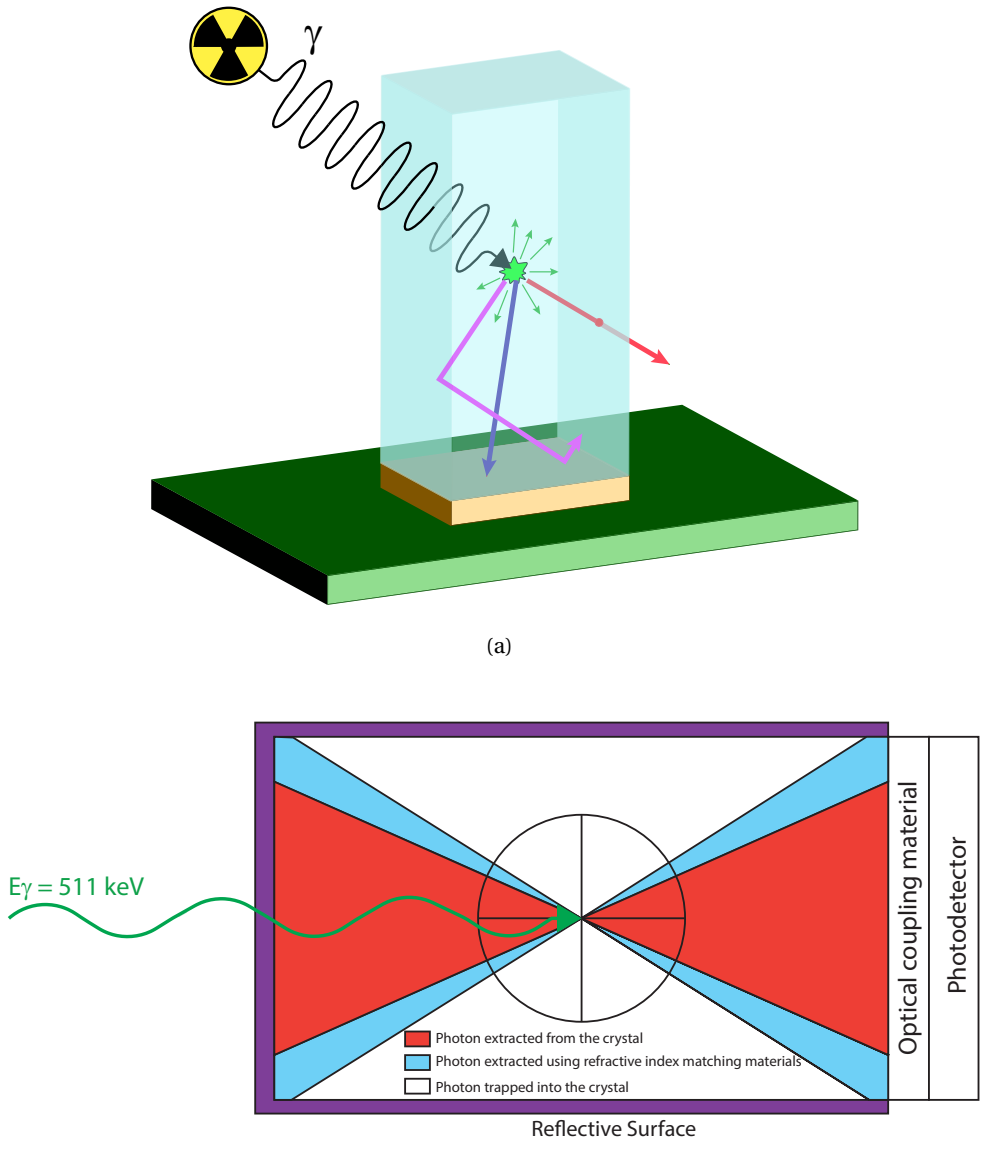


Figure 5.2: Main sources of light losses at the scintillator (*light blue*)-photodetector interface: a) light collection efficiency ($\propto L^2$, with L being the distance between the light source and the photodector); b) total internal reflection ($\theta_{angle} = sin(n_{crystal}/n_{air})$); c) Fresnel losses ($\propto \Delta(n)^2$).

redirect the optical photons toward the output surface. One possibility is to apply a specular reflector to the lateral surfaces (Fig. 5.3, *b*). However, in such a scenario, most of the light generated during scintillation cannot be extracted and remains trapped inside the crystal until it gets absorbed. This phenomenon is due to the relatively large difference of refractive index between the scintillator (e.g., 1.8 for LYSO and 2.15 for BGO at the peak emission), the air (n = 1) crystal and the photosensor window (typically 1.4-1.5 for glass) coupled on its surface. Optimization of the optical coupling is possible in order to mitigate this effect, e.g., employing refractive index matching compounds such as greases or waxes. The latter can increase the critical angle of total internal reflection, allowing the extraction of more light (Fig. 5.3, *b*). Another possibility is the use of diffusers on the scintillator surfaces, which redirect the impinging photons after each reflection (Fig. 5.4 *a* and *c*). This solution decreases the total



(a)

(b)

Figure 5.3: Simplified ionizing radiation detection scheme. (*Top*) The scintillator, attached to a photodetector, is irradiated with gamma rays. The interaction between the ionizing particles and the crystal structure generates optical photons isotropically. The latter can reach directly the output surface and be extracted from the crystal (*blue arrow*). Some photons will however be refracted at the other crystal surfaces and eventually escape from the crystal (*red arrow*). Other photons will undergo total internal reflection and be trapped within the crystal (*purple arrow*). (*Bottom*) Only a limited portion of photons is going to be extracted when employing specular reflective surfaces (*violet structures*). The situation improves by optimizing the optical coupling [68].

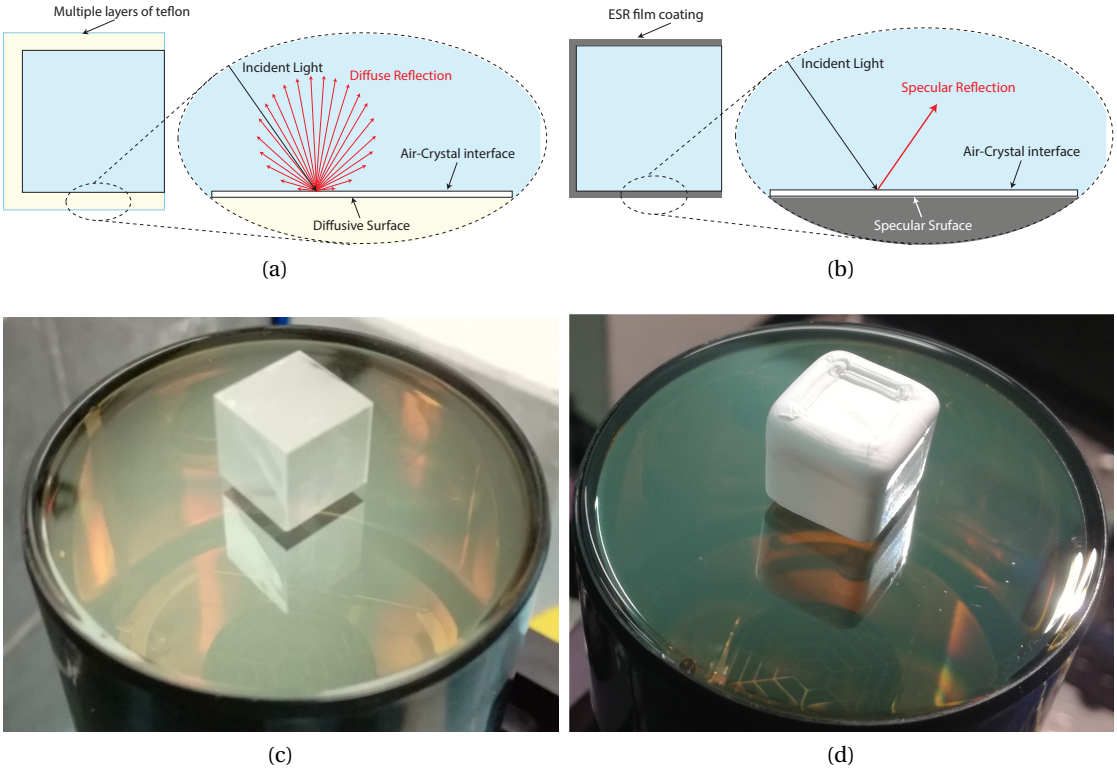


Figure 5.4: Typical scintillator wrapping examples. *Left*: Teflon (diffusive) wrapping vs. *Right* ESR (specular) wrapping. *Top row*: Schematic representation, *Bottom row*: actual crystals attached to the window of a PMT.

internal reflection probability at the output surface, with a higher chance of the scintillation photons being extracted at the first interaction. However, such an indirect redirection scheme does not prevent internal absorption from occurring. A more sophisticated approach has also been proposed, which relies on the redirection within the escape cone of the photons, which impinge on the lateral surfaces through diffraction gratings whose periodicity has been specifically designed [322]. However, this approach has not been implemented yet.

5.2 Conventional Light Extraction Enhancement Techniques

In current radiation detectors based on scintillators, one of the most commonly used techniques to overcome the lateral light losses consists in the use of Teflon wrapping (Fig. 5.4 *a* and *c*). This material is cheap and readily available. It is typically applied on five of the six surfaces of the crystal and acts as a light diffuser, whereby its properties also depend on the thickness of the applied layer [323], [324]. A detailed study of the influence of Teflon wrapping on scintillators can be found in [325].
The adoption of this solution, using a sufficiently thick layer of Teflon (larger than 100 μm) on the sidewalls to ensure enough reflectivity (> 80%) [323], [324], can dramatically improve the

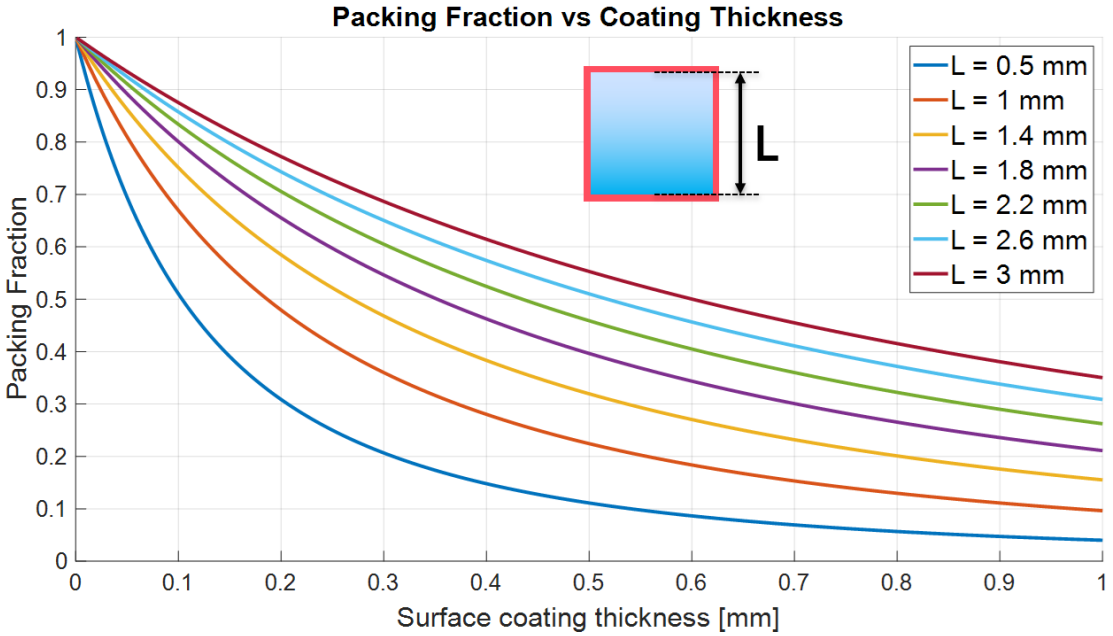


Figure 5.5: Coating thickness impact on the packing fraction. The coating thickness becomes non-negligible when reducing the crystal size (L), thus affecting the overall efficiency of the system.

light extraction efficiency of the system, even if it appears to be unsuitable for applications in which the crystal is pixelated in miniaturized needles. This scenario is, for example, the case for preclinical PET scanners [326], which typically rely on small pitch (sub-millimeter) scintillators to increase the spatial resolution. Indeed, as shown in Fig. 5.5, the efficiency of the system is quickly reduced when the crystal size becomes comparable to the size of the coating due to the reduced packing fraction (ratio of crystal cross-section without coating to cross-section with coating).

This geometric effect can be mitigated by changing the scattering behavior and resorting to other materials such as the enhanced specular reflector (ESR) [327], which features a typical thickness of 65 μm; its analysis is reported in [328]. The ESR is a film of dielectric material (Fig. 5.4 *b* and *d*) with high reflectivity over a vast spectrum and represents the current gold standard when reflective coatings are applied on scintillator surfaces. However, its reflectivity drops at 400 nm, limiting its use in those cases when the emission spectrum of scintillators is shifted toward the blue-near ultraviolet (NUV) range (Fig. 5.1). Fig. 5.4 illustrates a visual comparison between the two reflectors mentioned above.

However, the use of reflectors alone is not sufficient. Indeed, very few systems are air-coupled - almost all practical PET implementations rely on optimizing the optical interface between crystal and detector to mitigate the light losses at the photosensor interface. This coupling is typically achieved by applying a refractive index matching material to smooth the refractive index (n) transition and increase the critical angle, thereby reducing the amount of light that undergoes total internal reflection and Fresnel losses. Several solutions are available with

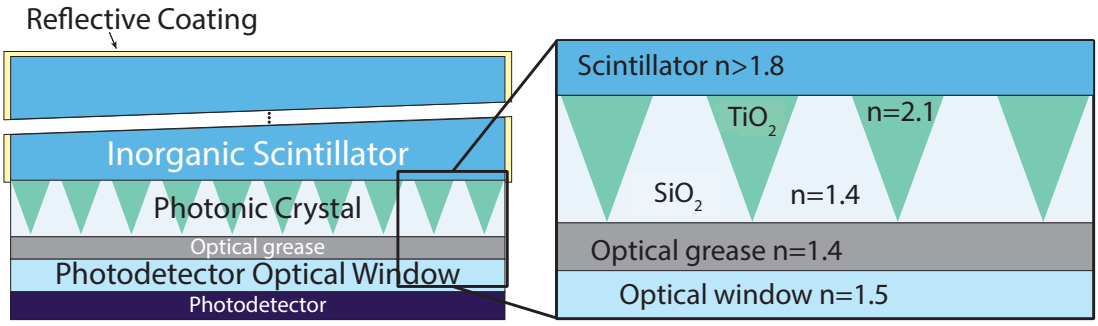
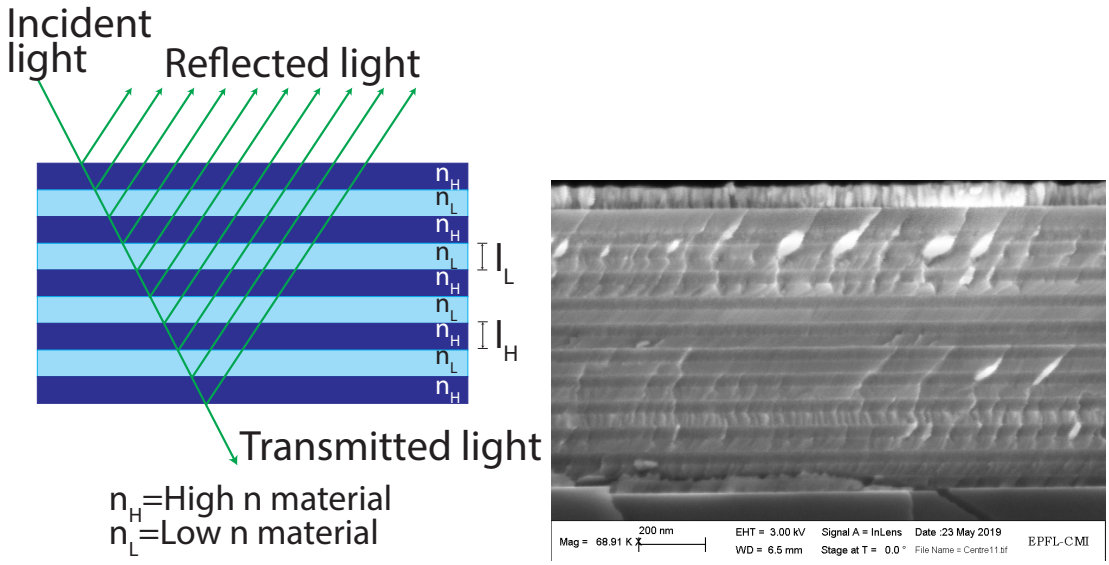


Figure 5.6: Proposed photonic crystal solution: the scintillator is covered by reflective material and the novel PhC structure is implemented on the output surface. The PhC is composed of titanium oxide nanocones surrounded by a buffer layer of silicon oxide with a refractive index of about 1.4. The solution allows a stable coupling with the photodetector by means of a refractive index matching material, such as grease.



(a) (b)

Figure 5.7: Distributed Bragg reflector (DBR) scheme. *Left*: Schematic overview. The DBR structure is composed of a periodic alternation of layers with different index of refraction (n_H and n_L). The number of layers and the difference in refractive index between the layers have an impact on the mirror performance. *Right*: SEM cross-section of a DBR implemented in the EPFL's clean rooms at the Center of MicroNanotechnology (CMI). The layer stack is clearly visible. Scale bar: 200 nm.

different refractive indices depending on the target scintillator. The compounds being used (especially during prototyping and research) are often waxes or greases because they are practical to employ and guarantee a reasonable degree of system stability. These compounds also provide high transmittance at a short wavelength and a refractive index typically between 1.4 and 1.6.

5.3 Novel Light Extraction Enhancement Techniques

Recently, the use of nanostructures applied to the output surface of the scintillator, commonly called photonic crystals (PhC), has been investigated for PET applications [329]–[331]. Photonic crystals are periodic structures of hundreds of nanometers in size directly built on the output surface of the scintillator. Their principle of operation relies on scattering the light photons at the interfaces of materials with different refractive indexes. The scattered waves interact constructively or destructively (depending on the structure) with each other. The coating acts as a diffraction grating and leads to higher extraction efficiency, *de facto* overcoming the limitations of total internal reflection described by Snell's law [330]. Initial developments have relied on e-beam lithography, which is accurate but usually slow and costly, moving then to nanoimprint with the potential of large area coverage, process simplification, and cost reduction; self-assembly techniques have also been employed more recently. These structures rely on high refractive index contrast between the nanostructure and the surrounding medium, leading to interesting optical properties, especially if applied on the surface of a high refractive index material. Indeed, PhCs have shown to be capable of increasing the amount of light extracted from inorganic scintillators, albeit to varying degrees and under often quite different experimental conditions, such as for LSO [330], LYSO [332]–[335] (self-assembly and e-beam, and nanoimprint), GYGAG and SrI_2 [336] (nanoimprint), CsI(Na) [337], [338] and BGO [339] (self-assembly). A more comprehensive description of the PhC physics is reported in [330], [331]. Their use has also been suggested by our group in combination with microlensed SiPMs [340], whereby the two components would work in synergy, with the PhCs reducing the angular spread of the scintillation light, thereby increasing the efficiency of the microlenses.

However, the PhC structures have been mostly optimized and tested using air coupling with the photodetector optical window [334], in some cases with techniques that do not scale to large patterning areas. In addition, their structure is extremely fragile and not suitable to be used easily and reliably in standard configurations, in which a refractive index matching compound is employed between the crystal and the sensor's optical window. Indeed, the shear stress created during the coupling process can easily damage the nanostructure, compromising its performance and measurement repeatability. Moreover, the PhC architecture, as mentioned, is susceptible to the refractive index contrast, and the use of optical grease can compromise the effectiveness of the solution [330]. To overcome this issue, in this study, we elaborated a nanoimprint-based solution that uses a hybrid organic/inorganic (O/I) silica sol-gel buffer layer (n=1.46) which encapsulates the nanostructure. This approach ensures structural robustness and maintains the PhC performance unvaried (Fig. 5.6). In addition to the core work on improved photonic crystals, the use of thin-film coatings, such as distributed Bragg reflectors [341]–[343] and metal coatings (Al and Ag), was explored to overcome the aforementioned limitations on the packing fraction. A DBR (Fig. 5.7) is a very thin (order of 1 μm) periodic structure formed by alternating quarter-wavelength stacks of dielectric layers. The reflectivity spectral width depends on the refractive index contrast of the materials used and is tunable. A DBR typically contains a large number of layers with a high refractive index

contrast. It can be used to achieve nearly total reflection within a range of wavelengths, and it is thus employed as reflector in waveguides and optical fibers, presenting extremely low losses compared to ordinary metallic mirrors [344]. Each interface between the two materials contributes a Fresnel reflection. At the design wavelength, the optical path length difference between reflections from subsequent interfaces is half the wavelength; in addition, the reflection coefficient amplitudes for the interfaces have alternating signs. Therefore, all reflected components interfere constructively, which results in a strong reflection.

The reflectivity achieved is determined by the number of layer pairs and by the refractive index contrast between the layer materials. However, because of their stratification, the reflectivity of the DBR turns out to be highly angular dependent. Indeed, the reflectivity changes significantly for incidence angles above 30°. It is worth mentioning here the possibility of using the so-called perfect or omnidirectional mirrors, which reflect light whatever the angle of incidence and polarization, assuming that the average refractive index of the DBR is higher than the one of the scintillating crystal [345]. This approach has however not yet been used for scintillating crystals. Finally, it is worth noting that DBRs can be combined with metal layers to enhance the reflectivity spectrum, especially at long wavelengths, while at the same time compensating the loss of reflectance below 450 nm typical of metal mirrors.

5.4 Coating Fabrication

In this section, we will discuss the fabrication of the coatings implemented in this work. First, we will discuss the photonic crystals fabrication on the scintillator surface. Then, the deposition of DBRs.

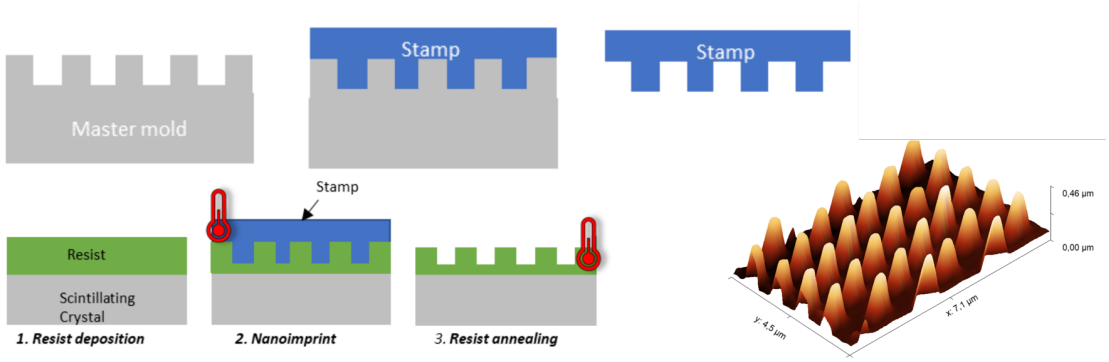
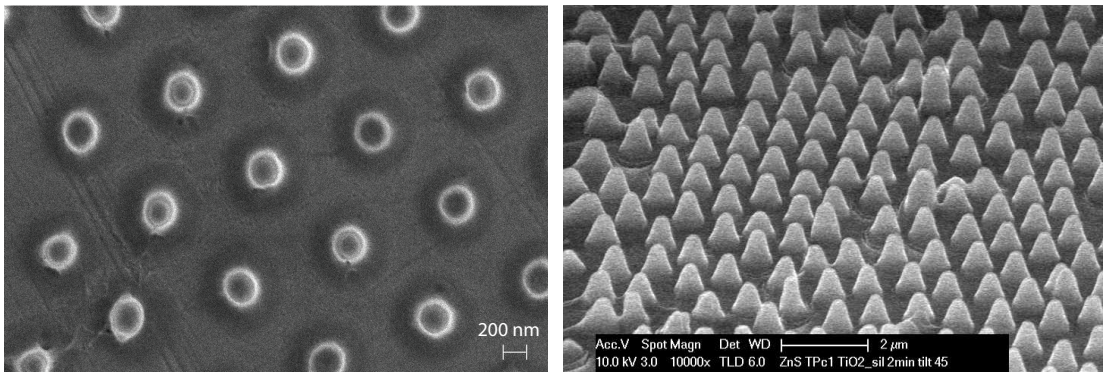


Figure 5.8: Simplified nanoimprinting process steps for pattern implementation (*left*) (for details see text). Atomic Force Microscope (AFM) image of a small area of the implemented pattern (*right*)[346].



(a) (b)

Figure 5.9: SEM photonic crystal nanostructure images before the implementation of the buffer layer. *Left*: top view (scale bar: 200 nm), *Right*: side view (scale bar: 2000 nm).

5.4.1 Photonic Crystals

In this work, PhCs were produced by Nanoimprint Lithography (Fig 5.8 *Left*). The patterns are reproduced on a resist using *soft stamps*, replicated from master molds made by different lithography techniques, such as e-beam, photo, laser interference, or colloidal lithography. In the present case, the master mold was made by colloidal lithography followed by Reactive Ion Etching (RIE) [347]. The soft stamp is then placed in contact with the resist and put under pressure. The nanoimprint process was used here with a TiO_2 sol-gel resist, which was deposited by spin-coating at 5500 rpm on the scintillator surface before being patterned via nanoimprint [348]–[352]. The patterned layer is then annealed at a temperature optimized to match the required height and refractive index (n=2.15). The pattern has a periodicity of 1000 nm and features cones of 560 nm height and 300 nm basal diameter as shown in Fig. 5.9. The pattern is encapsulated in a layer of silica sol-gel with a refractive index of 1.46, matching that of the used optical grease. The hybrid organic/inorganic encapsulation layer was deposited via spray coating using an airbrush spray gun with a 0.5 mm nozzle. The coating was first dried at room temperature, then annealed at 150°C for 30 minutes to ensure its mechanical properties. This encapsulation procedure also avoids the formation of air gaps, which could compromise the optical performance between the patterns' bottom and the grease if the latter were applied directly on top of the PhC. A schematic view of the implemented structure has already been shown in Fig. 5.6. Modifications can be introduced, if required, to the pitch of the periodic structures and their shapes. This flexibility allows structural optimizations, which can be combined with different refractive index materials and wavelength ranges tuned to the applications at hand.

5.4.2 Distributed Bragg Reflectors

Two different DBR structures were implemented in this work (Fig. 5.10). The first structure is shown on the *left* of Fig. 5.10 and consists of a conventional mirror fabricated with 13

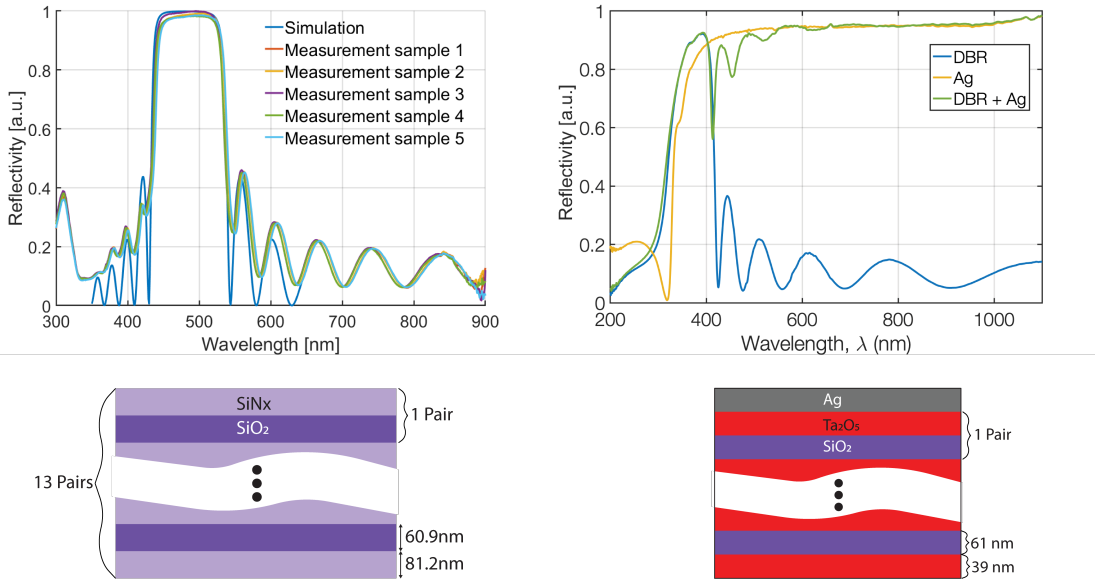
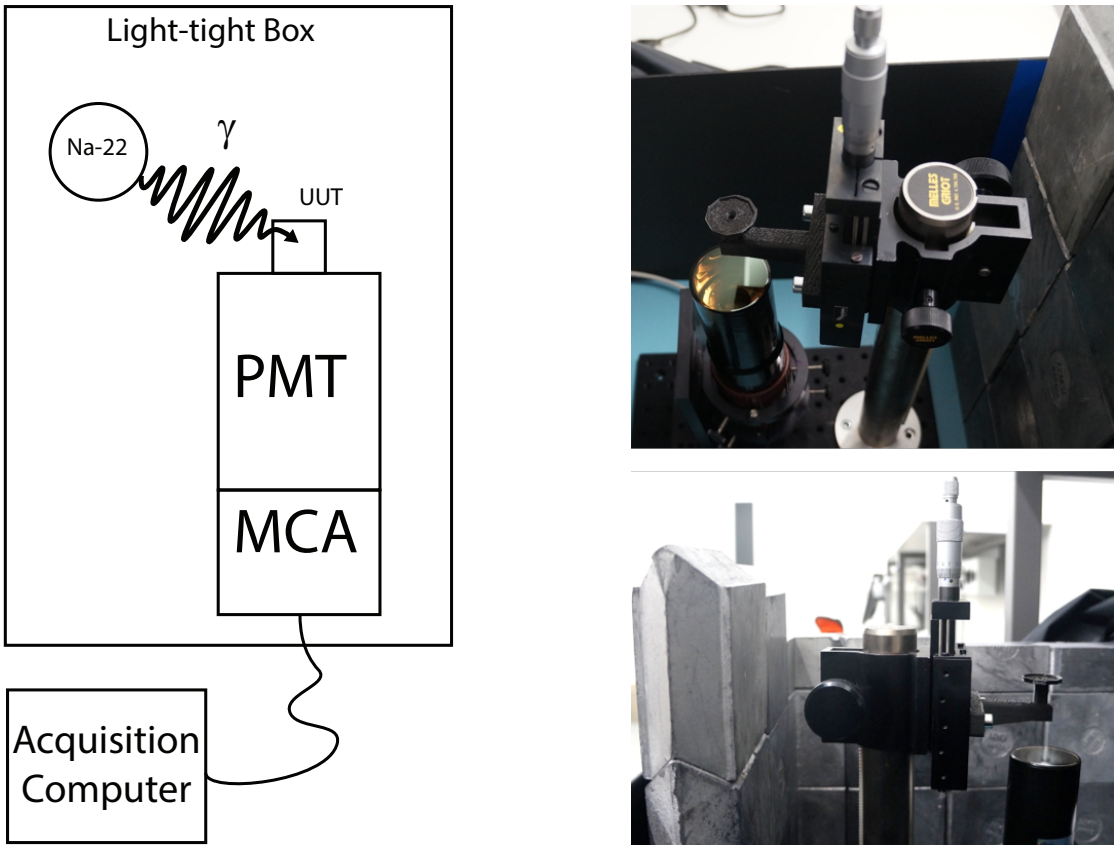


Figure 5.10: Reflectivity spectra of the DBRs implemented on glass substrates. *Left*: a conventional DBR designed with a center wavelength of 480 nm, to match the emission spectrum of BGO. *Right*: a combination of DBR (centered at 360 nm) and silver coating to extend the spectrum toward the red. *Top row, left*: simulation results vs. measurements of different samples, *Top row, right*: measurement results of the complete structure as well as its components. *Bottom row*: schematic representation of the respective DBR architectures.

layer pairs of SiO_2 and SiN_x (81.2 nm and 60.9 nm in thickness respectively, for a central wavelength of 475 nm). This mirror was designed to align its reflectivity spectrum with the emission spectrum of the BGO crystal [343], [353]. This coating was fabricated by means of a Plasma-enhanced Chemical Vapor Deposition (PECVD) process on a glass substrate. The reflectivity of this mirror was measured for several samples to validate the uniformity and reproducibility. All samples showed very similar behavior, which is also in agreement with the simulation results. The simulations were performed using a custom Python routing based on the CAMFr (Cavity Modelling Framework) [354] libraries. The second mirror, shown on the *right* of Fig. 5.10, was designed using Ta_2O_5 and SiO_2 layers (61 nm and 39 nm in thickness respectively, for a central wavelength of 360 nm). The design of this mirror aimed at extending the reflectivity also in the region where conventional reflectors (e.g., ESR) start to be lossy [355]. Another motivation for this choice is that several inorganic scintillators, such as LYSO:Ce, show an emission spectrum shifted toward the blue [353]. The design center wavelength was thus selected to be 360 nm and the reflectivity spectrum of such a DBR is shown in Fig. 5.10, *right* (blue line). This NUV-DBR, when coupled to a silver mirror, creates an enhanced reflective surface that ranges from 330 nm to the infrared with a reflectivity above 90% (Fig. 5.10, *right*, green line). In this case, the mirror has been fabricated on a glass substrate but using RF sputtering (Alliance-Concept DP650) and characterized using a spectroscopic reflectometer.



(a) (b)

Figure 5.11: (*a*): Experimental setup scheme used to measure the energy spectrum of the scintillator unit under test (UUT). The sample is coupled to a PMT inside a light-tight box to avoid any light background. Then, the PMT is read out with a multichannel analyzer (MCA), and the data are used to built the energy spectrum histogram in an acquisition computer. (*b*): pictures of the actual setup.

5.5 Experimental Setup

A set of ten $10\times10\times10$ mm^3 crystal samples (Epic-Crystal) with polished surfaces was used. The scintillator was coupled in all cases to the photodetector using optical grease. A bare crystal without any coating represented the most straightforward configuration (ID = 1). Two additional configurations relied on the addition of ESR or Teflon applied to the side walls (ID = 3 and ID = 4), whereas in one other case, a DBR was applied to the top surface only (ID = 2). Finally, three configurations employed a PhC on a bare crystal (ID = 5), as well as together with Teflon wrapping on all sides or in combination with a DBR (ID = 6 and ID = 7), again only on the top surface. The DBRs used for the crystals are those shown in Fig. 5.10, the left one for BGO, and the right one for LYSO. The DBRs, deposited on glass substrates, were air-coupled in both cases. For each experiment, a crystal was placed with the exit surface coupled to a PMT (ET Enterprises Electron Tubes 9266KB) with optical grease (n = 1.4) to measure the

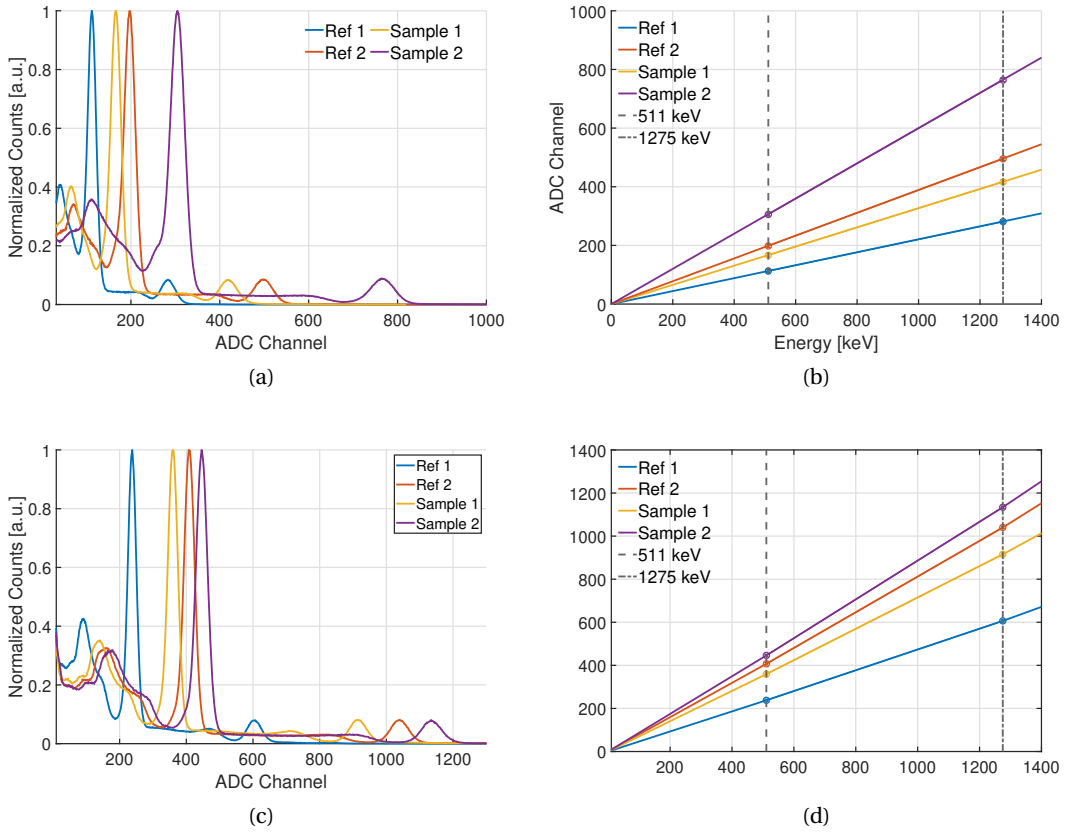


Figure 5.12: (*a*): Examples of energy spectra obtained with BGO samples in different configurations. *Ref1*: reference crystal with Teflon wrapping, air-coupled; *Sample 1*: crystal wrapped with Teflon with PhC on the output surface, air-coupled; *Ref 2*: reference crystal with Teflon wrapping and optical grease; *Sample 2*: crystal with Teflon wrapping, PhC on the output surface and optical grease. (*b*): corresponding calibration curves derived from the position of the two photopeaks (511 keV and 1275 keV) to eliminate any offset. The slope of the curves provides a visualization of the light gain improvement. (*c*) and (*d*): same results for LYSO.

output light yield. The choice of a PMT was motivated by its high linearity with respect to the detected light, ease of use, and measurement reproducibility; the latter was also enhanced by the use of a micropositioner to control the optical glue layer down to a thickness of around 100 μm. In addition, each sample was irradiated with a gamma source (^{22}Na) of 4 MBq to induce scintillation inside the material. Finally, the output signal from the PMT was injected in an Ortec *Digibase* multichannel analyzer (MCA), which provided the (uncalibrated) energy spectrum. Fig. 5.11 shows the setup scheme and the picture of the actual implementation.

Each sample was tested for all the configurations before and after implementing the PhC on the output surface. These two measurements allow verifying the relative performance improvement. In addition, all the crystals were taken from the same batch and measured in the standard configuration. This procedure was needed to avoid evaluation errors due to

possible light yield variations between crystals (especially from different batches) and ensure that the PhC fabrication process neither damaged the crystals nor affected their performance.

The measurement was performed by coupling the crystal sample to the PMT inside a light-tight box to avoid background noise given by environment light. The radioactive source was then placed on a custom-designed stand to keep a constant distance from the sample. The MCA was used to read the PMT photocurrent pulses, digitize them and create a histogram. The measurement was performed for enough time, typically a few minutes, to allow the accumulation of a sufficiently high count number in the energy spectrum histogram and reduce the errors in the statistical analysis to a negligible level.

The data acquired with the MCA were analyzed with custom software to extract the energy resolution value and the number of ADC channels corresponding to the 511 keV and 1275 keV photopeaks. The spectrum was calibrated by utilizing the two ^{22}Na photopeaks. The final energy resolution was evaluated on the calibrated spectrum as the FWHM at the first photopeak, calculated using a Gaussian fit divided by 511 keV. This value is independent of different systematic errors and variations between different samples of the same material. The comparison of the (calibrated) energy spectra resulting from different configurations allowed us to measure each configuration's performance in terms of variability, light gain, and energy resolution.

5.6 Experimental Results

The results obtained with the setup described in the previous section are summarized in Table 5.1. The table shows the measured light gain and energy resolution gain for seven crystal configurations, for both BGO and LYSO, taking as reference a standard configuration featuring a bare crystal wrapped in Teflon coupled with optical grease (ID = 4).
The use of a PhC pattern (Fig. 5.9) showed a light extraction improvement of ~41% and ~10% for BGO and LYSO respectively, measured with respect to the best configuration without PhC (i.e. Teflon wrapping and optical grease). Concerning the corresponding energy resolution, the improvement was of ~21% for BGO and ~4% for LYSO. Figure 5.12 shows the energy spectra obtained using BGO (*Top row*) and LYSO (*Bottom row*) crystals; the use of a PhC did clearly improve in a significant way the overall amount of light in output of the scintillator. Moreover, a closer look at the BGO *Sample 2* (Fig. 5.12, *Top left*) curve illustrates how the spectrum becomes more detailed (the K-shell peak starts to be visible), indicating as well an improvement in energy resolution.
The use of the previously described DBRs on the top surface of a bare crystal showed a modest improvement compatible with what was previously reported in [343]. This is mostly due to the light that escapes from the lateral surfaces and will be discussed in more detail in the next section. Conventional solutions based on ESR or Teflon (3rd and 4th configurations in Table 5.1) featured quite similar performance and in line with expectations.
The silica sol-gel structure was extensively tested: it featured a resistance to cracking on

Comparison of Experimental Results					
ID	Crystal	Configuration	Light gain	Energy Resolution (%)	Energy Resolution improvement
1	BGO	Bare crystal & Opt. Grease	0.55	20.8 ± 0.48	0.74
2	BGO	Bare crystal & Opt. Grease & DBR (top)	0.64	19.3 ± 0.26	0.80
3	BGO	ESR & Opt. Grease	0.98	15.6 ± 0.29	0.99
4	BGO	Teflon & Opt. Grease	1.00	15.4 ± 0.19	1.00
5	BGO	PhC Pattern & Opt. Grease	0.80	17.2 ± 0.58	0.90
6	BGO	PhC Pattern, Teflon & Opt. Grease	1.41	12.7 ± 0.36	1.21
7	BGO	PhC Pattern & Opt. Grease & DBR (top)	0.88	16.4 ± 0.57	0.94
1	LYSO	Bare crystal & Opt. Grease	0.74	12.2 ± 0.32	0.85
2	LYSO	Bare crystal & Opt. Grease & DBR (top)	0.79	11.8 ± 0.36	0.88
3	LYSO	ESR & Opt. Grease	1.00	10.4 ± 0.12	1.00
4	LYSO	Teflon & Opt. Grease	1.00	10.4 ± 0.15	1.00
5	LYSO	PhC Pattern & Opt. Grease	0.85	11.4 ± 0.33	0.91
6	LYSO	PhC Pattern, Teflon & Opt. Grease	1.10	10.0 ± 0.24	1.04
7	LYSO	PhC Pattern & Opt. Grease & DBR (top)	0.86	11.3 ± 0.34	0.92

Table 5.1: Summary of all experimental results on $10\times10\times10$ mm^3 BGO and LYSO crystals in seven configurations. Reference: a standard configuration featuring a bare crystal wrapped in Teflon (ID = 4). The scintillator was coupled in all cases to the photodetector using optical grease. The light gain and the energy resolution have been calculated on the first peak of the spectrum (511 keV).

bending higher than 1 T; an impact resistance of 18 J; a surface pencil hardness larger than 2 H, and an abrasion resistance (ISO3160) with no modification after 6 hours (EN 13523).

5.7 Discussion

The most important result achieved by this study was the significant light extraction enhancement when a modified nanoimprinted photonic crystal, enhanced with a sol-gel protection (hybrid organic/inorganic silica sol-gel buffer layer) to increase its robustness, was applied on the output surface of an inorganic scintillator. The best result was obtained by encapsulating the crystal with reflectors (e.g., Teflon) on the side surfaces to redirect the light toward the output. The PhC efficiency clearly improved with the growing scintillator refractive index. Thus, a substantial improvement in the BGO performance has indeed been obtained, whereas its use on LYSO crystals showed a milder gain. This is not surprising per se, given the large refractive index difference between the two scintillators (2.15 for BGO vs. 1.8 for LYSO), and that the simple use of a refractive index matching material is already providing good optical coupling for LYSO. This explanation is also accredited by similar performance trends for both LYSO and BGO in experiments featuring PhCs deposited on crystals that are air-coupled to photodetectors. Indeed, from measurements performed on our air-coupled samples with Teflon wrapping, we observed an improvement of ~25% and ~15% in energy resolution and ~50% and ~30% in light gain for LYSO and BGO, respectively. In this case, the gain provided by the PhC coating, without employing optical coupling compounds, was in line with what was reported in the literature [330]–[335], [339].

Concerning the use of DBR mirrors, only a modest result (in particular considering the energy

resolution) was obtained when using a single mirror applied on the top surface of the two kinds of crystals. This result is mainly because the crystal lateral surfaces are uncovered in this configuration (configuration 2 in Table 5.1), and a significant part of the scintillation light gets lost. On the other hand, the light gain does slightly increase when air-coupling the mirrors on five of the six surfaces, although still below our expectations. The main reasons lie in the technological hurdles that need to be overcome to deposit the reflective coating on the scintillators' surface directly. The corresponding process is indeed quite complex, suffering from edge effects or generating a sufficient amount of heat to compromise the crystals' performance. As a result, the performance of the coating (especially DBRs) changes close to the edges of the sample. Moreover, we observed a performance degradation when a plasma-based process was used; the underlying reasons are currently being investigated.

To conclude, we have added a silica buffer layer which encapsulates the nanostructures, thereby ensuring structural robustness and optimal coupling to the photodetector optical window, contrary to most previous air-coupled implementations, while ensuring unvaried PhC performance. The use of the PhC on the output surface showed a significant improvement in light extraction and energy resolution (see Table 5.1), in particular for BGO. This could possibly lead to reconsider the use of this scintillator in the development of future TOF-PET modules.

In addition to the core work on improved photonic crystals, we did also investigate the use of thin film coatings. Such reflective very thin structures (order of 1 μm), once applied to the sides of the scintillators, can allow a significant reduction of the dead space between crystals, and a corresponding improvement of the packing fraction (which in turn translates into an improvement of the overall detection efficiency of the system). This is of particular relevance when implementing tiles of miniaturized (sub-millimiter) crystal needles where the thickness of the crystal coating has a high impact, and can also contribute to suppression of optical crosstalk between neighboring needles. Several DBRs were simulated prior to fabrication and two different structures compared, both deposited on a glass substrate. One structure was optimised for BGO and another, whose reflectivity extended to the NUV, was coupled to a silver mirror, resulting in a combined very broad reflectivity range. Other wavelength ranges can be implemented if needed, thereby exploiting the full tunability potential of DBRs. The measured reflectivity spectra of the aforementioned DBRs agreed well with the simulated data, but the experimental results obtained when using them air-coupled to the crystals were somewhat below expectations. This was mostly due to several technological hurdles which still need to be overcome for the direct deposition of the reflective coating on the surface of the scintillators, and which we are currently exploring, together with the direct deposition of metal films.

We did indeed investigate the possibility of directly depositing metals on the crystal surface, either by metal evaporation, or by Physical Vapor Deposition (PVD). The former works at room temperature, i.e. without involving a plasma entering in contact with the crystal surface, and was implemented with success for the deposition of silver, although we encountered adhesion

issues. The latter was tested with aluminum, resulting in no apparent damage to the surface of the crystal, with an adhesion better than the one of silver, also without resorting to adhesion layers. The study of these additional treatments is still under development.

The next steps concerning the PhC will consist in extending the experimental characterization to crystal needles as well as measuring the possible impact of the PhC on the spatial reconstruction on one side, and the temporal on the other in terms of timing performance. We expect an improvement of CRT as also suggested in [331], [334], [356], [357]. Another interesting, though challenging perspective is represented by the simulation of the light interaction with the entire scintillator-photodetector interface and the scintillating crystal itself, to optimize the output gain. This will require various simulation tools and more specifically multiscale modeling. Such tools are currently available and will be tested in the near future [358].

6 Case Study: Direct Minimum Ionizing Particle Detection with SPADs

Considering the extraordinary timing jitter performance obtained with the SPAD-based sensor presented in Section 3.1, we decided to explore another TOF application for this kind of detector, such as direct particle detection. For this application, the system relies on the measurement of the TOF information collected by two sensors in coincidence. Although this case study is strictly linked to a TOF application suitable for high energy physics, there are some similarities with more exotic PET systems that are not using scintillator-based modules but rather lead layers for gamma conversion and direct detection of high energy electrons using solid state detectors. Examples of such systems were shown in [359]–[362].
In this work we focus the attention on minimum ionizing particles (MIP), i.e., charged particles with a momentum such that the ionizing energy released per unit length passing through matter is minimum. In other words, a MIP is a particle with an energy that falls into a small interval around the minimum of the Bethe-Bloch curve (see Fig. 6.1) [363].
After an introduction on the typical way to perform particle detection using solid-state devices, we go through the description of the system implemented for this application, and the experimental setup used on the test beam. The measurement results are discussed toward the end of the the chapter.

This case study is adapted from the work reported in [278]

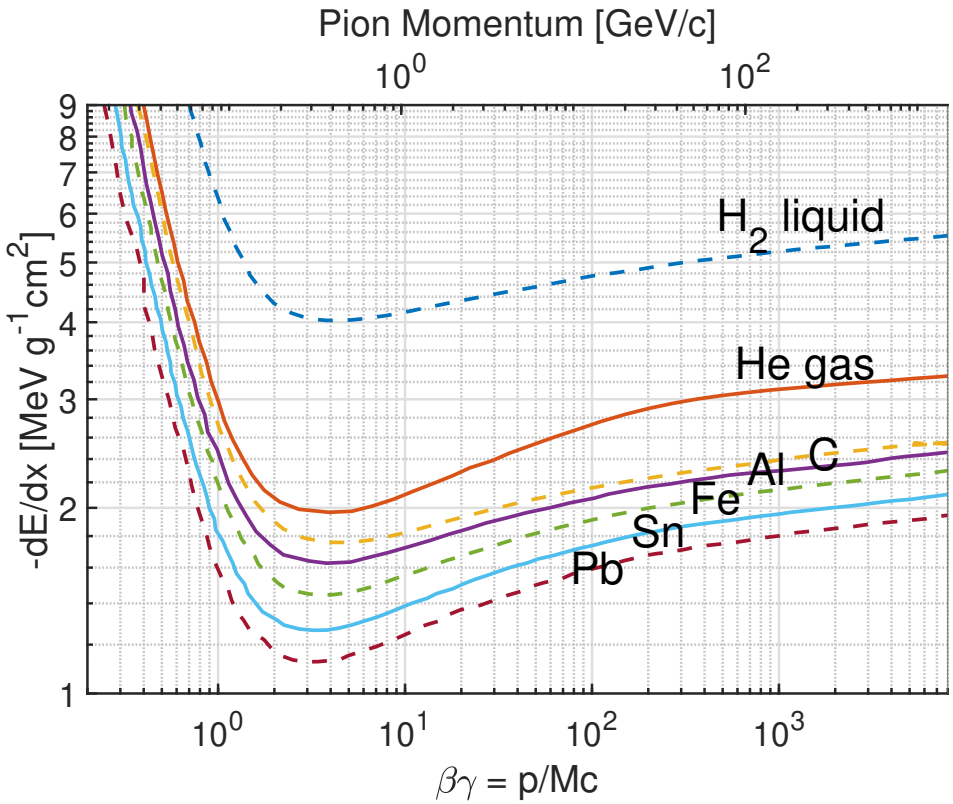


Figure 6.1: Bethe-Bloch curve for several materials. The graph shows the energy loss by relativistic charged particles crossing matter as a function of their momentum [363].

6.1 High timing resolution with silicon pixel detectors

Silicon pixel detectors have been developed in high-energy physics applications to provide precise position measurements thanks to their compactness and high spatial granularity. Recent developments have been focused on sub-100 ps timing measurements of optical photons and direct detection of charged particles.

When a particle passes through the detector, electron-hole pairs are generated. When these charges move in the depletion region, an induced current pulse is registered on one electrode. According to the Schockley-Ramo theorem [364], [365], this current is proportional to the free charge Q, to the speed of the charge carriers v, and to the weighting field, which can be expressed, to a first approximation, as $\frac{1}{d}$, where d is the thickness of the depletion region. Hence, we can calculate the induced current as:

$$i = kQv\frac{1}{d}, \tag{6.1}$$

where k is a proportionality factor. The signal ends when all charges have been collected. Moreover, in case of a MIP crossing a thin device, the charge Q is proportional to d. We thus have:

$$i = kNdv\frac{1}{d} = kNv, \tag{6.2}$$

where N is the number of electron-hole pairs generated per unit length. This result shows that the initial value of the induced current is constant and does not depend on the thickness of the depletion region. When reading out this current signal on a load, behaving like an ideal transimpedance amplifier, we observe a sharp voltage pulse. The time-of-arrival (ToA) of the charge is usually determined by comparing the voltage pulse with a threshold. The uncertainty of the voltage pulse σ_V is expressed as:

$$\sigma_V = \sigma_t \frac{dV}{dt}, \tag{6.3}$$

where σ_t is the jitter of the voltage pulse. By inverting Eq. 6.3, we find that:

$$\sigma_t = \frac{\sigma_V}{\frac{dV}{dt}}. \tag{6.4}$$

Eq. 6.4 shows that the signal fluctuation (σ_V) should be reduced to achieve a better timing jitter σ_t and the slew rate $\frac{dV}{dt}$ should be maximized. In case of a sensor with an internal finite gain G [287], [366], [367], the slew rate is proportional to $\frac{G}{d}$. This analysis suggests that thin sensors with large internal gain will in principle result in a better timing jitter. There is, however, another effect that can significantly affect the pulse shape while detecting MIPs: the charge collection noise. This phenomenon is caused by the variability of the profile of the deposited charge. As shown in [368], this effect introduces a timing jitter that is non-negligible at the 10 ps scale, and which increases with the detector thickness [369]. Various solutions have been proposed to reduce the contribution of this additional source of timing jitter, such

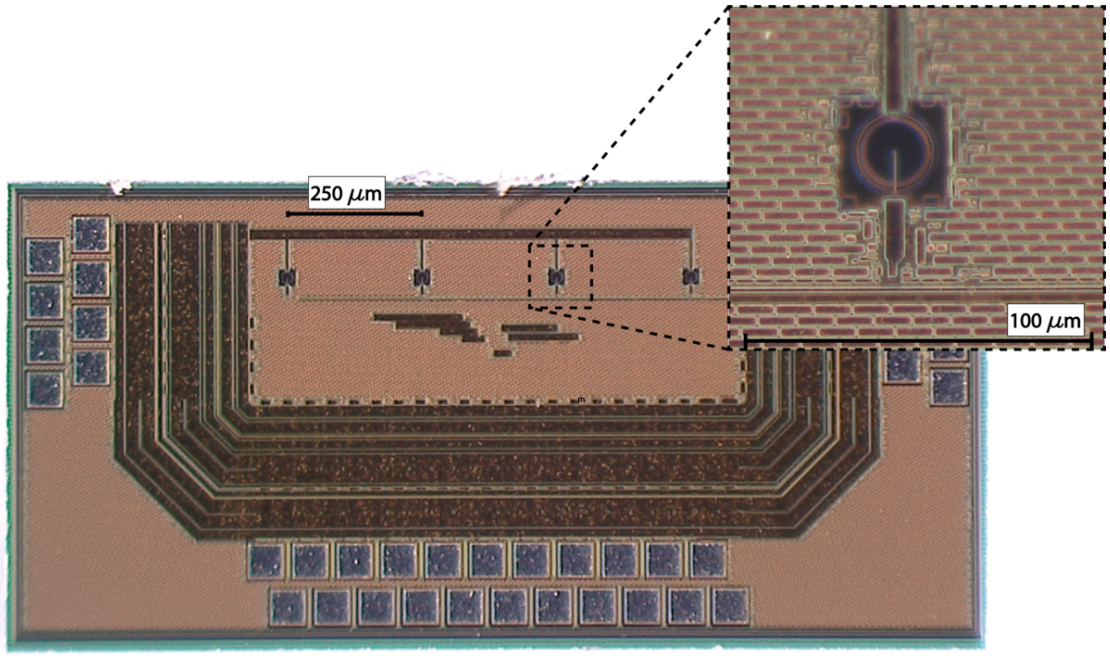


Figure 6.2: Micrograph of the implemented chip embedding 25 μm diameter SPADs with integrated pixel circuit [232].

as the detectors reported in [367], [370], [371].

All the effects mentioned above call for extremely high intrinsic gain and slew rate together with thin structures. Thus, SPADs could represent promising candidates for substantial timing jitter reduction [26]. In SPADs, unlike APDs, the avalanche is a self-sustaining process, and the timing jitter contributions are more related to the avalanche growth dynamics. In particular, the timing jitter improves, decreasing the avalanche current value needed by the front-end electronics to detect a pulse [288]. A comprehensive theoretical study of timing performance in SPADs when used in MIP detection is presented in [199]. First examples of such systems, detailed in [372]–[374], were following the concept proposed by [375].
Although these potential advantages are promising, some problems usually affect Geiger-mode devices in the framework of MIP detection. Indeed, typical SPADs have a long dead time if no properly designed active quenching and recharge techniques are used [203]. Another issue is the presence of noise in the form of spurious pulses even in the dark. This noise, known as dark count rate (DCR), could limit the suitability of SPADs for the target application because of significant degradation of the measurement signal-to-noise ratio (SNR). However, as shown by [376], [377], DCR can be drastically reduced by detecting MIPs with two SPADs operated in coincidence. Building on these elements, we present MIP time-of-flight (ToF) measurements resulting in unprecedented timing precision.

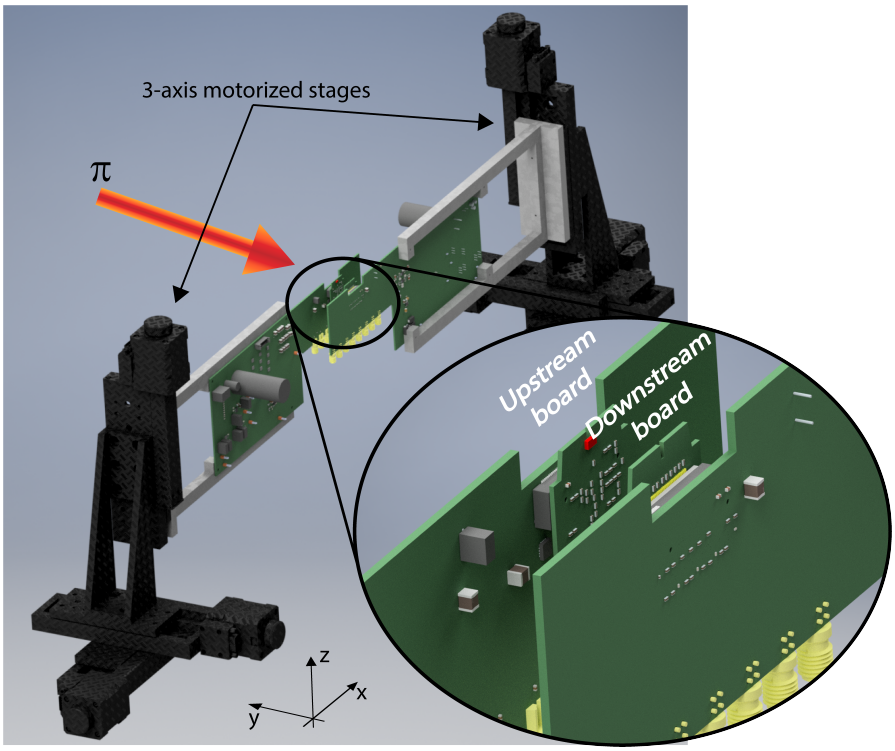


Figure 6.3: Experimental MIP coincidence measurement setup.

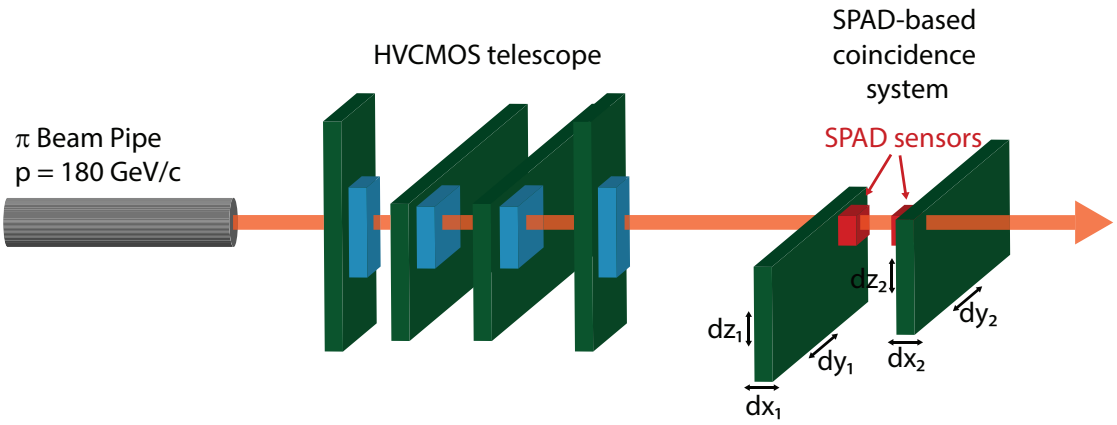
6.2 SPAD Detector and System-on-Board

The detector system used in this work relies on the SPAD-based sensor presented in Section 3.1. Fig. 6.2 shows a micrograph of the sensor used in this study. The sensor, as described, integrates four independent SPAD pixels with a diameter of 25 μm. A dedicated on-chip front-end circuitry, shown in Section 3.1.5, is implemented in close proximity to each SPAD. The circuit is designed to enable a tunable dead time, as short as 3 ns, supporting very high count rates while still maintaining very low afterpulsing [232].

In this work, we used the complete and optimized system-on-board described in Section 3.1.5, to take advantage of its improved perfomance and compactness. Moreover, the fast SiGe comparators (Analog Device ADCMP572) that drive 50 Ω lines (Fig. 3.16) allow to reduce the capacitive load at the chip's output (high impedance node) and propagate the signal through a fairly long high-frequency cable to the timestamping electronics without degrading the signal.

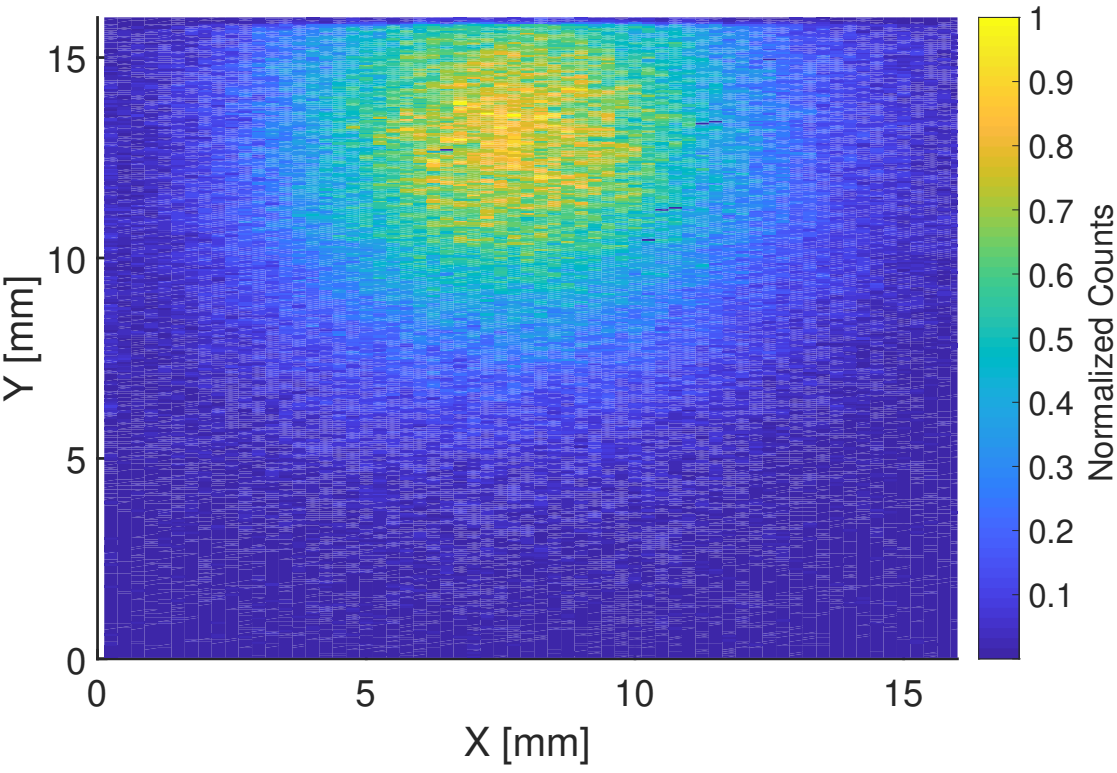
6.3 Minimum Ionizing Particles Detection

The setup used for the ToF measurement of MIPs is shown in Fig. 6.3. It consists of two systems-on-board (see Section 3.1.5), both mounted on motorized linear stages with sub-micrometer positioning resolution to allow a proper detector alignment and to guarantee



(a)

Figure 6.4: Simplified scheme of the testbeam line setup.



(a)

Figure 6.5: MIP beam profile acquired with the HVCMOS telescope described in [378].

the acquisition of coincidence measurements. We installed the setup on the H8 beamline in the CERN North Area. This beamline delivers 180 GeV/c momentum pions produced on a graphite target by the interaction of protons accelerated by the Super Proton Synchrotron (SPS). Fig. 6.4 show a simplified scheme of the full test bench installed on the beam line.

Bias (V)	FWHM (ps)	FWTM (ps)	σ (ps)	σ_{single} (ps)
24	27 ± 1	104 ± 4	11.5 ± 0.4	8.1 ± 0.3
27	22 ± 2	62 ± 3	9.4 ± 0.7	6.6 ± 0.5

Table 6.1: Summary of the MIP detection measurement results. The Gaussian sigma has been obtained by dividing the FWHM by $2\sqrt{2ln(2)}$. Assuming that the response is the same for both SPADs, the σ_{single} values have been obtained by dividing the σ values by $\sqrt{2}$. The errors have been evaluated using statistical error propagation.

The beam profile measurement is reported in Fig. 6.5. A good alignment between the two detectors has been achieved thanks to an HVCMOS telescope [378]. The two detectors were positioned at the center of the beam, where the intensity is the highest. Coincidence events were acquired for two bias voltages, 24 V and 27 V, corresponding to approximately 2.5 V and 5.5 V V_{ex}, respectively. Fig. 6.6 shows the ToF distributions for both V_{ex}. The MIP measurement results are summarized in Table 6.1.

6.4 Radiation Hardness

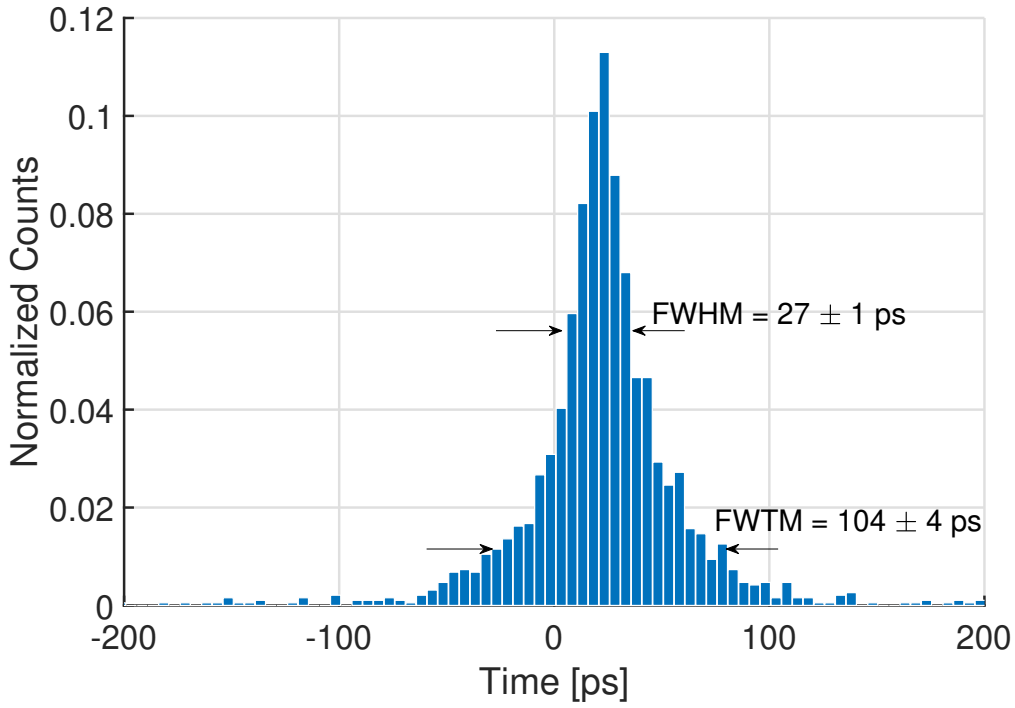
The radiation hardness of the SPAD detectors was characterized by using protons to induce ionizing damage and, more importantly, displacement damage, which causes structural permanent defects [379], [380]. The detectors were irradiated using the Proton Irradiation Facility (PIF) at the Paul Scherrer Institute (PSI, Villigen, Switzerland). We used a 100 MeV monoenergetic beam with a fluence of 1×10^8 protons per second to reach a 300 TeV/g displacement damage dose (DDD) and 9.4 krad total ionizing dose (TID). The DCR difference was measured two weeks after the exposure with the aforementioned setup. The DCR comparison before and after the exposure is reported in Fig.6.7. The characterization of the radiation hardness shows that the SPAD detector can maintain its functionality under the given radiation dose. The SPADs are not saturated by the DCR induced by the radiation damage thanks to their short dead time and high count rate. Moreover, as the ToF measurement is based on coincidence, even detectors reaching a DCR value of 1×10^5 counts per second will not affect their particle detection performance, as also shown in [376], [377]. No other degradation of the device performance was observed.

6.5 Discussion

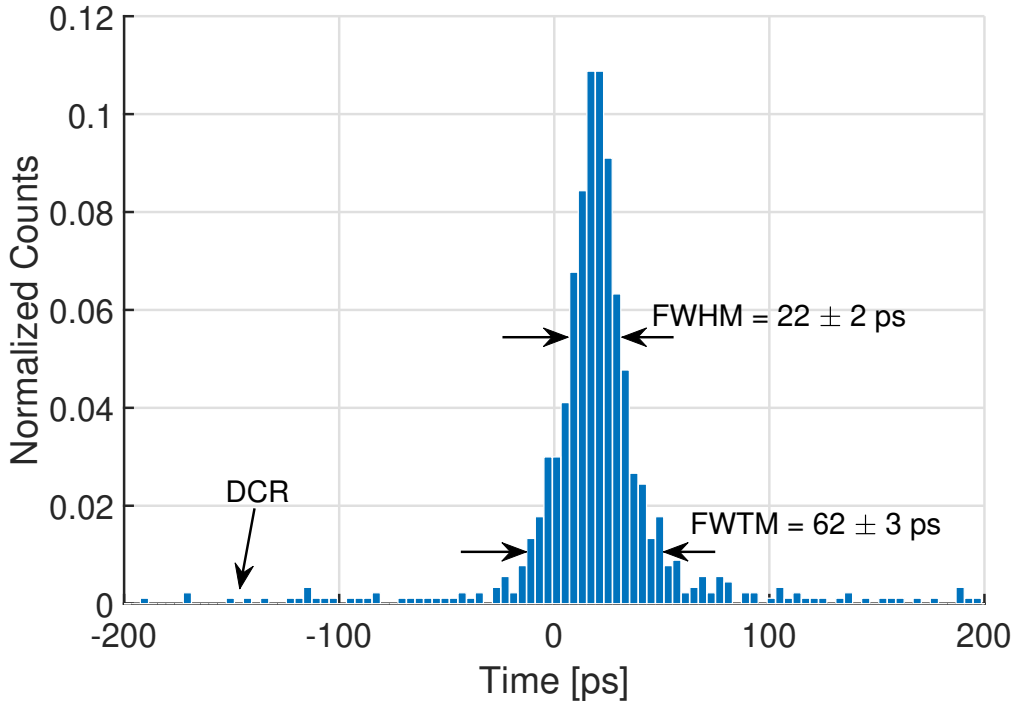
In this work we showed how Geiger-mode devices (i.e., SPADs) can detect MIPs with a sub-10 ps timing precision. This result paves the way to the implementation of future high timing resolution particle trackers based on this kind of detector. Moreover, the radiation hardness of the device was explored up to a DDD of 300 TeV/g. After exposure, DCR increased by about 3 orders of magnitude, which is comparable with [381]. Nevertheless, this increment does not affect the timing performance of the device. In fact, the number of accidental coincidences due to DCR is strongly suppressed by the logic AND between the two SPADs used for the ToF

measurement [376].

When analyzing Fig. 6.6 and the results in Table 6.1, we notice a dependency of the performance on the applied bias voltage. In particular, we can see a lower FWTM and an improvement in the FWHM for the ToF distribution when increasing the bias point from 24 V to 27 V. The higher field improves the avalanche buildup and lateral spread time dispersion [193]–[195]. Moreover, a higher bias voltage enlarges the drift region and increases the electric field inside it. This effect reduces the size of the neutral region and helps minimize the statistical spread of the diffusion and transit time needed by the primary charge carriers to reach the multiplication region [26].
Both distributions show a negligible flat background coming from random dark count coincidences. This is indicative of the efficacy of the noise filtering provided by measurements in coincidence.
Moreover, we can compare these results also with what obtained using optical photons. Assuming that the two detectors used for the coincidence measurement have the same performance, we can easily derive the contribution of one of the two by dividing the result by $\sqrt{2}$. With this assumption we can note how the measurement performed at 24 V reverse bias for both charged particles and optical photons produce the same result of about 20 ps FWHM. However, we can see a difference of about 5 ps in the results obtained at 27 V. Although this error is high in relative terms, in absolute terms it is quite low and it could attributed to some additional interference or noise on the signal. To get more understanding and verify the source of this error, further experimentation is planned as a future step.



(a)



(b)

Figure 6.6: Normalized distribution of the MIP time-of-flight between two SPADs in coincidence. (*a*) at 24 V and (*b*) at 27 V bias voltage.

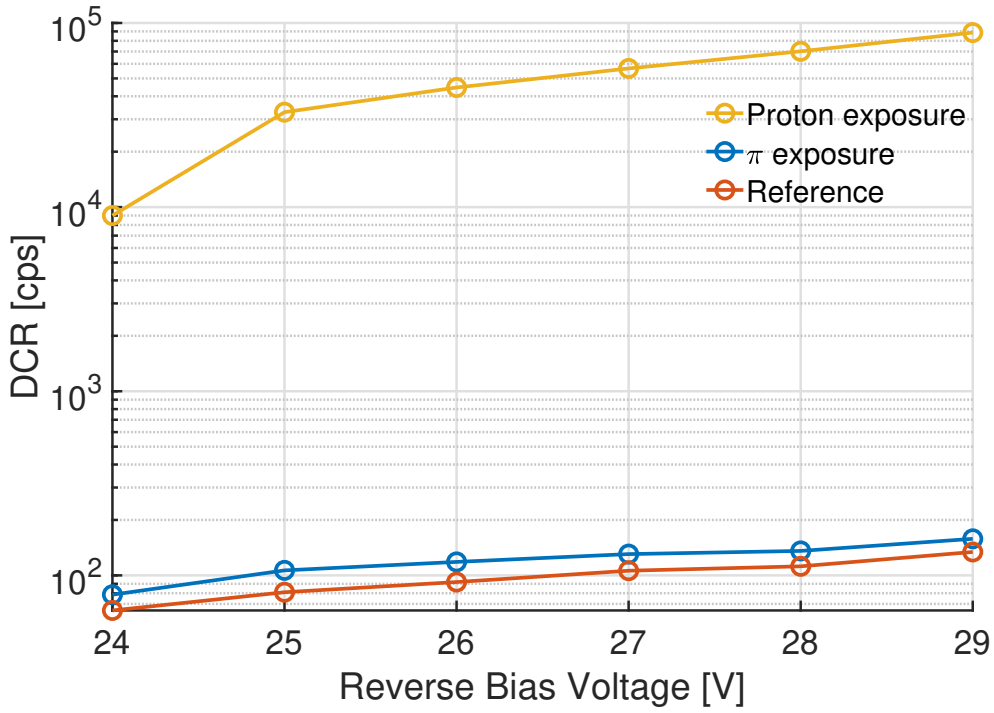


Figure 6.7: Radiation hardness experimental results. DCR trends over various excess bias voltages are reported for the same device. The reference with no irradiation (*orange*) is compared with the noise level after 100 MeV proton irradiation (*yellow*), and with the samples after 1 week of 180 GeV/c pions irradiation.

7 Conclusion

7.1 Conclusions

The primary aim of this thesis was the study, design, and characterization of SPAD-based sensors with improved performance for time-of-flight applications, with a particular focus on TOF-PET. The SPAD devices demonstrated in this thesis exhibit excellent timing performance and are compatible with large-size detectors. Various SPAD structures and front-end circuits were presented, together with a large-scale sensor relying on the 3D-stacking FSI approach and novel solutions on the application side.

The initial phase of this thesis involved SPAD design in two standard deep sub-micrometer technology nodes: 180 nm CMOS and 55 nm BCD.
As a first step, simple SPAD-based sensors were implemented in 180 nm CMOS technology, featuring an optimized pixel front-end integrated on-chip. The behavior of these devices behavior was studied as a function of device size, and the experimental results show how a proper design could improve performance to unprecedented levels. In particular, this study showed the direct dependency of timing jitter, afterpulsing, and DCR on the device size. As a result of this study, we could design a SPAD that can achieve an SPTR of 7.5ps FWHM at room temperature. Moreover, the use of a PQAR circuit directly integrated on a chip allowed the reduction of the dead time to the of 3 ns allowing about 300 MHz count rate and limiting at the same time the afterpulsing probability to well below 1%.
Next, the focus moved on a much more advanced technology node (55 nm BCD). Several SPAD structures were implemented, and four of them are thoroughly described in this thesis. We showed a sensitivity performance improvement of about a factor two on our devices, achieving a peak PDP of about 62% at 530 nm. Moreover, using an integrated pixel front-end based on a PQAR architecture, we demonstrated how to achieve an afterpulsing probability of ~0.13% at 1.5 ns dead time and an SPTR of about 30 ps. In addition, we also explored the use of shallower junctions to increase the sensitivity in the blu-NUV portion of the spectrum. This goal was achieved with two devices that show a sensitivity spectrum centered at shorter wavelengths (about 60% at 440 nm), together with low noise and good timing jitter.

The aforementioned large-scale sensor implemented in this thesis uses 3D integration to overcome some problems intrinsically linked to the 2D approach, such as a tradeoff between the sensor FF and the amount of electronic circuits that could be integrated on-chip. In contrast to conventional 3D-BSI approaches, this sensor exploits the FSI-approach. During this thesis, the 3D process was custom-designed concurrently to the sensor. The choice of using FSI is driven by the need to have a good sensitivity at short wavelengths (below 500 nm) that can meet the specifications of applications, such as TOF-PET.
The sensor is a MD-SiPM and features 128 channels (clusters) split in two independent cores of 8×8 clusters. Each cluster includes 64 SPADs-based pixels, a photon counting system used to track the number of photons detected, a SPAD address system that provides the information on what SPAD fired at each frame, and a TDC. The sensor structure was carefully analyzed, and a preliminary characterization of its functionality was performed. The results show the correct functionality of the chip and provide a first estimation of the 3D process yield (about 90%). The DCR was measured as a function of temperature, showing how cooling is very effective for noise reduction.

Next, we focus on the optical optimization of an inorganic scintillator-based module for gamma-ray detection. To this end, several configurations of wrapping and coating were employed on two commonly used inorganic scintillators, such as LYSO and BGO. We also PhC directly implemented on the scintillator surface to improve light extraction. The PhC structure was improved by adding a silica buffer layer that guarantees stability and robustness, also allowing the use of common optical coupling compounds, such as greases and waxes. The experimentation of this solution showed a significant improvement in light extraction and energy resolution. With BGO, we achieved a 40% improvement in light and about 25% in ER, thanks to the higher refractive index of this material (2.15) with respect to LYSO (1.8).
BGO is also interesting due to the lower cost and the possibility of detecting prompt Cherenkov photons.

Following the design and characterization of the device mentioned above in 180 nm CMOS technology, we demonstrated the unique time-of-flight coincidence performance of SPAD employed for direct particle detection on a MIP beamline. In this context, a coincidence system was implemented to cope with the issue of dark counts. The results of this experiment showed an outstanding coincidence time resolution of 9.4 ps sigma, which corresponds to 6.6 ps sigma on the single device. Moreover, using a coincidence system has proved to be highly effective in background reduction, improving the SNR of the measurements significantly. Moreover, we highlighted a similar behavior of the device compared to light detection. Indeed, the performance results are relatively close in terms of timing jitter, and it depends on the applied bias. Increasing the excess bias voltage of the device, we observe an improvement of the sigma and a reduction of the diffusion tails.
This notable result opens the door to SPAD for low-cost, high timing performance particle detectors meant for beamline facilities.

7.2 Future Work

As a first step, further analysis on the fundamental limit of SPAD timing performance could be carried out redesigning the system based on the SPAD in 180 nm, optimizing all those parameters that contribute to the timing jitter of the device. Further, the possibility of scaling up the device to a large array keeping the same timing performance, should be explored. For this step, the design of a high-performance timing circuit, such as a TDC, should be conducted to preserve the timing precision achieved in the thesis.

Another interesting open question is the real potential of the 55 nm BCD technology and whether other novel and high-performance devices could be implemented. Moreover, a close collaboration with the foundry can allow the introduction of some custom layers in the process that could be used to better shape the electric field profile inside the SPAD cross-section and guarantee further performance improvement.
Many other devices were implemented in this technology and are currently object of study and detailed characterization.

The characterization of our 3D-stacked FSI MD-SiPM is still ongoing, and more effort is needed before it can be completed. In this sense, further experiments aiming to exploit the chip functionality need to be performed.
Several critical points need to be addressed during a second implementation. A first point could be the redesign of the SPADs in order to obtain a device with a lower noise level. Then, the mechanical stabilization through filling with organic of the gap between the two tiers can help avoid the risk of cracks in the top tier and allow an easier and more reliable coupling of the sensor with inorganic scintillators as provided for PET application.

Although the light extraction enhancement and energy resolution improvement results are encouraging, further measurements need to be performed. Indeed, the study performed limited results to characterize the amount of light extracted with sets of crystals of a specific size. Therefore, the following natural steps will be to characterize this solution with crystals of different geometry and the expected improvement in timing performance (CTR).

Eventually, excellent performance shown by SPAD devices in direct MIPs detection should be further explored for very high-performance particle trackers. This point should be addressed in several ways, starting from the implementation of a larger SPAD array to the design of a dedicated device optimized for this kind of application.

Chip Gallery

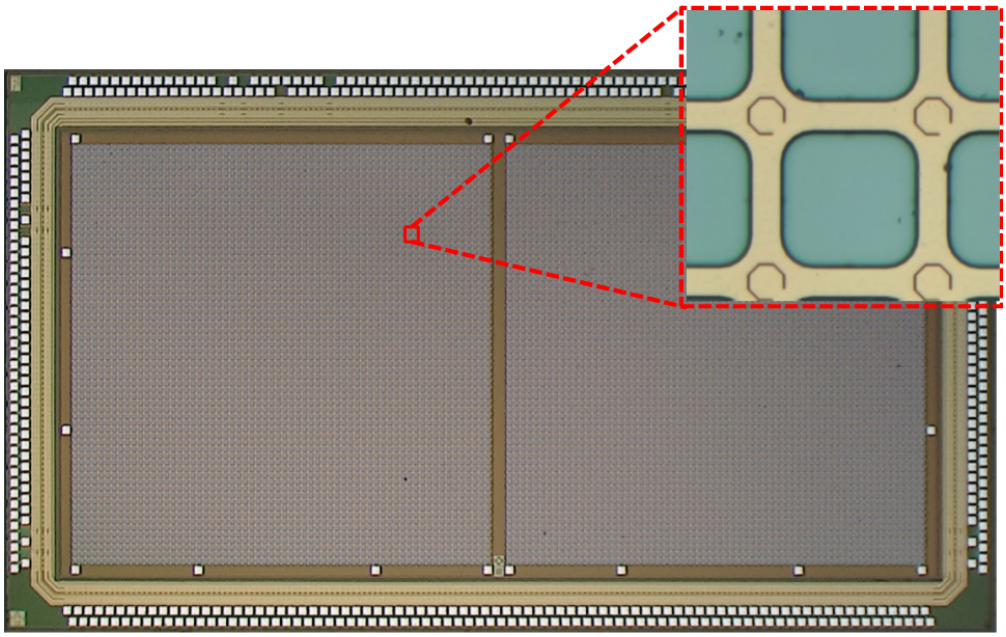
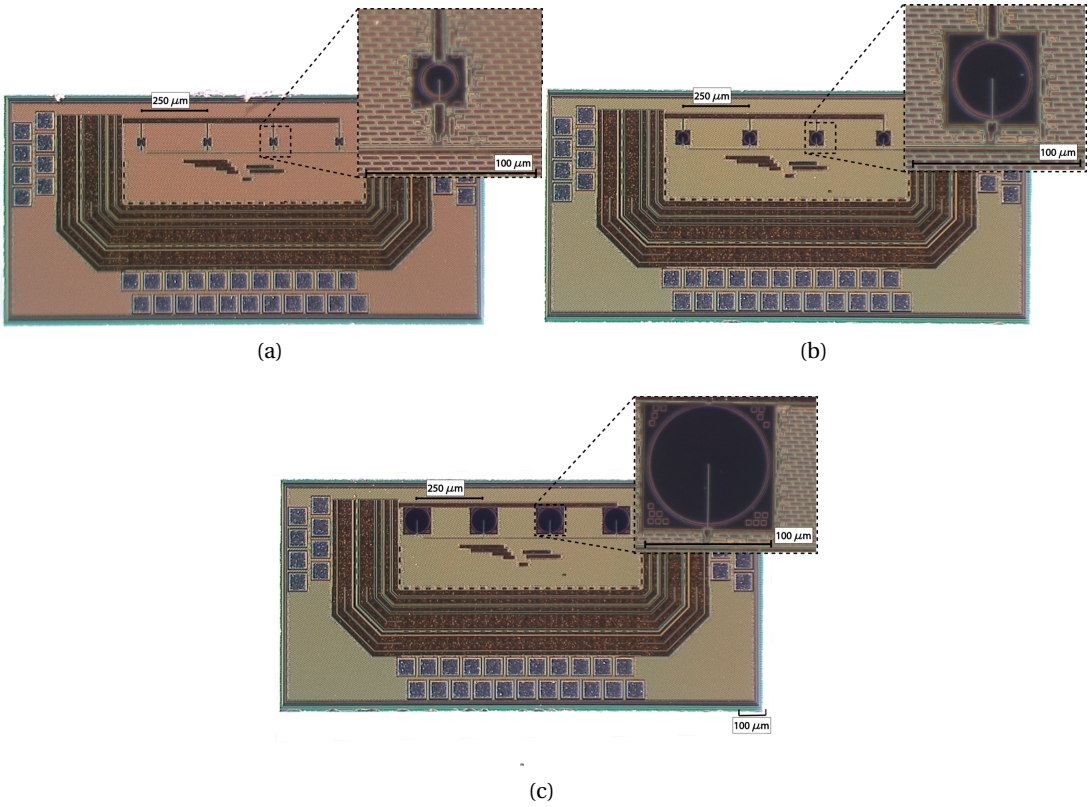


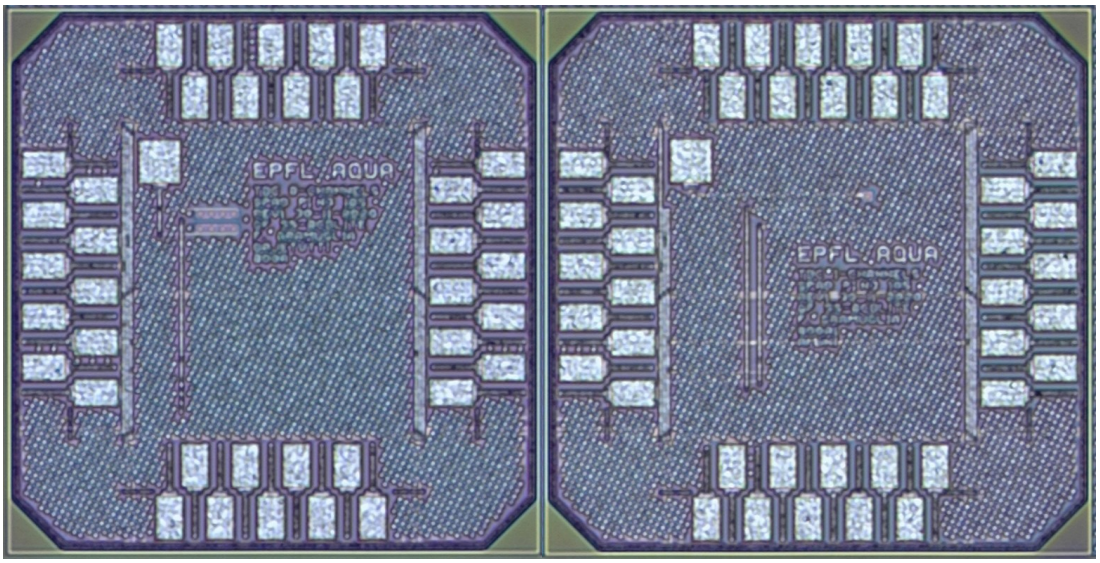
Figure 7.1: Chip gallery: 3D-stacked multi-digital SiPM.



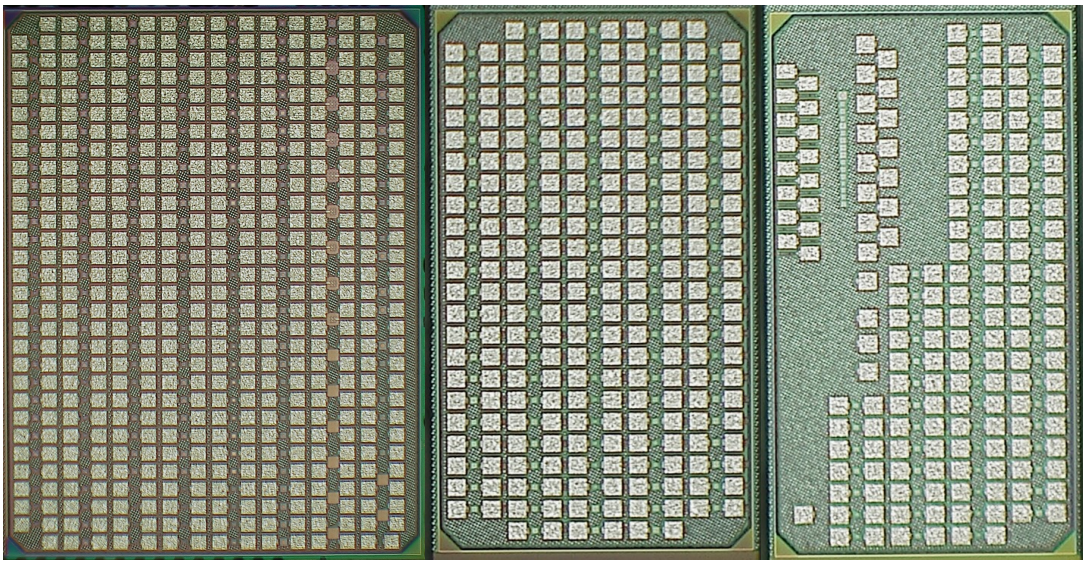
(a) (b)

(c)

Figure 7.2: SPAD-based sensors in 180 nm CMOS technology. 25 μm diam. (*a*), 50 μm diam. (*b*), and 100 μm diam. (*c*).



(a) (b)



(c) (d) (e)

Figure 7.3: (*a*) – (*b*) TDC structures, not presented in the thesis. (*c*) – (*e*) SPAD farms implemented.

A SPAD Metrology

This chapter provides the reader with a description of the setups and the measurements protocols needed for the complete characterization of SPAD devices. In addition, each section discusses one single performance parameter and how to extract it.

A.1 I-V Measurement and Breakdown Voltage Estimation

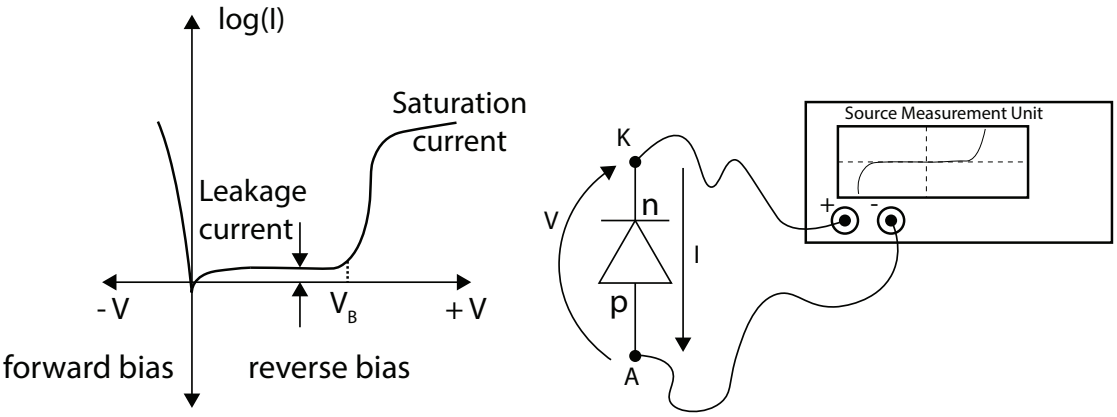


Figure A.1: The I-V characteristic of a SPAD is divided in three main region: the forward bias, the reverse bias below breakdown and the reverse bias above breakdown (*left*). Above breakdown the current increase up to a saturation value on the order of few milliampere. To measure the I-V characteristic, a SPAD anode and cathode are directly connected to a source and measurement unit. A voltage V is swept on the terminals and the resulting current I is measured (*right*).

The first measurement usually performed on a SPAD is the I-V characteristic. The I-V characteristic provides a wealth of information about the device under test (DUT). Indeed, it can be used to highlight the diode behavior of the device, showing the forward bias voltage threshold of the diode as well as its reverse breakdown. In addition, it can also provide a qualitative indication of the noise level of the device and its gain.
The measurement of the I-V curve is performed using a device capable to force a voltage and measure the current through the device, like a source measurement unit (SMU) (or a semiconductor analyzer). The anode and cathode terminals of the DUT are directly connected to the pins of the SMU. The voltage is swept across the SPAD with the SMU covering the entire range of interest. For the sweep, the integration time per point is set at around 100 ms to guarantee enough time to acquire a stable current value, approaching a steady state measurement regime. The measurement is performed with an unidirectional voltage sweep, starting from 0 V (or slightly direct bias) to strong reverse bias voltages. Fig. A.1 shows the expected I-V characteristic of a SPAD (*left*) together with a simplified scheme of the experimental setup (*right*). The SPAD breakdown can vary from less than 10 V up to more than 50 V. This value is strictly linked to the SPAD structure and is a function of the local doping level at the junction and temperature.
The voltage sweep is performed in two environmental conditions: under illumination and

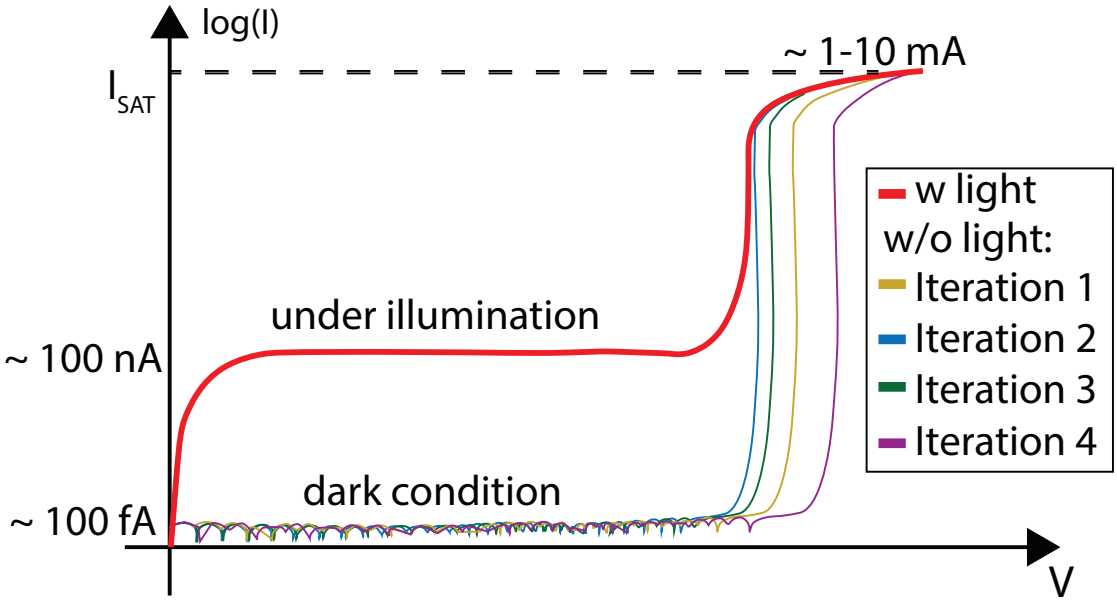


Figure A.2: Illustration of the SPAD I-V characteristic is measured under both illumination and dark condition. Under constant illumination the breakdown is clearly highlighted. In dark condition the leakage current is on the order of 100 fA. Without any source of light, depending of the noise of the device, the breakdown can be overestimated in some iterations of the measurement. This overestimation is linked to the speed of the voltage sweep and it depends on the probability that a noise carrier will trigger an avalanche. This effect is more evident for low DCR devices, and it can be eliminated with a longer integration time for each voltage step.

in the dark. Providing illumination during the measurement highlights the breakdown of the device, and, inside the voltage range where the device has a proportional behavior, will show a higher leakage current due to photo-generated carriers. On the other hand, when the measurement is performed in the dark, there are no photo-generated carriers in the depletion region, and the leakage current is minimal. Typical values for the leakage current in the dark are between 100 fA and 1 pA. The use of triaxial cables is suggested to reduce the leakage contribution of the cable itself when measuring such a small current.
Fig. A.2 shows a qualitative example of I-V characteristic measurements taken under illumination and in the dark. As mentioned, without illumination, there are no photogenerated carriers in the depletion region of the SPAD. So, when the bias voltage exceeds the breakdown in this configuration, the avalanche current is triggered by the noise, i.e., DCR. If the DCR rate is relatively low, it can happen that during the voltage sweep, the avalanche current is not triggered when the voltage reaches the breakdown but at a higher value. This overestimation of the breakdown voltage suggests in a qualitative way that the device has a low noise level. This overestimation is also related to the speed of the voltage sweep and has a statistical behavior between measurement iterations. This effect will therefore be less evident when increasing the integration time per point.
The device saturation current value, not limited by the SMU, provides indirect information

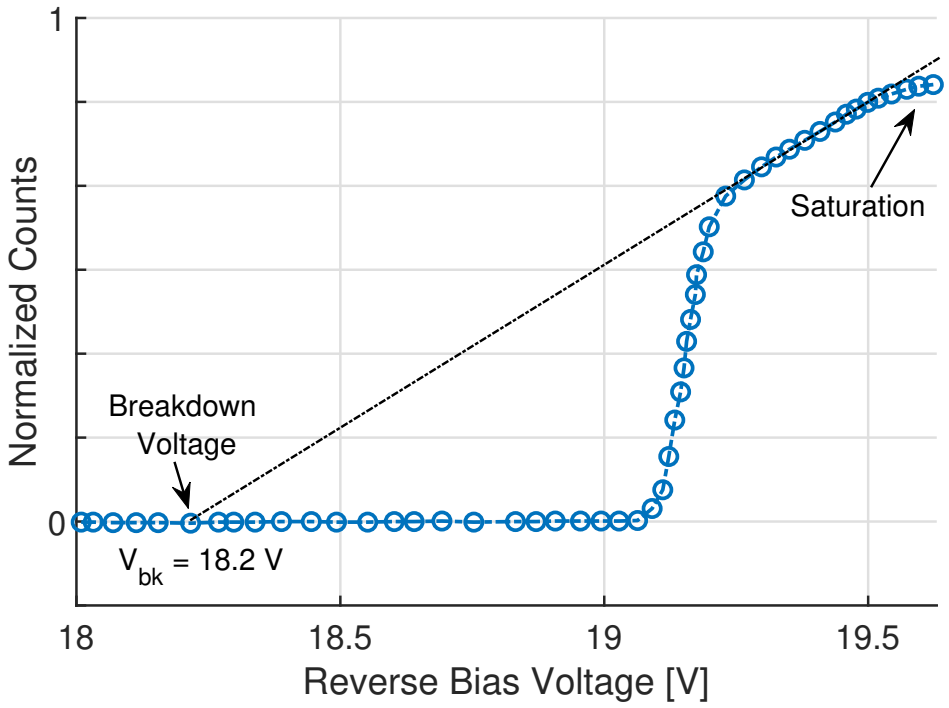


Figure A.3: Count rate dependency on the reverse bias voltage. The breakdown voltage can be extrapolated by linear fit of the data. In the example the breakdown, estimated using the zero crossing of the fit line, is 18.2V.

on the minimum quenching load that can be used to stop the avalanche process. Indeed, as mentioned in the previous chapter, the avalanche current stabilizes to a self-sustaining level and will keep flowing until the electric field level is high enough to guarantee a sufficiently high ionization rate. To turn off the avalanche, we then need to reduce the voltage across the SPAD terminals. If a standard resistor is used, its minimum value will be given by:

$$R_{L,min} = \frac{V_{ex}}{I_{sat}}, \tag{A.1}$$

where V_{ex} is the excess bias voltage that we want to apply to the device, and I_{sat} is the saturation current generated by the SPAD.
As an example, if the saturation current of our device is ~1 mA, we could use a 10 kΩ resistor to quench it. However, in order to limit the current flowing into the device and to reduce the risk of damage, it is suggested to use a higher value.
A straightforward and commonly used method to estimate the breakdown voltage of the SPAD when using the data from the I-V characteristic consists in applying, on a semi-log plot, a linear fit to the data in the region where the current increases abruptly, taking the intercept within the x-axis as the breakdown value. However, another essential point to consider while performing this kind of measurement is that the breakdown of a device is susceptible to temperature

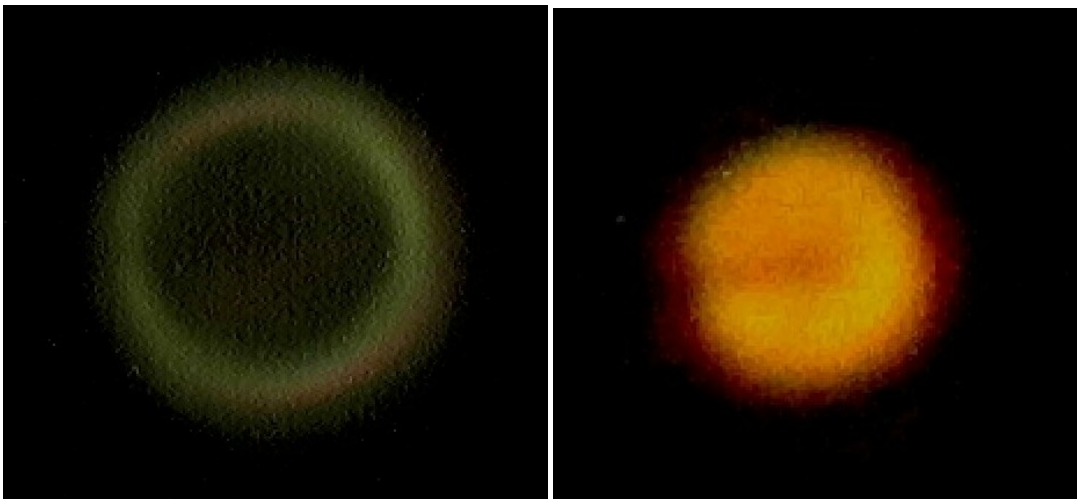


Figure A.4: Typical light emission test of two different SPAD devices, one suffering from premature edge breakdown (*left*) and one operating correctly (*right*).

variations [382]–[385]. In particular, it decreases with temperature with a rate between few millivolts to few tens of millivolts per degree kelvin. For this reason, it is suggested to perform the measurement in a controlled environment.

In some cases, for example, in pixels with integrated front-end or arrays, we do not have direct access to the SPAD, and we cannot measure the I-V characteristic. In these cases, another technique can be used to estimate the breakdown voltage. This technique is based on monitoring of the count rate dependency over the excess bias voltage. Indeed, by placing the detector in a controlled illumination environment and recording the count rate, changing its bias point, we can observe a variation of the number of pulses generated by the SPADs. In Fig. A.3, we can see an example of this trend. Below a specific excess bias voltage value, the SPAD signal does not reach the threshold of the discriminator used for its front-end, and no pulses are observed at the output. When the V_{ex} approaches the threshold of the discriminator, we can observe an abrupt change in the resulting count rate in the graph. On the other hand, by applying a too high excess bias voltage to the device, we observe a saturation effect due to the dead time of the device. In the voltage range between these two, the count rate can be treated as a linear function of the bias voltage [301]. By linear fitting the portion of the graph before saturation occurs, we can extrapolate the breakdown voltage value taking the zero crossing of the fit line.

A.2 Light Emission Test

The light emission test (LET) is another experiment usually performed on newly designed SPAD devices to verify their functionality. LET is very useful to check the active area of the SPAD and the functionality of the guard ring. Indeed, it can be used to highlight problems like

premature edge breakdown. This test, similarly to the previous one, forces a specific voltage (above breakdown) across the DUT and lets the avalanche current flow through the device. To avoid that the device gets destroyed by the heat generated at the junction (see Section 2.1 for more details), the current should be limited to a reasonable value, usually between 1 mA and 5 mA depending on the SPAD. To perform this measurement, we set a bias point and let the current flowing through the device. While the current flows into the device, the SPAD behaves similarly to an LED, emitting light from the high field region. Moreover, observing the light profile emitted by the SPAD during the LET can provide helpful information to the designer about the uniformity of the DC current flow and the electric field across the SPAD active area. Indeed, the value of the field is related to the amount of emitted light [386]. This effect was studied and also used to implement light sources using the avalanche mechanism [386], [387]. Fig. A.4 shows an example of LET for two devices. The one on the left presents premature edge breakdown, and it is possible to identify an annular shape of the emitted light. On the other hand, a well-functioning device shows a uniform light emission across the entire active area.

A.3 Uniformity Laser Scan

The sensitivity scan can be a valuable test to verify the extension of the SPAD active area and its uniformity. However, this test requires a pretty complex and costly measurement setup. Indeed, it requires an optical setup aimed to focus the laser beam to a spot of a size much smaller than the size featured by the SPAD under test (e.g., a laser spots of $\leq 1\ \mu m$).
The measurement is performed by scanning the surface of the SPAD, with this focused laser, recording the count rate. The scan is performed on a 2D surface, and each step of the laser represents one pixel of the final image. In addition, the integration time per step is set to a specific value to ensure that the recorded counts' average is stable. Thus, this value represents the intensity of the pixel in the final image.
With this method, it is possible to measure the sensitivity of a device all over its active area. The measurement protocol corresponds to the one used for the PDP measurement with the pulsed light technique (see Section A.6) [388], [389].
One of the advantages of this technique is the possibility to check the uniformity of the device's performance. Moreover, it can be helpful to extract the actual active area of the device and to compare it with the geometrical one.
The same setup used for this scan, with minor modifications, can also be used to measure the timing jitter of the DUT and relate it with the position in the active area where the optical photons hit the device [390].
However, the cost of the setup, together with the pretty long measurement time, makes this technique impractical. For this reason, so far, the scanning measurement is not commonly performed during the characterizations.

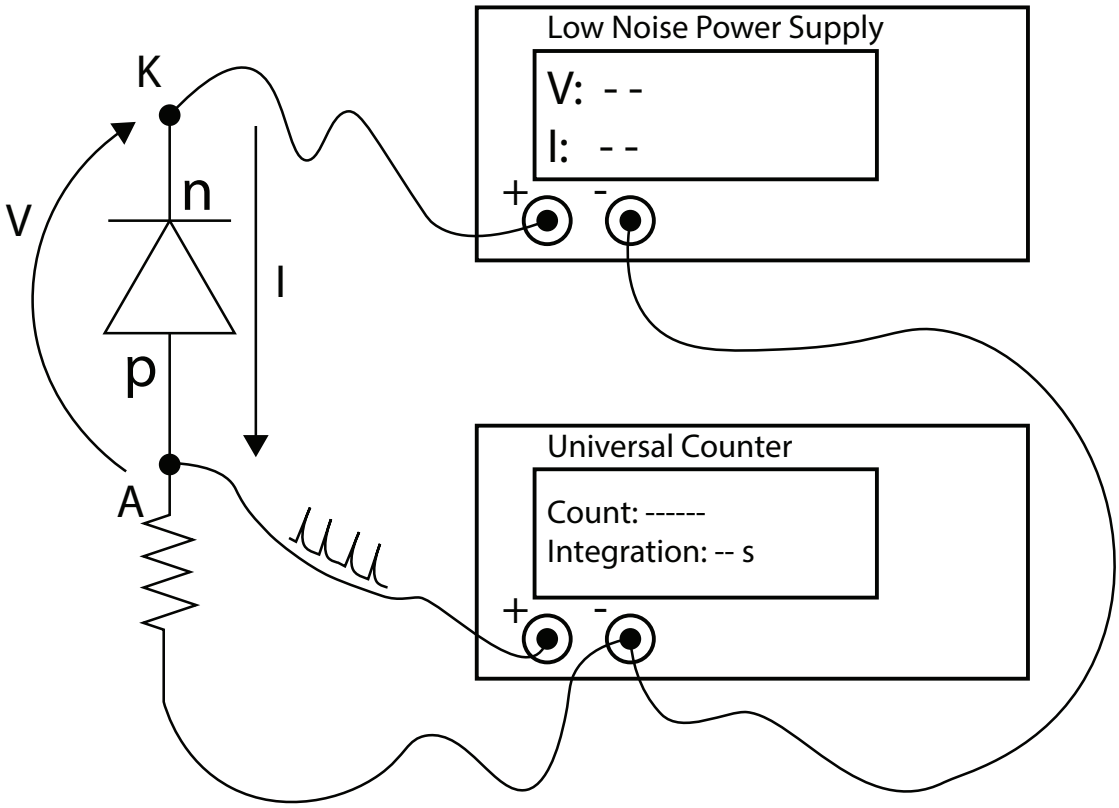
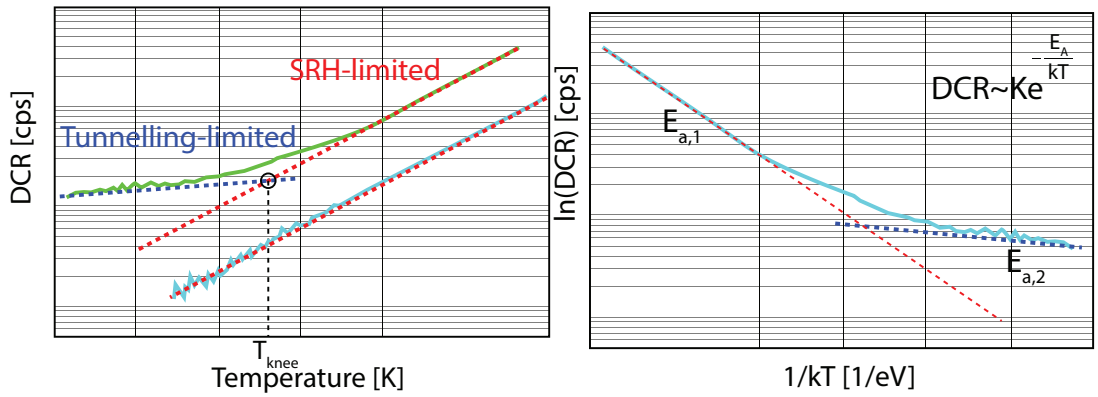


Figure A.5: Dark count rate measurement setup scheme. The SPAD is biased using a low noise power supply and the anode is connected to a load resistor to guarantee a proper quenching. An universal counter is used to count the voltage pulses across the load resistor generated by the avalanche current pulses.

A.4 Dark Count Rate

The measurement of the noise level of the device is one of the most important tests to evaluate the functionality and performance of a SPAD. The dark count rate (DCR), as described in the previous chapter, is composed of a series of pulses generated by avalanche events when the device is in the dark.

The most standard and simple way to measure DCR is to connect the SPAD to a quenching resistor, bias the device to a specific excess bias voltage, and count the voltage pulses across the load. Fig. A.5 shows a scheme of the simple DCR measurement circuit. In this scheme, the pulses are counted using a universal counter, while a low noise power supply provides the bias. The use of a universal counter to perform such measurements brings a major advantage over an oscilloscope in terms of measurement speed and setup cost. Indeed, it overcomes the bottleneck of high-speed sampling and visualization typical of an oscilloscope thanks to a continuous-time operation, and provides a solution at a lower cost. It should also be noted that this simple setup has some intrinsic limitations linked to the passive quenching mechanism used. The long dead time caused by the passive quenching, indeed limit the use



(a) (b)

Figure A.6: (*a*): DCR behavior over temperature for two different cases: tunneling-limited and SHR-limited noise. In the first case, there is a change below a certain temperature in the DCR trend and the noise results to be much less dependent on temperature variations. This temperature (T_{knee}) sets the point where band-to-band tunneling becomes the dominant noise source. In the other case, the trend remains the same allover the temperature range observed, indicating that thermal generation remains the dominant noise source which can be reduced by lowering the temperature. (*b*): The Arrhenius plot calculated from the measurements of DCR over temperature. This plot is derived from the Arrhenius equation and can be used to extract the activation energy for a certain device.

of this setup to low noise devices, i.e., with a DCR level on the order of a few kcps. For higher noise devices, an active reset solution should be used to reduce the SPAD dead time and allow a higher count rate. However, since the noise of a SPAD (particularly the thermal carrier generation) is strongly dependent on temperature, it must be monitored and kept at a precise and known temperature value during the measurement. Indeed, already a temperature shift of 5°C can significantly change the noise value of a device.
For this reason, in the measurement scheme shown in Fig. A.5, the sample is often placed in a temperature chamber. Moreover, the use of a temperature chamber is mandatory to characterize the device's noise over a wide range of temperatures. This kind of characterization can be used to decouple two main noise contributions: thermal generation and band-to-band tunneling. Indeed, although the first one strongly depends on the temperature, the second one does not. When reducing the operating temperature of the device, a *knee* point can thus be found. From this point downward, the temperature trend changes and the DCR becomes almost constant (Fig. A.6 (*a*)). This value does therefore represents the temperature from which the dominant noise contribution changes: above it thermal generation prevails, below it the band-to-band tunneling is the main contribution.
The extraction of the temperature trend of the noise in a SPAD can be helpful to understand the quality of the process. Indeed, from the DCR data, it is possible to construct the so-called Arrhenius plot (Fig. A.6 (*b*)), from which it is possible to extract the activation energy of the dominant element that contributes to the noise. As mentioned in Section 2.1.5, the activation energy is defined as the energy difference between the trap's energy level and the bottom of the

conduction band for electron traps (or the top of the valence band for hole traps). Therefore, the extraction of this parameter can give information on the primary source of noise and can help in identifying the presence of contamination [175]. The activation energy can be extracted from the Arrhenius plot considering the following model [283]:

$$DCR = K \cdot e^{-\frac{E_A}{kT}}. \tag{A.2}$$

A.5 Afterpulsing Probability

Various ways of characterizing the afterpulsing probability of SPAD devices were proposed over the years [136], [147], [172], [176], [178], [180], [183], [184], [391], [392]. However, a unanimous consensus seems not to have been achieved on the methodology to be used. This chapter presents two techniques that have received a good support in the community.

Time Correlated Carrier Counting Technique

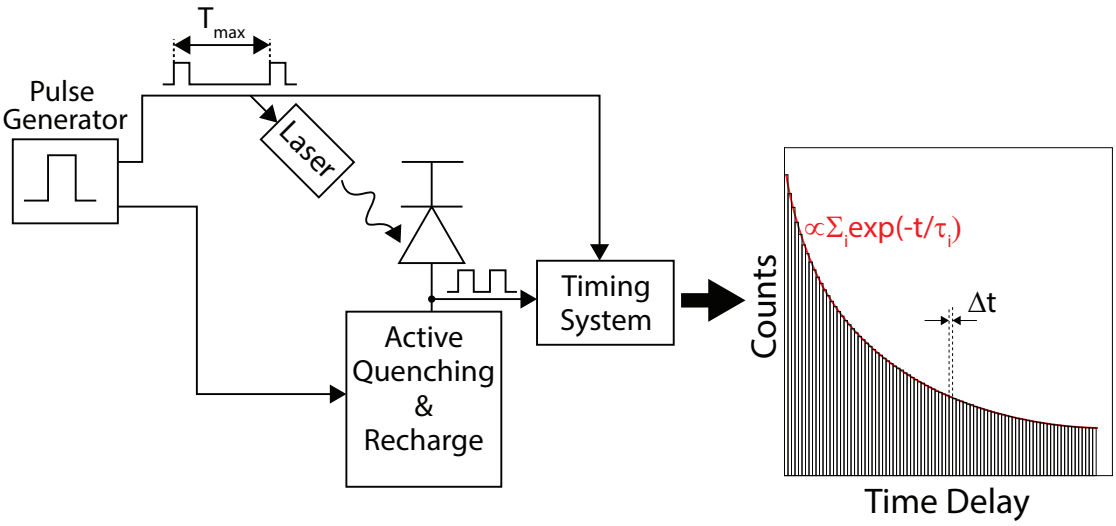


Figure A.7: Time-correlated carrier counting measurement setup.

The first and older method is the so-called time-correlated carrier counting (TCCC). This technique, introduced in 1991, was the first capable of measuring the release of minority carriers emitted from deep level traps [172]. In addition, this technique allows the extraction of the lifetime and activation energy of deep levels in high electric field regions. This technique is actually similar to the one used for fluorescence decays. Indeed, afterpulsing and fluorescence show some similarities linked to their correlated behavior. The TCCC technique consists of the following steps. First of all, a light pulse is used as a stimulus to induce an avalanche in the SPAD and make sure that the traps are populated by carriers. Next, the time interval between

the triggering pulse and the first released carrier detection is recorded. Then, the process is repeated, collecting a histogram of the carrier emission counts versus time.
Fig. A.7 shows the setup used for this measurement. In this setup, the SPAD is connected to a circuit that allows a controlled quenching and recharge, employing active solutions for both. Thanks to this, the SPAD is initially biased below breakdown for a long time (in the millisecond range). This long off-phase ensures that all the traps are emptied. The operating point is then quickly set to a specific excess bias voltage. In this phase, the SPAD is active, and an avalanche is triggered through a laser pulse. The avalanche current is left flowing by the active quenching circuit for a set time to ensure that all the traps are populated, before quenching the device. After this timelapse, the reverse bias is restored across the SPAD until the first released carrier will trigger another avalanche. Finally, the time difference between the optically generated pulse used to populate the traps and the avalanche pulse started by the first released carrier is recorded using a suitable timing circuit. These steps are repeated several times to construct a histogram that, normalized to the total number of cycles, corresponds to the time decay of the carrier emission probability in a specified time window. The histogram data recorded with this method can be analyzed and fitted with a model composed of a sum of exponentials. In [172], where the technique was initially proposed, the authors suggest that a model based on the sum of four exponential decays was the best to fit the data. Moreover, also the background contribution given by the thermally generated carriers should be measured. We can thus write the afterpulsing probability as:

$$P_{ap} = \sum_{i=1}^{4} \left(A_i e^{-\frac{t}{\tau_i}} \right) + B, \tag{A.3}$$

where A_i are amplitudes of the exponential components at ($t = 0$), τ_i are the lifetimes of the traps, B is the background rate, Δt is the histogram bin width. This measurement can be done similarly to what was described above but without the use of the laser. By performing the described measurement at different temperature points it is possible to build the Arrhenius plot of the lifetimes of the various components and then extract their activation energies.

Inter-arrival Time Technique

The Inter-arrival time technique was first proposed in 2003 [176], to simplify the setup needed for the TCCC method [172].
This technique consists of the construction of a histogram using the inter-arrival time between two consecutive avalanche pulses [136], [147], [176]. In SPADs, both the DCR and the photo-generated pulses recorded under dim illumination, which are composed only by primary pulses ($C_{a,pri}$), can be modeled as a Poissonian process. In these cases, the inter-time trend follows an exponential distribution. In a histogram built measuring the inter-arrival time between primary pulses, the area under the exponential fit of the distribution corresponds to the primary pulse count. When we also consider the contribution given by the afterpulsing, we can highlight a distortion in the counts' distribution. In this case, the histogram will

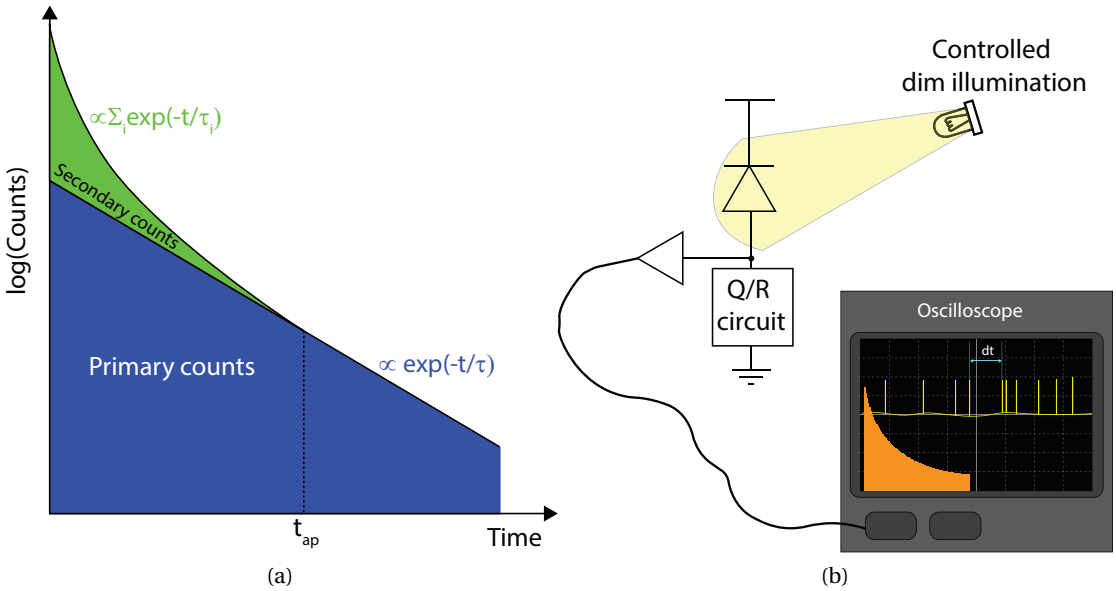


Figure A.8: (*a*): Inter-arrival time histogram example. (*b*): Inter-arrival time measurement setup.

present a multi-exponential behavior (see Fig.A.8 (*a*)). In typical cases, the traps' lifetime in silicon devices does not exceed tens of microseconds. Therefore, if we fit the histogram data with a single exponential starting from more than 10 μs inter-arrival time, the contribution of primary pulses will be the dominant one, and the contribution of the secondary pulses completely negligible. Then, by subtracting the primary pulse count area from the area under the entire histogram (total avalanche count, $C_{a,tot}$), we obtain the secondary pulse count ($C_{a,sec}$). The ratio between the secondary pulse count and the total avalanche count represents the afterpulsing probability:

$$P_{ap} = \frac{C_{a,tot} - C_{a,pri}}{C_{a,tot}} = \frac{C_{a,sec}}{C_{a,tot}}. \tag{A.4}$$

The afterpulsing probability decreases with larger inter-arrival time. This trend can easily be seen in the histogram, and it can identify one point in time (t_{ap}) from where the primary counts become the dominant contribution. Indeed, by setting the dead time of the device to this t_{ap}, the afterpulsing contribution should be practically negligible. To demonstrate that, it is essential to repeat the measurement with different SPAD dead times. This tuning can be easily done using an active solution for the pixel circuit (see Section 2.2.2). For the measurement of the afterpulsing probability by the inter-arrival time technique, the setup is described in the following part of this section (see Fig. A.8 (*b*)). First, a SPAD under test is connected to a quenching and recharge circuit that precisely sets the SPAD dead time. Next, The SPAD is biased to a specific excess bias voltage level. Then, the output of the pixel is connected to the input of an oscilloscope equipped with histogramming functions (or an analogous instrument), used to measure the inter-arrival time and to populate a histogram. Finally, the

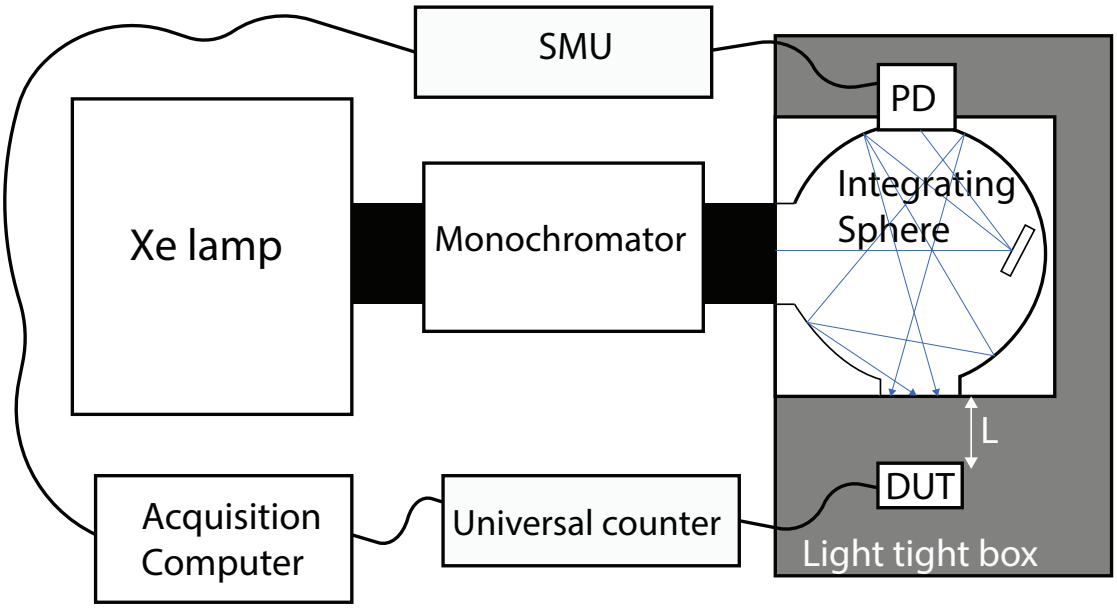


Figure A.9: PDP setup based on the continuous light technique. A Xenon lamp generates a broad spectrum photon flux. Narrow wavelength bands are selected with a monochromator, while a integrating sphere diffuses the photons equally to the reference PD and the DUT. The latter is placed at a calibrated distance L. The setup is enclosed in a light tight box and an universal counter is used to evaluate the SPAD output. Figure adapted from [232]

analysis of the data and the fit is performed offline. The setup is shown in Fig. A.8. This measurement should be performed using the DCR as the source of primary pulses. However, in silicon SPAD, the noise can be as low as few counts per second. The low noise makes the acquisition of significant statistics in the inter-arrival time histogram very slow. To speed up the measurement, the DUT can be illuminated with a constant dim light. The addition of continuous light preserves the Poissonian nature of the primary pulses but increases the count rate. Moreover, the light level should be low enough not to induce too high a pulse rate limiting the maximum inter-arrival time.

A.6 Photon Detection Probability

The light sensitivity of single devices, as discussed in Section 1.2.6, is usually expressed in terms of PDP. This quantity is independent of the geometrical parameters of the device and provides quantitative information on a device's capability of detecting single photons at a specific wavelength. In this section, two ways of measuring this parameter are discussed. The first method is based on the use of a continuous light source to illuminate the device. The second relies on a pulsed light source, such as an LED or a laser.

Continuous light

The first method presented here is also the most commonly used one for measuring the PDP on silicon-based devices. In general, this technique uses a continuous light source to create an area with a uniform photon flux of a specific wavelength. Next, a device is placed inside this area, and its response is acquired. Then, the exact process is repeated with a calibrated reference device (typically a photodiode). Finally, the responsivity of the SPAD under test is compared with the reference calibrated device extracting the PDP. Let us discuss now more in detail the experimental setup (Fig. A.9) and the protocol used for such measurements. A description of the setup used to measure PDP based on the continuous light technique can be found in [232], [290], and it is schematically shown in Fig. A.9. The setup comprises a wide-spectrum Xenon lamp, a monochromator, an integrating sphere, a calibrated reference photodiode (PD) with a precision source-and-measurement-unit (SMU) to measure its photocurrent, and a universal counter connected to the device under test (DUT). The Xenon lamp emits light with a spectrum that ranges from ~200 nm to more than 1000 nm. The lamp output light is then input to a monochromator that allows a precise wavelength selection. After the monochromator, ideally, only a single wavelength is transmitted. Usually the wavelength resolution of a monochromator depends on the slit size, but normally this value is well below 1 nm. This monochromatic light is then injected inside an integrating sphere that uniformly diffuses it. An output window on the integrating sphere allows creating a light beam with a quite uniform photon flux. The DUT is placed at a distance L (about 10 to 15 cm) from the output window of the integrating sphere to ensure a lower light level and even higher flux uniformity [393]. The reason to have a lower light level on the DUT is that the SPADs (sensitive to single photons) can be saturated if exposed to high light levels, thus causing pile-up, which distorts their sensitivity curve, causing an underestimation of the PDP. On another output window of the integrating sphere, a calibrated reference PD is placed to monitor the light level. A higher light level is guaranteed on the reference device by placing the photodiode directly at the output of the integrating sphere. A stronger light impinging on the reference photodiode can actually improve its SNR. In addition, the integrating sphere and the DUT are enclosed in a light-tight box to eliminate any source of background noise that would affect the measurement.

The output of the DUT is connected to a universal counter (or an oscilloscope) that counts the pulses generated by the optically triggered avalanches. The acquisition lasts for the time needed to have an accurate estimation of the mean count value. We found that for most of the SPADs, a 45 s integration time was sufficient at each wavelength position. This value was estimated monitoring the average number of pulses generated by the SPAD every second and after how much time the average converges to a stable value.

Before starting with the sensitivity analysis, the DCR is measured for each excess bias, as explained in section 2.3. These values are then used to compute the PDP at each bias point as shown in [290]:

$$PDP(\lambda) = \eta \frac{S - DCR}{A_{SPAD}} \frac{I_{signal} - I_{dark}}{r(\lambda) A_{PD}} = \eta \frac{S - DCR}{A_{SPAD} F_{PD}(\lambda)}, \tag{A.5}$$

where η is the light power ratio. This parameter is computed during the calibration phase by measuring the light power at the integrating sphere output port (where the PD is located) and at the location of the DUT with another calibrated reference photodiode. In the equation (A.5), S is the number of pulses at the SPAD output when exposed to light; A_{SPAD} is the active area of the SPAD; A_{PD} is the area of the PD used; I_{signal} is the photocurrent generated by the PD under illumination; I_{dark} is the dark current of the PD; $r(\lambda)$ is the responsivity of the PD; $F_{PD}(\lambda)$ is the photon flux detected by the reference photodiode.
This technique allows performing measurements with a small wavelength step ($\leqslant$ 10 nm) on the entire spectrum. This gives the possibility to extract a detailed curve describing the PDP wavelength dependency.

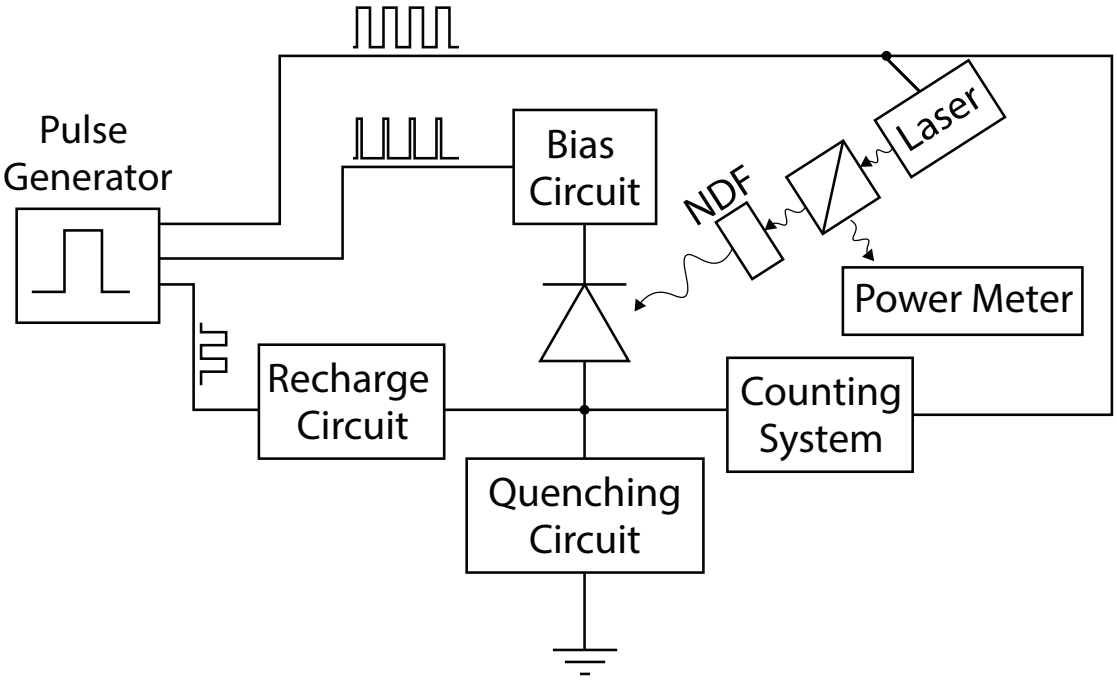


Figure A.10: PDP setup based on the pulsed light technique. The SPAD is connected to a quenching circuit to stop the avalanche. Next, the cathode of the SPAD is connected to a bias circuit that takes as input the output of a pulse generator employed to set and tune the detection window. Then, a reset circuit is used to provide a fast recovery phase. The operation is synchronous to a laser system pulsing light on top of the SPAD within the detection window. A beam splitter splits the laser's beam into two branches. One branch is monitored using a power meter, and a second one is attenuated to a low power level through neutral density filters (NDFs) and aligned with the SPAD.

Pulsed Light Technique

The second technique presented in this section to measure the SPAD PDP is based on a pulsed light source. Although it is more common that such an apparatus is employed to test (III-V)-based SPADs, it is helpful when the DUT presents a relatively high afterpulsing. Indeed, the

afterpulsing increases when the silicon-based device's operating temperature is lowered to a cryogenic level ($\leqslant$ 77 K) [162], [394]. In these cases, gated measurements can still allow precise measurement of the PDP of the devices.

Let us now discuss the measurement procedure used in such a technique. First, the SPAD is connected to a circuit that provides the quenching. Next, a pulse generator provides the bias through a periodic gate. In this configuration, the SPAD is initially biased at a DC value below the breakdown (V_{off}). In this way, all the trapped carriers are released. Then, the pulse quickly brings the bias point to a specific excess bias level and keeps it for a determined timelapse that defines the detection window. After the end of the detection window, the bias is restored to the V_{off} value and the diode is quickly discharged. The gate pulse is kept as short as possible (a typical value is on the order of 100 ps) to help reduce the avalanche current [395].

Moreover the gate pulse is also synchronous to a laser trigger, when measuring the PDP of the SPAD, sending light pulses within the detection window. The laser power that arrives on the DUT should be reduced to avoid distortion of the resulting sensitivity measurement. Thus, an attenuation up to an average level (μ) of 0.1 photons per pulse has been proposed in [395] to obtain accurate sensitivity estimation results. Fig. A.10 shows a schematic view of the presented setup. In this scheme a power meter is added on one branch of the laser to monitor its power and allow a precise calculation of μ. The first step in measuring the sensitivity of a DUT is to measure its DCR. For this purpose, the same setup system is used except for the laser. Assuming that the DCR, defined as the number of pulses per second in dark conditions, is independent of the gate window length, we can measure it by counting the number of pulses detected over a specific timelapse composed of separate gate windows. We can also define the dark count probability (P_d) as:

$$P_d = DCR\tau_g, \tag{A.6}$$

where τ_g is the gate width. Therefore, by acquiring a large number of gates and calculating the percentage of them that contains a dark pulse becomes possible to extract the dark count probability. Then, repeating the measurement with different gate lengths makes it possible to do a linear fit of the data and compute the trend's slope. This value corresponds to the DCR.

Now that we have the value of DCR, let us move to the actual measurement and estimation of the light sensitivity of the detector. In this context, we can define another important quantity, the photon count probability (P_p), which characterize the probability that photons trigger an avalanche. P_p can be expressed as:

$$P_p = \frac{P_t - P_d}{1 - P_d}, \tag{A.7}$$

where P_t is the total count probability, i.e., the total probability of recording a pulse inside a gate, either from the laser, either from the noise. P_t can be easily measured using the same procedure described above for the P_d but with the addition of the pulsed laser synchronous to the gate. Then, we can define the PDP of the DUT with the following expression:

$$PDP = -\frac{1}{\mu} ln\left(1 - P_p\right) = \frac{1}{\mu} ln\left(\frac{1 - P_d}{1 - P_t}\right), \tag{A.8}$$

where μ, as mentioned before, is the average number of photons per laser pulse.

A.7 Single Photon Time Resolution

The precise characterization of the time precision (or jitter) of a single-photon detector, typically expressed in terms of SPTR (see Section 2.1.7), is essential to understand if a device fits the specifications of a specific application. Indeed, in some applications (e.g., TOF-PET), the time resolution of the detector is one of the critical parameters that the designer should try to preserve. As explained in Section 2.1.7, several processes contribute to the jitter of a SPAD. These contributions can be related to the operating condition of the device (bias and temperature) and the illumination [99], [147], [163], [394]. Indeed, the timing jitter of a detector depends both on the excess bias voltage and the photon wavelength adopted during the measurement [396], [397]. Therefore, we need to ensure a controlled illumination and a stable bias to the device, within a controlled environment.
Lasers are the most commonly used light source to measure the STPR performance of a SPAD. Nowadays, lasers are available at almost every wavelength and pulse specification, enabling an accurate characterization of these devices.

General Concepts and Measurement Setup Description

The most straightforward scheme of a typical setup used to measure a SPAD's timing jitter is composed of a laser of a specific wavelength λ pulsed at a repetition rate f, illuminating a DUT. The output pulses generated by the DUT and a reference trigger signal of the laser are input to an oscilloscope. Next, the oscilloscope is used to build the histogram of the IRF, taking the time difference between the DUT pulses and the reference trigger.
Many problems can be highlighted in this simple scheme. Indeed, the first one is related to the amount of light that illuminates the detector. As mentioned in Section 2.1.7, the typical SPAD IRF has a non-Gaussian profile, and it can be distorted if the number of photoelectrons generated in the SPAD (related to the amount of light that illuminates it) is too high. In this case, the typical exponential tail of the distribution given by the carrier diffusion tends to disappear, and the Gaussian behavior of the peak becomes dominant (Fig. 2.12).
For this reason, the laser average power needs to be reduced. A light attenuation can be performed by employing NDF placed along the laser path. Although the ideal case would be to have a single photon impinging on the SPAD active area per laser pulse, this turns out to be very unpractical to implement with standard laser systems. So, a good rule of thumb consists of reducing the laser power until the device detects a minimal percentage of the pulses sent by the laser ($\sim$ 1% being a good compromise) [4], [232]. This allows obtaining an average number of photons per pulse that is less than, or close to, one.
Another critical point to analyze concerns the properties of the laser pulse itself. Indeed, the laser pulse shape will contribute to the timing dispersion measured at the output of the DUT. Assuming a low enough light level and that on average one photon contained in the laser

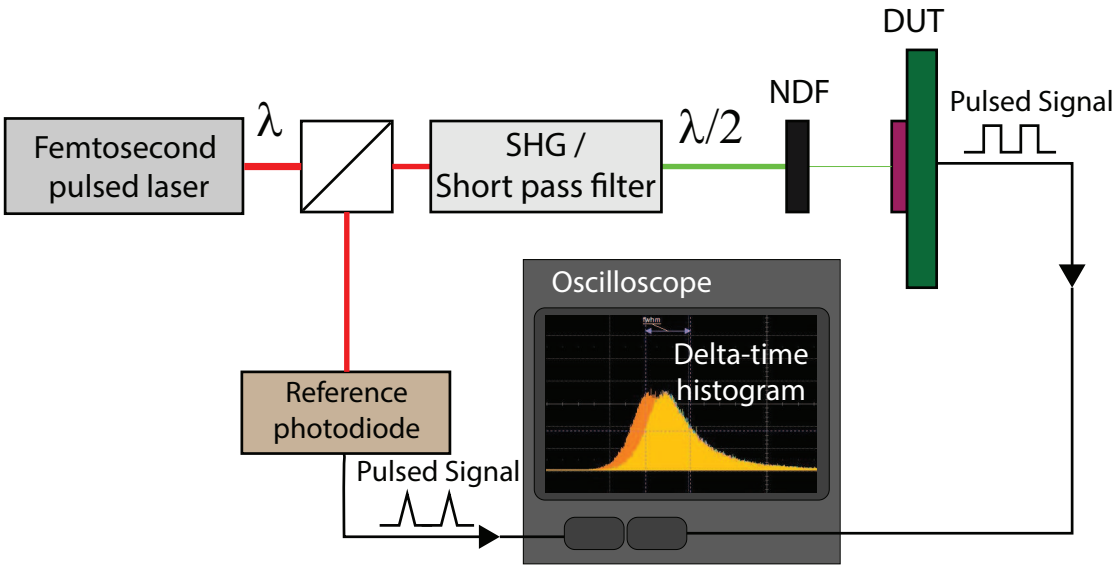


Figure A.11: Timing jitter measurement setup. The setup is composed of an infrared femptosecond laser used as controlled light source. A second harmonic generation system (SHG) is used to obtain visible laser pulses. The laser is then sent to the device under test (DUT) after proper attenuation. The reference, optical rather than electrical, is generated employing a fast photodiode. The output signal of the DUT and the fast photodiode are then send in input to an oscilloscope that builds the inter time histogram. The image is adapted from [232].

pulse will trigger an avalanche, the detection of a pulse can be described as a random process with mean μ_{laser} and variance σ^2_{laser}. If we also assume that the causes of jitter in the device (σ^2_{SPAD}) are statistically independent of the laser pulse, we can write the total time dispersion as:

$$\sigma^2_{tot} = \sigma^2_{laser} + \sigma^2_{SPAD}. \tag{A.9}$$

From equation (A.9), we can understand how the contribution of the laser pulse on the jitter will become more relevant when it approaches the intrinsic SPAD jitter value. For this reason, the pulse shape of the laser used for this measurement should be chosen as narrow as possible. Commonly, lasers with a pulse shape of ~50 ps FWHM or less are employed. Considering a typical SPAD jitter value of about 80 ps, the laser pulse will contribute about 15% to the measured result. These are usually solid-state lasers triggered by a controller that allows tuning the pulse repetition rate and the power. This flexibility helps when the device used is operated employing passive quenching, and it presents a relatively long deadtime (see Section 2.2.1). However, recent works demonstrated how SPAD could achieve SPTR performance well below 50 ps [144], [224], [232]. In these cases, it becomes mandatory to employ more sophisticated devices such as femtosecond pulsed lasers. The latter can provide a pulse shape as narrow as few hundreds of femtoseconds (or even tens of femtoseconds), making the contribution of the laser on the measured timing jitter completely negligible, and allowing to measure the real impulse response of the detector under test.

Moreover, to ensure that the dominant jitter is that of the detector, it is also important to have

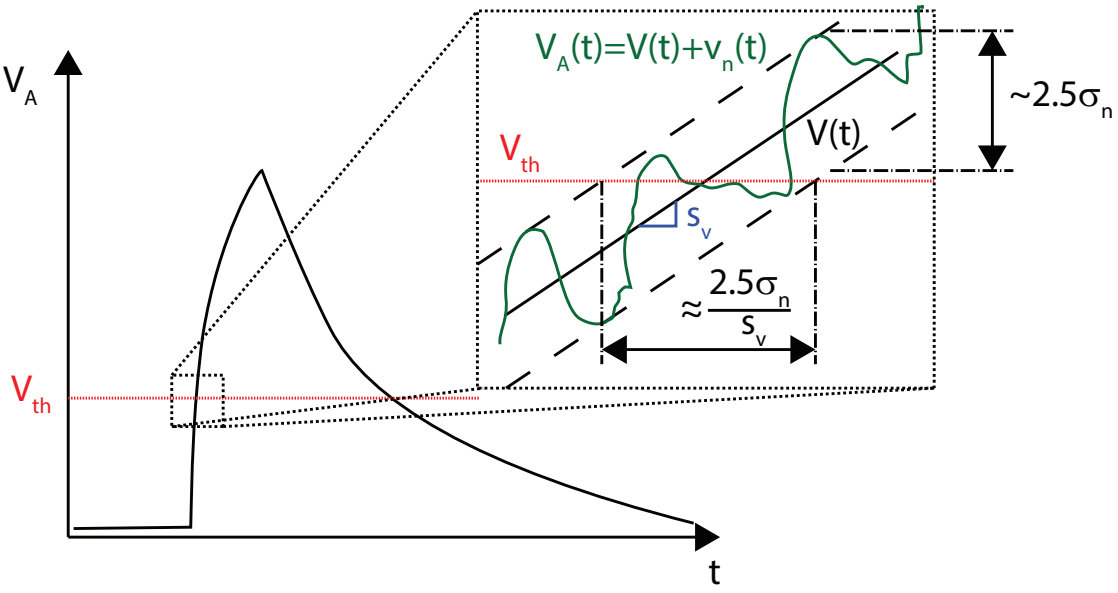


Figure A.12: The noise on a voltage signal can cause deviations from the original waveform. This deviation will cause uncertainty on the time when this signal will cross a fixed threshold V_{th}. This uncertainty can be modeled as a random process and the jitter added to the signal can be calculated to first approximation as the ratio of the amplitude of the noise ($\approx 2.5\sigma_n$) and the slope of the signal (s_v).

a precise reference signal, optical rather than electrical. This signal is generally provided by using a fast photodiode (e.g., InGaAs high-speed optical receiver) connected to a secondary laser branch obtained using a beam splitter [224], [232]. The scheme of the described setup is shown in Fig. A.11. In this example, we consider a typical 1030 nm wavelength femtosecond laser is employed. Since the silicon-based SPADs have an extremely low sensitivity at that wavelength, a second harmonic generation (SHG) block is inserted to obtain a laser pulse at 515 nm, around the peak of the typical sensitivity spectrum of silicon-based SPADs.
One disadvantage of this kind of setup is the typical high pulse repetition rate of these lasers (usually $\geqslant$ 50 MHz) that imposes the use of active solutions (see Section 2.2.2) for the SPAD front-end. Another possible solution is to rely on so-called pulse picking systems. These commercial systems can pick only some of the pulses sent by the laser, de facto reducing the actual pulse frequency at its output. The pulse picking alternative can also allow the use of passive solutions for the SPAD front-end (see Section 2.2.1)

Additional Jitter Sources

It has been demonstrated that the timing jitter of a SPAD is strongly influenced by the probabilistic behavior of the avalanche build-up and spreading phase (see Section 2.1.7). Therefore, the effect on signal time jitter is more significant when the detector size increases. Lowering

the front-end electronics threshold was shown to be effective in mitigating this effect [288]. Indeed, a lower threshold helps detect the pulses when the avalanche is still at the beginning of the build-up phase, reducing the timing spread of the current pulse and therefore the overall jitter [195].

Another critical point to take into account when the aim is to characterize the SPTR of a SPAD detector precisely is the electronic noise present in the system. In parallel to the analysis provided in [398], we could ideally model a SPAD device as a current source charging a capacitor, following the dynamic shown in Section 2.2.1. In this ideal case, if we read out the SPAD pulses with a fixed threshold V_{th}, we can consider that one event will take a time t_{th} to reach the threshold voltage following an exponential envelope. In reality, neglecting for a moment the sources of jitter discussed in Section 2.1.7, a second contribution that should be taken into account includes the electronic noise. This contribution can be modeled with an additional current source in parallel to the one mentioned above, modeling the noise current. This generator will cause the voltage across the output capacitor to deviate from its original envelope. This deviation can be modeled as a random process, which creates an additional source of jitter in the system. To first approximation, the corresponding timing error is obtained by dividing the deviation amplitude ($\approx 2.5\sigma_n$) by the slope of the signal (s_v). Fig. A.12 shows a representation of the jitter introduced by such a noise source. The dependency on the signal slope suggests an evident proportionality between the noise and the jitter and an inverse proportionality between the pulse rise slope and the jitter. We suggest to review [398] for a more detailed analysis of the jitter and its causes in electronic circuits.
Considering what is explained in this section, it should be understood how, during the measurement of the SPTR of a SPAD, it is essential to minimize the sources of noise in the system (e.g., readout noise and supply noise) as well as minimizing the capacitive load applied to the output node of our SPAD device. Indeed, the latter can cause slower rising edge of the signal, which in turn can be seriously affected by the noise contribution along the signal path, increasing the overall jitter. An example is a single SPAD operated with an external load resistor used for quenching. In this case, the avalanche current needs to charge a large load (typically a few picofarads), and the output voltage shows a rise time that often reaches several tens of nanoseconds. Ideally, the best case would be to measure the jitter of a SPAD for which a pixel circuit is integrated next to it. However, this is not always possible. For this reason, some groups developed a discrete version of the pixel circuit. These modules are capable of early detection of the avalanche current and can provide a fast signal at the output. A few examples of these modules are detailed in [136], [206], [212]–[214], [216], [217], [221].

Bibliography

[1] A. Einstein, "On a heuristic point of view about the creation and conversion of light, 1905, translated by wikisource", *Annalen der Physik*, vol. 322, no. 6, pp. 132–148, 1905. DOI: https://doi.org/10.1002/andp.19053220607. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.19053220607. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19053220607.

[2] R. Loudon and L. Hunter, *The quantum theory of light*, 1986.

[3] C. Cohen-Tannoudji, J. Dupont-Roc, and G. Grynberg, *Photons and Atoms-Introduction to Quantum Electrodynamics*. Wiley-Interscience, 1997.

[4] R. H. Hadfield, "Single-photon detectors for optical quantum information applications", *Nature Photonics*, vol. 3, no. 12, pp. 696–705, Dec. 1, 2009, ISSN: 1749-4893. DOI: 10.1038/nphoton.2009.230. [Online]. Available: https://doi.org/10.1038/nphoton.2009.230.

[5] G. N. LEWIS, "The conservation of photons", *Nature*, vol. 118, no. 2981, pp. 874–875, Dec. 1, 1926, ISSN: 1476-4687. DOI: 10.1038/118874a0. [Online]. Available: https://doi.org/10.1038/118874a0.

[6] P. A. M. Dirac and N. H. D. Bohr, "The quantum theory of the emission and absorption of radiation", *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 114, no. 767, pp. 243–265, 1927. DOI: 10.1098/rspa.1927.0039. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1927.0039. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1927.0039.

[7] E. Fermi, "Quantum theory of radiation", *Rev. Mod. Phys.*, vol. 4, pp. 87–132, 1 Jan. 1932. DOI: 10.1103/RevModPhys.4.87. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.4.87.

[8] A. B. Arons and M. B. Peppard, "Einstein's proposal of the photon concept-a translation of the annalen der physik paper of 1905", *American Journal of Physics*, vol. 33, no. 5, pp. 367–374, 1965. DOI: 10.1119/1.1971542. eprint: https://doi.org/10.1119/1.1971542. [Online]. Available: https://doi.org/10.1119/1.1971542.

[9] In *The Old Quantum Theory*, ser. The Commonwealth and International Library: Selected Readings in Physics, D. T. Haar, Ed., Pergamon, 1967, pp. 205–206, ISBN: 978-0-08-012102-4. DOI: https://doi.org/10.1016/B978-0-08-012102-4.50020-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080121024500206.

[10] S. K. Poultney, "Single photon detection and timing: experiments and techniques", in ser. Advances in Electronics and Electron Physics, L. Marton and C. Marton, Eds., vol. 31, Academic Press, 1972, pp. 39–117. DOI: https://doi.org/10.1016/S0065-2539(08)60222-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S006525390860222X.

[11] M. D. Eisaman, J. Fan, A. Migdall, and S. V. Polyakov, "Invited review article: single-photon sources and detectors", *Review of Scientific Instruments*, vol. 82, no. 7, p. 071 101, 2011. DOI: 10.1063/1.3610677. eprint: https://doi.org/10.1063/1.3610677. [Online]. Available: https://doi.org/10.1063/1.3610677.

[12] M. A. Green and M. J. Keevers, "Optical properties of intrinsic silicon at 300 K", *Progress in Photovoltaics: Research and Applications*, vol. 3, pp. 189 –192, 1995,
, ISSN: 1099159X. DOI: 10.1002/pip.4670030303.

[13] E. Nocerino, "The semiconductor multiplication system for photoelectrons in a vacuum silicon photomultiplier tube and related front end electronics", Ph.D. dissertation, Nov. 2016.

[14] D. Renker, "Geiger-mode avalanche photodiodes, history, properties and problems", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 567, no. 1, pp. 48–56, 2006, Proceedings of the 4th International Conference on New Developments in Photodetection, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2006.05.060. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900206008680.

[15] X Michalet, O. SIEGMUND, J. Vallerga, P Jelinsky, J. MILLAUD, and S. Weiss, "Detectors for single-molecule fluorescence imaging and spectroscopy", *Journal of modern optics*, vol. 54, p. 239, Jan. 2007. DOI: 10.1080/09500340600769067.

[16] J. Hamamatsu Photonics K. K., *Photomultiplier tubes basics and applications - fourth edition*, 2017.

[17] Q. Peng, W.-S. Choong, and W. W. Moses, "Evaluation of the timing properties of a high quantum efficiency photomultiplier tube", *IEEE Transactions on Nuclear Science*, vol. 60, no. 5, pp. 3212–3219, 2013. DOI: 10.1109/TNS.2013.2281176.

[18] T. Szczesniak, M. Moszynski, L. Swiderski, *et al.*, "A comparative study of fast photomultipliers for timing experiments and TOF PET", *Nuclear Science, IEEE Transactions on*, vol. 56, pp. 1017 –1023, Jul. 2009. DOI: 10.1109/TNS.2008.2011482.

[19] G. N. Gol'tsman, O. Okunev, G. Chulkova, *et al.*, "Picosecond superconducting single-photon optical detector", *Applied Physics Letters*, vol. 79, no. 6, pp. 705–707, 2001. DOI: 10.1063/1.1388868. eprint: https://doi.org/10.1063/1.1388868. [Online]. Available: https://doi.org/10.1063/1.1388868.

[20] C. M. Natarajan, M. G. Tanner, and R. H. Hadfield, "Superconducting nanowire single-photon detectors: physics and applications", *Superconductor Science and Technology*, vol. 25, no. 6, p. 063 001, 2012. DOI: 10.1088/0953-2048/25/6/063001. [Online]. Available: https://doi.org/10.1088/0953-2048/25/6/063001.

[21] I. Holzman and Y. Ivry, "Superconducting nanowires for single-photon detection: progress, challenges, and opportunities", *Advanced Quantum Technologies*, vol. 2, no. 3-4, p. 1 800 058, 2019. DOI: https://doi.org/10.1002/qute.201800058. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qute.201800058. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/qute.201800058.

[22] I. Esmaeil Zadeh, J. W. N. Los, R. B. M. Gourgues, *et al.*, "Single-photon detectors combining high efficiency, high detection rates, and ultra-high timing resolution", *APL Photonics*, vol. 2, no. 11, p. 111 301, 2017. DOI: 10.1063/1.5000001. eprint: https://doi.org/10.1063/1.5000001. [Online]. Available: https://doi.org/10.1063/1.5000001.

[23] E. E. Wollman, V. B. Verma, A. D. Beyer, *et al.*, "UV superconducting nanowire single-photon detectors with high efficiency, low noise, and 4 K operating temperature", *Opt. Express*, vol. 25, no. 22, pp. 26 792–26 801, 2017. DOI: 10.1364/OE.25.026792. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-25-22-26792.

[24] B. Korzh, Q.-Y. Zhao, J. P. Allmaras, *et al.*, "Demonstration of sub-3 ps temporal resolution with a superconducting nanowire single-photon detector", *Nature Photonics*, vol. 14, no. 4, pp. 250–255, 2020.

[25] A. Vetter, S. Ferrari, P. Rath, *et al.*, "Cavity-enhanced and ultrafast superconducting single-photon detectors", *Nano letters*, vol. 16, no. 11, pp. 7085–7092, 2016.

[26] 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*, p. 2 000 102, 2020.

[27] C. L. Lv, H Zhou, H Li, *et al.*, "Large active area superconducting single-nanowire photon detector with a 100 μm diameter", *Superconductor Science and Technology*, vol. 30, no. 11, p. 115 018, 2017. DOI: 10.1088/1361-6668/aa8e28. [Online]. Available: https://doi.org/10.1088/1361-6668/aa8e28.

[28] F. Marsili, V. B. Verma, J. A. Stern, *et al.*, "Detecting single infrared photons with 93% system efficiency", *Nature Photonics*, vol. 7, no. 3, pp. 210–214, 2013.

[29] F. Flamini, N. Spagnolo, and F. Sciarrino, "Photonic quantum information processing: a review", *Reports on Progress in Physics*, vol. 82, no. 1, p. 016 001, 2018. DOI: 10.1088/1361-6633/aad5b2. [Online]. Available: https://doi.org/10.1088/1361-6633/aad5b2.

[30] B. F. Aull, E. K. Duerr, J. P. Frechette, K. A. McIntosh, D. R. Schuette, and R. D. Younger, "Large-Format Geiger-Mode Avalanche Photodiode Arrays and Readout Circuits", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 2, pp. 1–10, 2018. DOI: 10.1109/JSTQE.2017.2736440.

[31] M. A. Itzler, X. Jiang, M. Entwistle, *et al.*, "Advances in InGaAsP-based avalanche diode single photon detectors", *Journal of Modern Optics*, vol. 58, no. 3-4, pp. 174–200, 2011.

[32] A. Tosi, F. Acerbi, M. Anti, and F. Zappa, "InGaAs/InP single-photon avalanche diode with reduced afterpulsing and sharp timing response with 30 ps tail", *IEEE Journal of quantum electronics*, vol. 48, no. 9, pp. 1227–1232, 2012.

[33] A. Tosi, N. Calandri, M. Sanzaro, and F. Acerbi, "Low-noise, low-jitter, high detection efficiency InGaAs/InP single-photon avalanche diode", *IEEE Journal of selected topics in quantum electronics*, vol. 20, no. 6, pp. 192–197, 2014.

[34] P. Vines, K. Kuzmenko, J. Kirdoda, *et al.*, "High performance planar germanium-on-silicon single-photon avalanche diode detectors", *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.

[35] P. Padmanabhan, "Direct time-of-flight spad image sensors for light detection and ranging", p. 161, 2021. DOI: 10.5075/epfl-thesis-8231. [Online]. Available: http://infoscience.epfl.ch/record/282653.

[36] P. F. McManamon, *LiDAR Technologies and Systems*. Spie Press Book, 2019.

[37] M. E. Warren, "Automotive LIDAR technology", in *2019 Symposium on VLSI Circuits*, 2019, pp. C254–C255. DOI: 10.23919/VLSIC.2019.8777993.

[38] A. R. Ximenes, "Modular time-of-flight image sensor for light detection and ranging", Ph.D. dissertation, Delft University of Technology, Delft, Netherlands, Jul. 2019.

[39] A. J. Carimatto, "High speed electronics for spad image sensors used in Time-of-Flight applications", Ph.D. dissertation, Delft University of Technology, Delft, Netherlands, Jul. 2020.

[40] S. W. Hutchings, N. Johnston, I. Gyongy, *et al.*, "A reconfigurable 3-D-Stacked SPAD imager with in-pixel histogramming for flash lidar or high-speed time-of-flight imaging", *IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 2947–2956, 2019. DOI: 10.1109/JSSC.2019.2939083.

[41] C. Niclass, M. Soga, H. Matsubara, S. Kato, and M. Kagami, "A 100-m range 10-frame/s 340×96-pixel time-of-flight depth sensor in 0.18 μm cmos", *IEEE Journal of Solid-State Circuits*, vol. 48, no. 2, pp. 559–572, 2013. DOI: 10.1109/JSSC.2012.2227607.

[42] P. A. Schubiger, L. Lehmann, and M. Friebe, *PET chemistry*. Springer, 2007, vol. 64.

[43] H. Young, P. Carnochan, J. Zweit, J. Babich, S. Cherry, and R. Ott, "Evaluation of copper(II)-pyruvaldehyde bis (N-4-methylthiosemicarbazone) for tissue blood flow measurement usina a trapped tracer model", *European Journal of Nuclear Medicine*, vol. 21, pp. 336–341, 2004.

[44] C Burtea, S Laurent, L. Elst, and R. Muller, "Molecular imaging I: handbook of experimental pharmacology", 2008.

[45] T. Massoud and S. Gambhir, "Molecular imaging in living subjects: seeing fundamental biological processes in a new light.", *Genes & development*, vol. 17 5, pp. 545–80, 2003.

[46] C. S. Levin, "New imaging technologies to enhance the molecular sensitivity of positron emission tomography", *Proceedings of the IEEE*, vol. 96, no. 3, pp. 439–467, 2008. DOI: 10.1109/JPROC.2007.913504.

[47] M. L. James and S. S. Gambhir, "A molecular imaging primer: modalities, imaging agents, and applications", *Physiological Reviews*, vol. 92, no. 2, pp. 897–965, 2012, PMID: 22535898. DOI: 10.1152/physrev.00049.2010. eprint: https://doi.org/10.1152/physrev.00049.2010. [Online]. Available: https://doi.org/10.1152/physrev.00049.2010.

[48] W. Semmler and M. Schwaiger, "Molecular imaging II", *Springer*, pp. 145–163, 2008.

[49] E. Venialgo, "PET detector technologies for next-generation molecular imaging: from single-positron counting to single-photoelectron counting", Ph.D. dissertation, Delft University of Technology, Delft, Netherlands, Apr. 2019.

[50] D. L. Bailey, D. W. Townsend, P. E. Valk, *et al.*, *Positon emission tomography. basic sciences| springer-verlag london 2005*, 2006.

[51] S. R. Cherry and M. Dahlbom, "PET: physics, instrumentation, and scanners", in *PET*, Springer, 2006, pp. 1–117.

[52] M. F. L'Annunziata, "1 - nuclear radiation, its interaction with matter and radioisotope decay", in *Handbook of Radioactivity Analysis (Second Edition)*, M. F. L'Annunziata, Ed., Second Edition, San Diego: Academic Press, 2003, pp. 1–121, ISBN: 978-0-12-436603-9. DOI: https://doi.org/10.1016/B978-012436603-9/50006-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124366039500065.

[53] T. K. Lewellen, "Time-of-flight [PET", *Seminars in Nuclear Medicine*, vol. 28, no. 3, pp. 268–275, 1998, The Coming Age of Pet (Part 1), ISSN: 0001-2998. DOI: https://doi.org/10.1016/S0001-2998(98)80031-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001299898800317.

[54] S. Surti, "Update on time-of-flight PET imaging", *Journal of Nuclear Medicine*, vol. 56, no. 1, pp. 98–105, 2015, ISSN: 0161-5505. DOI: 10.2967/jnumed.114.145029. eprint: https://jnm.snmjournals.org/content/56/1/98.full.pdf. [Online]. Available: https://jnm.snmjournals.org/content/56/1/98.

[55] M. M. Ter-Pogossian, D. C. Ficke, J. T. Hood Sr, M. Yamamoto, and N. A. Mullani, "PETT VI: a positron emission tomograph utilizing cesium fluoride scintillation detectors.", *Journal of computer assisted tomography*, vol. 6, no. 1, pp. 125–133, 1982.

[56] M. M. Ter-Pogossian, D. C. Ficke, M. Yamamoto, and J. T. Hood, "Super PETT I: a positron emission tomograph utilizing photon time-of-flight information", *IEEE transactions on medical imaging*, vol. 1, no. 3, pp. 179–187, 1982.

[57] R Gariod, R Allemand, E Cormoreche, M Laval, and M Moszynski, "The LETI positron tomograph architecture and time of flight improvements", in *Proceedings of The Workshop on Time of Flight Tomography*, 1982, pp. 25–29.

[58] W. Wong, N. Mullani, E. Philippe, *et al.*, "Performance characteristics of the University of Texas TOFPET-I PET camera", *Journal of Nuclear Medicine*, vol. 25, no. 5, P46–P47, 1986.

[59] T. Lewellen, A. Bice, R. Harrison, M. Pencke, and J. Link, "Performance measurements of the SP3000/UW time-of-flight positron emission tomograph", *IEEE Transactions on Nuclear Science*, vol. 35, no. 1, pp. 665–669, 1988.

[60] B Mazoyer, R Trebossen, C Schoukroun, *et al.*, "Physical characteristics of TTV03, a new high spatial resolution time-of-flight positron tomograph", *IEEE Transactions on Nuclear Science*, vol. 37, no. 2, pp. 778–782, 1990.

[61] S. I. Kwon, R. Ota, E. Berg, *et al.*, "Ultrafast timing enables reconstruction-free positron emission imaging", *Nature Photonics*, pp. 1–5, 2021.

[62] P. Lecoq, C. Morel, and J. Prior, "Case for setting up a 10 ps challenge: a step toward reconstruction-less tof-pet", *Il nuovo cimento C*, vol. 43, no. 1, pp. 1–7, 2020.

[63] P. Lecoq, C. Morel, J. Prior, *et al.*, "Roadmap toward the 10 ps time-of-flight pet challenge", *Physics in Medicine & Biology*, vol. 65, no. 21, 21RM01, May 2020. DOI: 10.1088/1361-6560/ab9500. [Online]. Available: https://doi.org/10.1088/1361-6560/ab9500.

[64] D. S. McGregor, "Materials for gamma-ray spectrometers: inorganic scintillators", *Annual Review of Materials Research*, vol. 48, no. 1, pp. 245–277, 2018. DOI: 10.1146/annurev-matsci-070616-124247. eprint: https://doi.org/10.1146/annurev-matsci-070616-124247. [Online]. Available: https://doi.org/10.1146/annurev-matsci-070616-124247.

[65] R. Mao, L. Zhang, and R.-Y. Zhu, "Optical and scintillation properties of inorganic scintillators in high energy physics", *IEEE Transactions on Nuclear Science*, vol. 55, no. 4, pp. 2425–2431, 2008. DOI: 10.1109/TNS.2008.2000776.

[66] ——, "Crystals for the HHCAL detector concept", *Journal of Physics: Conference Series*, vol. 404, p. 012 029, 2012. DOI: 10.1088/1742-6596/404/1/012029. [Online]. Available: https://doi.org/10.1088/1742-6596/404/1/012029.

[67] F. Acerbi and S. Gundacker, "Understanding and simulating SiPMs", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 926, pp. 16–35, 2019, Silicon Photomultipliers: Technology, Characterisation and Applications, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2018.11.118. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900218317704.

[68] F. Gramuglia, S. Frasca, E. Ripiccini, *et al.*, "Light extraction enhancement techniques for inorganic scintillators", *Crystals*, vol. 11, no. 4, 2021, ISSN: 2073-4352. DOI: 10.3390/cryst11040362. [Online]. Available: https://www.mdpi.com/2073-4352/11/4/362.

[69] S. Gundacker and A. Heering, "The silicon photomultiplier: fundamentals and applications of a modern solid-state photon detector", *Physics in Medicine & Biology*, vol. 65, no. 17, 17TR01, 2020. DOI: 10.1088/1361-6560/ab7b2d. [Online]. Available: https://doi.org/10.1088/1361-6560/ab7b2d.

[70] S. Brunner, L Gruber, A Hirtl, K Suzuki, J Marton, and D. Schaart, "A comprehensive characterization of the time resolution of the philips digital photon counter", *Journal of Instrumentation*, vol. 11, no. 11, P11004, 2016.

[71] *SensL Technologies, High PDE and Timing Resolution SiPM Sensors in a TSV Package ((accessed April 20, 2018))*.

[72] V Puill, C Bazin, D Breton, *et al.*, "Single photoelectron timing resolution of sipm as a function of the bias voltage, the wavelength and the temperature", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 695, pp. 354–358, 2012.

[73] 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)*, 2009, pp. 1959–1965. DOI: 10.1109/NSSMIC.2009.5402143.

[74] 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, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2015.10.078. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900215013121.

[75] H Wieczorek, A Thon, T Dey, V Khanin, and P Rodnyi, "Analytical model of coincidence resolving time in TOF-PET", *Physics in Medicine and Biology*, vol. 61, no. 12, pp. 4699–4710, 2016. DOI: 10.1088/0031-9155/61/12/4699. [Online]. Available: https://doi.org/10.1088/0031-9155/61/12/4699.

[76] S. Vinogradov, "Approximations of coincidence time resolution models of scintillator detectors with leading edge discriminator", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 912, pp. 149–153, 2018, New Developments In Photodetection 2017, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2017.11.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900217311956.

[77] G Nelson and D. Reilly, "Gamma-ray interactions with matter", *Passive nondestructive analysis of nuclear materials*, vol. 2, pp. 27–42, 1991.

[78] R Dolenec, "Time-of-flight positron emission tomography using cherenkov radiation", Ph.D. dissertation, University of Ljubljana, 2012.

[79] E. Venialgo, S. Mandai, T. Gong, D. R. Schaart, and E. Charbon, "Time estimation with multichannel digital silicon photomultipliers", *Physics in Medicine and Biology*, vol. 60, no. 6, pp. 2435–2452, 2015. DOI: 10.1088/0031-9155/60/6/2435. [Online]. Available: https://doi.org/10.1088/0031-9155/60/6/2435.

[80] S. Mandai, "Multichannel digital silicon photomultipliers for Time-of-Flight PET", Ph.D. dissertation, Delft University of Technology, Delft, Netherlands, Jul. 2014.

[81] W. W. Moses, "Fundamental limits of spatial resolution in PET", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 648, S236–S240, 2011.

[82] T. M. Shaffer, E. C. Pratt, and J. Grimm, "Utilizing the power of cerenkov light with nanotechnology", *Nature nanotechnology*, vol. 12, no. 2, pp. 106–117, 2017.

[83] P. A. Wolff, "Theory of electron multiplication in silicon and germanium", *Phys. Rev.*, vol. 95, pp. 1415–1420, 6 Sep. 1954. DOI: 10.1103/PhysRev.95.1415. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.95.1415.

[84] W. Shockley, "Problems related top-n junctions in silicon", *Czechoslovak Journal of Physics*, vol. 11, no. 2, pp. 81–121, Feb. 1, 1961, ISSN: 1572-9486. DOI: 10.1007/BF01688613. [Online]. Available: https://doi.org/10.1007/BF01688613.

[85] G. A. Baraff, "Distribution functions and ionization rates for hot electrons in semiconductors", *Phys. Rev.*, vol. 128, pp. 2507–2517, 6 Dec. 1962. DOI: 10.1103/PhysRev.128.2507. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.128.2507.

[86] M. Razeghi, *Technology of quantum devices*. Springer, 2010.

[87] P. Webb and R. McIntyre, "Single photon detection with avalanche photodiodes", *Bulletin of the American Physical Society*, vol. 15, p. 813, 1970.

[88] C. L. Anderson and C. R. Crowell, "Threshold energies for electron-hole pair production by impact ionization in semiconductors", *Phys. Rev. B*, vol. 5, pp. 2267–2272, 6 Mar. 1972. DOI: 10.1103/PhysRevB.5.2267. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.5.2267.

[89] A. Antognetti, S. Cova, and A. Longoni, "A study of the operation and performances of an avalanche diode as a single photon detector", in *Proceedings 2nd ISPRA Nuclear Electronics Symposium*, 1975, pp. 453–456.

[90] S. Cova, A. Longoni, and A. Andreoni, "Avalanche photodiodes for single-photon detection", in *9th International Symposium on Photon Detectors*, Budapest: J.Schanda, IMEKO-TC2, 1980, pp. 82–91.

[91] S. Cova, A. Longoni, and G. Ripamonti, "Physical phenomena and performance of avalanche photodiodes for single-photon detection", 1982, pp. 15–16.

[92] P. Webb, R. McIntyre, and J. Conradi, "Properties of avalanche photodiodes", *R.C.A. Review*, vol. 35, pp. 234–278, Jun. 1974.

[93] R. McIntyre, "The distribution of gains in uniformly multiplying avalanche photodiodes: theory", *IEEE Transactions on Electron Devices*, vol. 19, no. 6, pp. 703–713, 1972. DOI: 10.1109/T-ED.1972.17485.

[94] T. Kaneda, H. Matsumoto, and T. Yamaoka, "A model for reach-through avalanche photodiodes (rapd's)", *Journal of Applied Physics*, vol. 47, no. 7, pp. 3135–3139, 1976. DOI: 10.1063/1.323107. eprint: https://doi.org/10.1063/1.323107. [Online]. Available: https://doi.org/10.1063/1.323107.

[95] A. Lacaita, M. Ghioni, and S. Cova, "Double epitaxy improves single-photon avalanche diode performance", English, *Electronics Letters*, vol. 25, 841–843(2), 13 Jul. 1989, ISSN: 0013-5194. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/el_19890567.

[96] S Cova, G Ripamonti, A. L. Lacaita, and M Ghioni, "Epitaxial silicon avalanche photodiodes for single-photon detection with picosecond resolution", in *Solid State Device Research Conference, 1987. ESSDERC '87. 17th European*, 1987, pp. 511–514.

[97] M Ghioni, S Cova, A. L. Lacaita, and G Ripamonti, "Epitaxial silicon single photon detector with reduced carrier diffusion effects and picosecond resolution", in *Electron Devices Meeting, 1987 International (Volume:33)*, 1987, pp. 452–455.

[98] A Lacaita, M Ghioni, and S Cova, "Ultrafast single-photon detector with double epitaxial structure for minimum carrier diffusion effect", in *Solid State Device Research Conference, 1988. ESSDERC '88. 18th European*, vol. C4-49, 1988, pp. 633–636.

[99] S. Cova, A. Lacaita, M. Ghioni, G. Ripamonti, and T. Louis, "20-ps timing resolution with single-photon avalanche diodes", English, *Review of Scientific Instruments*, vol. 60, no. 6, pp. 1104–1110, 1989, ISSN: 0034-6748. DOI: 10.1063/1.1140324.

[100] A. Goetzberger, B. McDonald, R. H. Haitz, and R. M. Scarlett, "Avalanche effects in silicon p-n junctions. II. structurally perfect junctions", *Journal of Applied Physics*, vol. 34, no. 6, pp. 1591–1600, 1963. DOI: 10.1063/1.1702640. eprint: https://doi.org/10.1063/1.1702640. [Online]. Available: https://doi.org/10.1063/1.1702640.

[101] R. McIntyre, "Recent developments in silicon avalanche photodiodes", *Measurement*, vol. 3, no. 4, pp. 146–152, 1985, ISSN: 0263-2241. DOI: https://doi.org/10.1016/0263-2241(85)90024-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0263224185900247.

[102] M Ghioni, S. Cova, A Lacaita, and G. Ripamonti, "New silicon epitaxial avalanche diode for single-photon timing at room temperature", *Electronics Letters*, vol. 24, no. 24, pp. 1476–1477, 1988.

[103] A Lacaita, S Cova, M Ghioni, and F Zappa, "Single-photon avalanche diode with ultrafast pulse response free from slow tails", *IEEE Electron Device Letters*, vol. 14, pp. 360–362, 1993. DOI: 10.1109/55.225573.

[104] B. Aull, A. Loomis, J. Gregory, and D. Young, "Geiger-mode avalanche photodiode arrays integrated with CMOS timing circuits", in *56th Annual Device Research Conference Digest (Cat. No.98TH8373)*, 1998, pp. 58–59. DOI: 10.1109/DRC.1998.731120.

[105] J. Jackson, A. Morrison, D. Phelan, and A. Mathewson, "A novel silicon geiger-mode avalanche photodiode", in *Digest. International Electron Devices Meeting,*, 2002, pp. 797–800. DOI: 10.1109/IEDM.2002.1175958.

[106] B. F. Aull, A. H. Loomis, D. J. Young, *et al.*, "Geiger-mode avalanche photodiodes for three-dimensional imaging", *Lincoln laboratory journal*, vol. 13, no. 2, pp. 335–349, 2002.

[107] E. Sciacca, A. Giudice, D. Sanfilippo, *et al.*, "Silicon planar technology for single-photon optical detectors", *IEEE Transactions on Electron Devices*, vol. 50, no. 4, pp. 918–925, 2003. DOI: 10.1109/TED.2003.812095.

[108] B. Aull, J. Burns, C. Chen, *et al.*, "Laser radar imager based on 3D integration of geiger-mode avalanche photodiodes with two SOI timing circuit layers", in *2006 IEEE International Solid State Circuits Conference - Digest of Technical Papers*, 2006, pp. 1179–1188. DOI: 10.1109/ISSCC.2006.1696163.

[109] M Ghioni, A Gulinatti, P Maccagnani, I Rech, and S Cova, "Planar silicon SPADs with 200 μm diameter and 35 ps photon timing", in *Procedings of SPIE Vol. 6372, Advanced Photon Counting Techniques*, vol. 6372, SPIE (Wolfgang Becker), 2006, 63720R–1–63720R–9. DOI: 10.1117/12.685834.

[110] S. Cova, M. Ghioni, F. Zappa, I. Rech, and A. Gulinatti, "A view on progress of silicon single-photon avalanche diodes and quenching circuits", in *Advanced Photon Counting Techniques*, W. Becker, Ed., International Society for Optics and Photonics, vol. 6372, SPIE, 2006, pp. 123 –134. [Online]. Available: https://doi.org/10.1117/12.685963.

[111] M. Ghioni, A. Gulinatti, I. Rech, F. Zappa, and S. Cova, "Progress in silicon single-photon avalanche diodes", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 13, no. 4, pp. 852–862, 2007. DOI: 10.1109/JSTQE.2007.902088.

[112] M. Ghioni, G. Armellini, P. Maccagnani, I. Rech, M. K. Emsley, and M. S. Unlu, "Resonant-cavity-enhanced single-photon avalanche diodes on reflecting silicon substrates", *IEEE Photonics Technology Letters*, vol. 20, no. 6, pp. 413–415, 2008. DOI: 10.1109/LPT.2008.916926.

[113] A. Gulinatti, F. Panzeri, I. Rech, P. Maccagnani, M. Ghioni, and S. Cova, "Planar silicon SPADs with improved photon detection efficiency", in *Advanced Photon Counting Techniques IV*, SPIE, 2010, pp. 1–12. DOI: 10.1117/12.849664.

[114] A. Gulinatti, F. Panzeri, I. Rech, P. Maccagnani, M. Ghioni, and S. D. Cova, "Planar silicon SPADs with improved photon detection efficiency", in *Quantum Sensing and Nanophotonic Devices VIII*, M. Razeghi, R. Sudharsanan, and G. J. Brown, Eds., International Society for Optics and Photonics, vol. 7945, SPIE, 2011, pp. 634 –643. DOI: 10.1117/12.874685. [Online]. Available: https://doi.org/10.1117/12.874685.

[115] A. Gulinatti, I. Rech, F. Panzeri, *et al.*, "New silicon SPAD technology for enhanced red-sensitivity, high-resolution timing and system integration", *Journal of Modern Optics*, vol. 59, no. 17, pp. 1489–1499, 2012.

[116] A. Gulinatti, F. Ceccarelli, M. Ghioni, and I. Rech, "Custom silicon technology for SPAD-arrays with red-enhanced sensitivity and low timing jitter", *Opt. Express*, vol. 29, no. 3, pp. 4559–4581, 2021. DOI: 10.1364/OE.413821. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-29-3-4559.

[117] E. M. Fisher, "Principles and early historical development of silicon avalanche and geiger-mode photodiodes", in Mar. 2018, ISBN: 978-953-51-3907-2. DOI: 10.5772/intechopen.72148.

[118] E. A. G. Webster, "Single-photon avalanche diode theory, simulation, and high performance CMOS integration", Ph.D. dissertation, The University of Edinburgh, Edinburgh, Scotland, Mar. 2013.

[119] S. Pellegrini, B. Rae, A. Pingault, *et al.*, "Industrialised SPAD in 40 nm technology", in *IEEE International Electron Devices Meeting (IEDM)*, Dec. 2017, pp. 16.5.1–16.5.4. DOI: 10.1109/IEDM.2017.8268404.

[120] A Rochas, M Gani, B Furrer, *et al.*, "Single photon detector fabricated in a complementary metal–oxide–semiconductor high-voltage technology", *Review of Scientific Instruments*, vol. 74, no. 7, pp. 3263–3270, 2003.

[121] C. Niclass, A. Rochas, P. Besse, and E. Charbon, "A CMOS single photon avalanche diode array for 3D imaging", in *2004 International Conference Solid-State Circuits*, ser. 2004 IEEE International Solid-State Circuits Conference, IEEE, 2004. DOI: 10.1109/ISSCC.2004.1332623.

[122] C. Niclass, M. Sergio, and E. Charbon, "A single photon avalanche diode array fabricated in 0.35-μm CMOS and based on an event-driven readout for TCSPC experiments", in *Advanced Photon Counting Techniques*, International Society for Optics and Photonics, vol. 6372, 2006, 63720S.

[123] C. Niclass, M. Gersbach, R. Henderson, L. Grant, and E. Charbon, "A Single Photon Avalanche Diode Implemented in 130-nm CMOS Technology", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 13, no. 4, pp. 863–869, 2007. DOI: 10.1109/JSTQE.2007.903854.

[124] D. Stoppa, F. Borghetti, J. Richardson, *et al.*, "A 32x32-pixel array with in-pixel photon counting and arrival time measurement in the analog domain", in *2009 Proceedings of ESSCIRC*, 2009, pp. 204–207. DOI: 10.1109/ESSCIRC.2009.5325970.

[125] M. Gersbach, Y. Maruyama, E. Labonne, *et al.*, "A parallel 32×32 time-to-digital converter array fabricated in a 130 nm imaging CMOS technology", in *2009 Proceedings of ESSCIRC*, 2009, pp. 196–199. DOI: 10.1109/ESSCIRC.2009.5326021.

[126] J. Richardson, R. Walker, L. Grant, *et al.*, "A 32×32 50 ps resolution 10 bit time to digital converter array in 130 nm CMOS for time correlated imaging", in *2009 IEEE Custom Integrated Circuits Conference*, 2009, pp. 77–80. DOI: 10.1109/CICC.2009.5280890.

[127] C. Veerappan, J. Richardson, R. Walker, *et al.*, "A 160×128 single-photon image sensor with on-pixel 55 ps 10b time-to-digital converter", in *2011 IEEE International Solid-State Circuits Conference*, 2011, pp. 312–314. DOI: 10.1109/ISSCC.2011.5746333.

[128] A. Rochas, M. Gosch, A. Serov, *et al.*, "First fully integrated 2-D array of single-photon detectors in standard CMOS technology", *IEEE Photonics Technology Letters*, vol. 15, no. 7, pp. 963–965, 2003. DOI: 10.1109/LPT.2003.813387.

[129] F. Zappa, A. Gulinatti, P. Maccagnani, S. Tisa, and S. Cova, "SPADA: single-photon avalanche diode arrays", *IEEE Photonics Technology Letters*, vol. 17, no. 3, pp. 657–659, 2005. DOI: 10.1109/LPT.2004.840920.

[130] S. Tisa, A. Tosi, and F. Zappa, "Fully-integrated CMOS single photon counter", *Opt. Express*, vol. 15, no. 6, pp. 2873–2887, Mar. 2007. DOI: 10.1364/OE.15.002873.

[131] R. J. Walker, J. A. Richardson, and R. K. Henderson, "A 128×96 pixel event-driven phase-domain $\Delta\Sigma$-based fully digital 3D camera in 0.13μm CMOS imaging technology", in *2011 IEEE International Solid-State Circuits Conference*, 2011, pp. 410–412. DOI: 10.1109/ISSCC.2011.5746374.

[132] D. Tyndall, B. Rae, D. Li, J. Richardson, J. Arlt, and R. Henderson, "A 100Mphoton/s time-resolved mini-silicon photomultiplier with on-chip fluorescence lifetime estimation in 0.13μm CMOS imaging technology", in *2012 IEEE International Solid-State Circuits Conference*, 2012, pp. 122–124. DOI: 10.1109/ISSCC.2012.6176946.

[133] F. Villa, B. Markovic, S. Bellisai, *et al.*, "SPAD smart pixel for time-of-flight and time-correlated single-photon counting measurements", *IEEE Photonics Journal*, vol. 4, no. 3, pp. 795–804, 2012. DOI: 10.1109/JPHOT.2012.2198459.

[134] S Mandai and E Charbon, "Timing optimization of a H-tree based digital silicon photomultiplier", *Journal of Instrumentation*, vol. 8, no. 09, P09016–P09016, Sep. 2013. DOI: 10.1088/1748-0221/8/09/p09016. [Online]. Available: https://doi.org/10.1088/1748-0221/8/09/p09016.

[135] E. Charbon, "Single-photon imaging in complementary metal oxide semiconductor processes", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 372, no. 2012, p. 20 130 100, 2014. DOI: 10.1098/rsta.2013.0100. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2013.0100. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2013.0100.

[136] C. Veerappan, "Single-photon avalanche diodes for cancer diagnosis", Ph.D. dissertation, Delft University of Technology, Delft, Netherlands, Jul. 2016.

[137] C. Bruschini, H. Homulle, I. M. Antolovic, S. Burri, and E. Charbon, "Single-photon avalanche diode imagers in biophotonics: review and outlook", *Light: Science & Applications*, vol. 8, no. 1, p. 87, 2019, ISSN: 2047-7538. DOI: 10.1038/s41377-019-0191-5.

[138] 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. DOI: 10.1109/JSEN.2015.2483565.

[139] C. Veerappan and E. Charbon, "A substrate isolated CMOS SPAD enabling wide spectral response and low electrical crosstalk", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 299–305, 2014.

[140] C. Niclass, H. Matsubara, M. Soga, M. Ohta, M. Ogawa, and T. Yamashita, "A NIR-sensitivity-enhanced single-photon avalanche diode in 0.18 μm CMOS", in *Proceedings of the International Image Sensor Workshop, Vaals, The Netherlands*, 2015, pp. 8–11.

[141] E. Kamrani, F. Lesage, and M. Sawan, "Premature edge breakdown prevention techniques in CMOS APD fabrication", in *10th IEEE International NEWCAS Conference*, 2012, pp. 345–348. DOI: 10.1109/NEWCAS.2012.6329027.

[142] J. A. Richardson, E. A. G. Webster, L. A. Grant, and R. K. Henderson, "Scaleable Single-Photon Avalanche Diode Structures in Nanometer CMOS Technology", *IEEE Transactions on Electron Devices*, vol. 58, no. 7, pp. 2028–2035, 2011. DOI: 10.1109/TED.2011.2141138.

[143] H. Finkelstein, M. Hsu, and S. Esener, "STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology", *IEEE Electron Device Letters*, vol. 27, no. 11, pp. 887–889, 2006. DOI: 10.1109/LED.2006.883560.

[144] M. Sanzaro, P. Gattari, F. Villa, A. Tosi, G. Croce, and F. Zappa, "Single-photon avalanche diodes in a 0.16 μm BCD technology with sharp timing response and red-enhanced sensitivity", *IEEE Journal of Selected Topics in Quantum Electronics*, pp. 1–1, Oct. 2017. DOI: 10.1109/JSTQE.2017.2762464.

[145] G. A. Petrillo, R. J. McIntyre, R. Lecomte, G. Lamoureux, and D. Schmitt, "Scintillation detection with large-area reach-through avalanche photodiodes", *IEEE Transactions on Nuclear Science*, vol. 31, no. 1, pp. 417–423, 1984. DOI: 10.1109/TNS.1984.4333290.

[146] F. Conti and M. Conti, "Surface breakdown in silicon planar diodes equipped with field plate", *Solid-State Electronics*, vol. 15, no. 1, pp. 93–105, 1972, ISSN: 0038-1101. DOI: https://doi.org/10.1016/0038-1101(72)90070-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0038110172900706.

[147] M. W. Fishburn, "Fundamentals of CMOS single-photon avalanche diodes", Ph.D. dissertation, Delft University of Technology, Delft, Netherlands, Sep. 2012.

[148] I. Janeković, T. Knezević, T. Suligoj, and D. Grubisić, "Optimization of floating guard ring parameters in separate-absorption-and-multiplication silicon avalanche photodiode structure", in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015, pp. 37–41. DOI: 10.1109/MIPRO.2015.7160234.

[149] C. Richard, T. Courcier, P. Pittet, *et al.*, "CMOS buried quad p-n junction photodetector for multi-wavelength analysis", *Opt. Express*, vol. 20, no. 3, pp. 2053–2061, 2012. DOI: 10.1364/OE.20.002053. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-20-3-2053.

[150] R. K. Henderson, E. A. G. Webster, and L. A. Grant, "A dual-junction single-photon avalanche diode in 130-nm CMOS technology", *IEEE Electron Device Letters*, vol. 34, no. 3, pp. 429–431, 2013. DOI: 10.1109/LED.2012.2236816.

[151] Y.-C. Tsai, Y.-N. Zhong, F.-P. Chou, C.-A. Huang, and Y.-M. Hsin, "Silicon photodetectors with triple p–n junctions in CMOS technology at 650- and 850-nm wavelengths", *Electronics Letters*, vol. 52, no. 20, pp. 1707–1708, 2016. DOI: https://doi.org/10.1049/el.2016.2260. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/el.2016.2260. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/el.2016.2260.

[152] S. Ejdehakosh and M. A. Karami, "Electrical μ-lens synthesis using dual-junction single-photon avalanche diode", *iranian journal of electrical and electronic engineering*, vol. 15, pp. 502–508, 2019.

[153] K. Morimoto, "Megapixel SPAD cameras for time-resolved applications", p. 172, 2021. DOI: 10.5075/epfl-thesis-8773. [Online]. Available: http://infoscience.epfl.ch/record/283481.

[154] G. Jegannathan, T. Van den Dries, and M. Kuijk, "Current-assisted SPAD with improved p-n junction and enhanced NIR performance", *Sensors*, vol. 20, no. 24, 2020, ISSN: 1424-8220. DOI: 10.3390/s20247105. [Online]. Available: https://www.mdpi.com/1424-8220/20/24/7105.

[155] W. Jiang, Y. Chalich, R. Scott, and M. J. Deen, "Time-gated and multi-junction SPADs in standard 65 nm CMOS technology", *IEEE Sensors Journal*, vol. 21, no. 10, pp. 12 092–12 103, 2021. DOI: 10.1109/JSEN.2021.3063319.

[156] J. M. Pavia, M. Scandini, S. Lindner, M. Wolf, and E. Charbon, "A 1 × 400 backside-illuminated SPAD sensor with 49.7 ps resolution, 30 pJ/Sample TDCs fabricated in 3D CMOS technology for near-infrared optical tomography", *IEEE Journal of Solid-State Circuits*, vol. 50, no. 10, pp. 2406–2418, 2015. DOI: 10.1109/JSSC.2015.2467170.

[157] T. Al Abbas, N. A. W. Dutton, O. Almer, S. Pellegrini, Y. Henrion, and R. K. Henderson, "Backside illuminated SPAD image sensor with 7.83μm pitch in 3D-stacked CMOS technology", in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 8.1.1–8.1.4. DOI: 10.1109/IEDM.2016.7838372.

[158] S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf, and E. Charbon, "A high-PDE, backside-illuminated SPAD in 65/40-nm 3D IC CMOS pixel with cascoded passive quenching and active recharge", *IEEE Electron Device Letters*, vol. 38, no. 11, pp. 1547–1550, 2017. DOI: 10.1109/LED.2017.2755989.

[159] M.-J. Lee, A. R. Ximenes, P. Padmanabhan, *et al.*, "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. DOI: 10.1109/JSTQE.2018.2827669.

[160] R. K. Henderson, N. Johnston, S. W. Hutchings, *et al.*, "5.7 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)*, 2019, pp. 106–108. DOI: 10.1109/ISSCC.2019.8662355.

[161] K. Ito, Y. Otake, Y. Kitano, *et al.*, "A back illuminated 10μm SPAD pixel array comprising full trench isolation and Cu-Cu bonding with over 14% PDE at 940 nm", in *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 16.6.1–16.6.4. DOI: 10.1109/IEDM13553.2020.9371944.

[162] R. H. Haitz, "Mechanisms contributing to the noise pulse rate of avalanche diodes", *Journal of Applied Physics*, vol. 36, no. 10, pp. 3123–3131, 1965. DOI: 10.1063/1.1702936. eprint: https://doi.org/10.1063/1.1702936. [Online]. Available: https://doi.org/10.1063/1.1702936.

[163] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection", *Applied Optics*, vol. 35, no. 12, pp. 1956–1976, 1996.

[164] W. Shockley and W. T. Read, "Statistics of the recombinations of holes and electrons", *Phys. Rev.*, vol. 87, pp. 835–842, 5 1952. DOI: 10.1103/PhysRev.87.835. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.87.835.

[165] G. A. M. Hurkx, D. B. M. Klaassen, and M. P. G. Knuvers, "A new recombination model for device simulation including tunneling", *IEEE Transactions on Electron Devices*, vol. 39, no. 2, pp. 331–338, Feb. 1992. DOI: 10.1109/16.121690.

[166] T. McLean, *Progress in Semiconductors, vol. 5,edited by A. F. Gibson*. Heywood, 1960.

[167] A. Rochas, "Single photon avalanche diodes in CMOS technology", Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, 2003.

[168] M. A. Karami, M. Gersbach, H.-J. Yoon, and E. Charbon, "A new single-photon avalanche diode in 90nm standard CMOS technology", *Opt. Express*, vol. 18, no. 21, pp. 22 158–22 166, Oct. 2010.

[169] W. J. Kindt, "Geiger mode avalanche photodiode arrays: for spatially resolved single photon counting", Ph.D. dissertation, Delft University of Technology, Delft, Netherlands, Mar. 1999.

[170] S. Sze and K. K. Ng, in *Physics of Semiconductor Devices*. John Wiley & Sons, Ltd, 2006, ISBN: 9780470068328. DOI: https://doi.org/10.1002/9780470068328.fmatter. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470068328.fmatter. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470068328.fmatter.

[171] G. Giustolisi, R. Mita, and G. Palumbo, "Behavioral modeling of statistical phenomena of single-photon avalanche diodes", *International Journal of Circuit Theory and Applications*, vol. 40, no. 7, pp. 661–679, 2012. DOI: https://doi.org/10.1002/cta.748. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cta.748. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cta.748.

[172] S. Cova, A. Lacaita, and G. Ripamonti, "Trapping phenomena in avalanche photodiodes on nanosecond scale", *IEEE Electron Device Letters*, vol. 12, no. 12, pp. 685–687, 1991. DOI: 10.1109/55.116955.

[173] G Ripamonti, F Zappa, and S Cova, "Effects of trap levels in single-photon optical time-domain reflectometry: evaluation and correction", *Journal of Lightwave Technology*, vol. 10, pp. 1398–1402, 1992.

[174] Z. Cheng, X. Zheng, D. Palubiak, M. J. Deen, and H. Peng, "A comprehensive and accurate analytical spad model for circuit simulation", *IEEE Transactions on Electron Devices*, vol. 63, no. 5, pp. 1940–1948, 2016. DOI: 10.1109/TED.2016.2537879.

[175] R. Hull, *Properties of crystalline silicon*, 20. IET, 1999.

[176] A. Giudice, M. Ghioni, S. Cova, and F. Zappa, "A process and deep level evaluation tool: afterpulsing in avalanche junctions", in *ESSDERC '03. 33rd Conference on European Solid-State Device Research, 2003.*, 2003, pp. 347–350. DOI: 10.1109/ESSDERC.2003.1256885.

[177] C. Wang, J. Wang, Z. Xu, *et al.*, "Afterpulsing effects in SPAD-based photon-counting communication system", *Optics Communications*, vol. 443, pp. 202–210, 2019, ISSN: 0030-4018. DOI: https://doi.org/10.1016/j.optcom.2019.03.039. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0030401819302342.

[178] A. W. Ziarkash, S. K. Joshi, M. Stipčević, and R. Ursin, "Comparative study of afterpulsing behavior and models in single photon counting avalanche photo diode detectors", *Scientific Reports*, vol. 8, no. 1, p. 5076, Mar. 22, 2018, ISSN: 2045-2322. DOI: 10.1038/s41598-018-23398-z. [Online]. Available: https://doi.org/10.1038/s41598-018-23398-z.

[179] D. Bronzi, F. Villa, S. Bellisai, *et al.*, "Large-area CMOS SPADs with very low dark counting rate", in *Quantum Sensing and Nanophotonic Devices X*, M. Razeghi, Ed., International Society for Optics and Photonics, vol. 8631, SPIE, 2013, pp. 241 –248. [Online]. Available: https://doi.org/10.1117/12.2004209.

[180] G. Humer, M. Peev, C. Schaeff, S. Ramelow, M. Stipčević, and R. Ursin, "A simple and robust method for estimating afterpulsing in single photon detectors", *Journal of Lightwave Technology*, vol. 33, no. 14, pp. 3098–3107, 2015. DOI: 10.1109/JLT.2015.2428053.

[181] M. Anti, A. Tosi, F. Acerbi, and F. Zappa, "Modeling of afterpulsing in single-photon avalanche diodes", in *Physics and Simulation of Optoelectronic Devices XIX*, B. Witzigmann, F. Henneberger, Y. Arakawa, and A. Freundlich, Eds., International Society for Optics and Photonics, vol. 7933, SPIE, 2011, pp. 371 –378. DOI: 10.1117/12.875143. [Online]. Available: https://doi.org/10.1117/12.875143.

[182] M Ghioni, A Giudice, S Cova, and F Zappa, "High-rate quantum key distribution at short wavelength:performance analysis and evaluation of silicon single photon-avalanche diodes", *Journal of Modern Optics*, vol. 50, pp. 2251–2269, 2003. DOI: 10.1080/0950034031000113677.

[183] M. A. Itzler, X. Jiang, and M. Entwistle, "Power law temporal dependence of InGaAs/InP SPAD afterpulsing", *Journal of Modern Optics*, vol. 59, no. 17, pp. 1472–1480, 2012.

[184] D. B. Horoshko, V. N. Chizhevsky, and S. Y. Kilin, "Afterpulsing model based on the quasi-continuous distribution of deep levels in single-photon avalanche diodes", *Journal of Modern Optics*, vol. 64, no. 2, pp. 191–195, 2017. DOI: 10.1080/09500340.2016.1220643. eprint: https://doi.org/10.1080/09500340.2016.1220643. [Online]. Available: https://doi.org/10.1080/09500340.2016.1220643.

[185] B. Korzh, T. Lunghi, K. Kuzmenko, G. Boso, and H. Zbinden, "Afterpulsing studies of low-noise InGaAs/InP single-photon negative-feedback avalanche diodes", *Journal of Modern Optics*, vol. 62, no. 14, pp. 1151–1157, 2015. DOI: 10.1080/09500340.2015.1024294. eprint: https://doi.org/10.1080/09500340.2015.1024294. [Online]. Available: https://doi.org/10.1080/09500340.2015.1024294.

[186] P. Seitz and A. J. Theuwissen, *Single-photon imaging*. Springer Science & Business Media, 2011, vol. 160.

[187] 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. DOI: 10.1109/T-ED.1972.17544.

[188] A. Gulinatti, I. Rech, M. Assanelli, M. Ghioni, and S. Cova, "A physically based model for evaluating the photon detection efficiency and the temporal response of SPAD detectors", *Journal of Modern Optics*, vol. 58, no. 3-4, pp. 210–224, 2011. DOI: 10.1080/09500340.2010.536590. eprint: https://doi.org/10.1080/09500340.2010.536590. [Online]. Available: https://doi.org/10.1080/09500340.2010.536590.

[189] R. McIntyre, "On the avalanche initiation probability of avalanche diodes above the breakdown voltage", *IEEE Transactions on Electron Devices*, vol. 20, no. 7, pp. 637–641, 1973. DOI: 10.1109/T-ED.1973.17715.

[190] Y. Okuto and C. R. Crowell, "Ionization coefficients in semiconductors: a nonlocalized property", *Phys. Rev. B*, vol. 10, pp. 4284–4296, 10 1974. DOI: 10.1103/PhysRevB.10.4284. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.10.4284.

[191] 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. DOI: 10.1063/1.1147156. eprint: https://doi.org/10.1063/1.1147156. [Online]. Available: https://doi.org/10.1063/1.1147156.

[192] G. Ripamonti and S. Cova, "Carrier diffusion effects in the time-response of a fast photodiode", *Solid-State Electronics*, vol. 28, no. 9, pp. 925–931, 1985, ISSN: 0038-1101. DOI: https://doi.org/10.1016/0038-1101(85)90086-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0038110185900863.

[193] A. Spinelli and A. Lacaita, "Physics and numerical simulation of single photon avalanche diodes", *IEEE Transactions on Electron Devices*, vol. 44, no. 11, pp. 1931–1943, 1997. DOI: 10.1109/16.641363.

[194] C. H. Tan, J. S. Ng, G. J. Rees, and J. P. R. David, "Statistics of avalanche current buildup time in single-photon avalanche diodes", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 13, no. 4, pp. 906–910, 2007. DOI: 10.1109/JSTQE.2007.903843.

[195] A. Ingargiola, M. Assanelli, A. Gallivanoni, I. Rech, M. Ghioni, and S. Cova, "Avalanche buildup and propagation effects on photon-timing jitter in Si-SPAD with non-uniform electric field", in *Advanced Photon Counting Techniques III*, M. A. Itzler and J. C. Campbell, Eds., International Society for Optics and Photonics, vol. 7320, SPIE, 2009, pp. 103–114. [Online]. Available: https://doi.org/10.1117/12.818521.

[196] A. Lacaita, M. Mastrapasqua, M. Ghioni, and S. Vanoli, "Observation of avalanche propagation by multiplication assisted diffusion in p-n junctions", *Applied Physics Letters*, vol. 57, no. 5, pp. 489–491, 1990. DOI: 10.1063/1.103629. eprint: https://doi.org/10.1063/1.103629. [Online]. Available: https://doi.org/10.1063/1.103629.

[197] A. Lacaita, S. Cova, A. Spinelli, and F. Zappa, "Photon-assisted avalanche spreading in reach-through photodiodes", *Applied Physics Letters*, vol. 62, no. 6, pp. 606–608, 1993. DOI: 10.1063/1.108870. eprint: https://doi.org/10.1063/1.108870. [Online]. Available: https://doi.org/10.1063/1.108870.

[198] A. Gulinatti, P. Maccagnani, I. Rech, M. Ghioni, and S. Cova, "35 ps time resolution at room temperature with large area single photon avalanche diodes", *Electronics Letters*, vol. 41, pp. 272 –274, Apr. 2005. DOI: 10.1049/el:20047445.

[199] 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. 165 265, 2021, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2021.165265. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900221002497.

[200] P. Windischhofer and W. Riegler, "The statistics of electron–hole avalanches", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1003, p. 165 327, 2021, ISSN: 0168-9002. DOI:

https://doi.org/10.1016/j.nima.2021.165327. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900221003119.

[201] S. Cova, A. Longoni, and A. Andreoni, "Towards picosecond resolution with single-photon avalanche diodes", *Review of Scientific Instruments*, vol. 52, pp. 408–412, 1981.

[202] S. Cova, A. Longoni, and G. Ripamonti, "Active-quenching and gating circuits for single-photon avalanche diodes (SPADs)", *IEEE Transactions on Nuclear Science*, vol. 29, pp. 599–601, 1982.

[203] 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. DOI: 10.1109/TNS.2010.2074213.

[204] S. Tisa, F. Zappa, A. Tosi, and S. Cova, "Electronics for single photon avalanche diode arrays", *Sensors and Actuators A: Physical*, vol. 140, no. 1, pp. 113–122, 2007, ISSN: 0924-4247. DOI: https://doi.org/10.1016/j.sna.2007.06.022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924424707004979.

[205] R. H. Haitz, "Model for the electrical behavior of a microplasma", *Journal of Applied Physics*, vol. 35, no. 5, pp. 1370–1376, 1964. DOI: 10.1063/1.1713636. eprint: https://doi.org/10.1063/1.1713636. [Online]. Available: https://doi.org/10.1063/1.1713636.

[206] G. Acconcia, M. Ghioni, and I. Rech, "37ps-precision time-resolving active quenching circuit for high-performance single photon avalanche diodes", *IEEE Photonics Journal*, vol. 10, pp. 1–13, 2018.

[207] R. Mita and G. Palumbo, "High-speed and compact quenching circuit for single-photon avalanche diodes", *IEEE Transactions on Instrumentation and Measurement*, vol. 57, pp. 543–547, 2008.

[208] J. Richardson, R. Henderson, and D. Renshaw, "Dynamic quenching for single photon avalanche diode arrays", 2007.

[209] S. Tisa, F. Guerrieri, and F. Zappa, "Variable-load quenching circuit for single-photon avalanche diodes", *Opt. Express*, vol. 16, no. 3, pp. 2232–2244, 2008. DOI: 10.1364/OE.16.002232. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-16-3-2232.

[210] H. Pahlavanzadeh and M. A. Karami, "Low-power compact tunable quenching configuration for minimizing afterpulsing in single-photon avalanche diodes", *International Journal of Circuit Theory and Applications*, vol. 48, no. 9, pp. 1534–1543, 2020. DOI: https://doi.org/10.1002/cta.2805. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cta.2805. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cta.2805.

[211] S Cova, "Active quenching circuit for avalanche photodiodes.", *U.S. patent 4,963,727, (Italian patent 22367A/88); licensed for industrial production to Silena SpA, Milano, Italy*, Oct. 1990.

[212] F. Zappa, M. Ghioni, S. Cova, C. Samori, and A. Giudice, "An integrated active-quenching circuit for single-photon avalanche diodes", *IEEE Trans. Instrum. Meas.*, vol. 49, pp. 1167–1175, 2000.

[213] F. Zappa, A. Giudice, M Ghioni, and S. Cova, "Fully-integrated active-quenching circuit for single-photon detection", in *Proceedings of the 28th European solid-state circuits conference*, IEEE, 2002, pp. 355–358.

[214] F. Zappa, A. Lotito, A. Giudice, S. Cova, and M. Ghioni, "Monolithic active-quenching and active-reset circuit for single-photon avalanche detectors", *IEEE Journal of Solid-state Circuits*, vol. 38, pp. 1298–1301, 2003.

[215] I. Prochazka, K. Hamal, and L. Kral, "Single photon counting module for space applications", *Journal of Modern Optics*, vol. 54, no. 2-3, pp. 151–161, 2007. DOI: 10.1080/09500340600791756. eprint: https://doi.org/10.1080/09500340600791756. [Online]. Available: https://doi.org/10.1080/09500340600791756.

[216] M. Stipčević, "Active quenching circuit for single-photon detection with Geiger mode avalanche photodiodes", *Appl. Opt.*, vol. 48, no. 9, pp. 1705–1714, 2009. DOI: 10.1364/AO.48.001705. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-48-9-1705.

[217] F. Acerbi, A. D. Frera, A. Tosi, and F. Zappa, "Fast active quenching circuit for reducing avalanche charge and afterpulsing in InGaAs/InP single-photon avalanche diode", *IEEE Journal of Quantum Electronics*, vol. 49, no. 7, pp. 563–569, 2013. DOI: 10.1109/JQE.2013.2260726.

[218] D. Bronzi, S. Tisa, F. Villa, S. Bellisai, A. Tosi, and F. Zappa, "Fast Sensing and Quenching of CMOS SPADs for Minimal Afterpulsing Effects", *IEEE Photonics Technology Letters*, vol. 25, no. 8, pp. 776–779, 2013. DOI: 10.1109/LPT.2013.2251621.

[219] M. A. Wayne, A. Restelli, J. C. Bienfang, and P. G. Kwiat, "Afterpulse reduction through prompt quenching in silicon reach-through single-photon avalanche diodes", *Journal of Lightwave Technology*, vol. 32, no. 21, pp. 4097–4103, 2014. DOI: 10.1109/JLT.2014.2346736.

[220] I. Vornicu, R. Carmona-Galán, and. Rodríguez-Vázquez, "A SPAD-based 3D imager with in-pixel TDC for 145ps-accuracy ToF measurement", in *Image Sensors and Imaging Systems 2015*, R. Widenhorn and A. Dupret, Eds., International Society for Optics and Photonics, vol. 9403, SPIE, 2015, pp. 106 –111. [Online]. Available: https://doi.org/10.1117/12.2078777.

[221] M. Stipčević, B. G. Christensen, P. G. Kwiat, and D. J. Gauthier, "Advanced active quenching circuit for ultra-fast quantum cryptography", *Opt. Express*, vol. 25, no. 18, pp. 21 861–21 876, 2017. DOI: 10.1364/OE.25.021861. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-25-18-21861.

[222] L. Zheng, H. Hu, Z. Weng, Q. Yao, J. Wu, and W. Sun, "Compact active quenching circuit for single photon avalanche diodes arrays", *J. Circuits Syst. Comput.*, vol. 26, 1750149:1–1750149:14, 2017.

[223] R. Enne, B. Steindl, M. Hofbauer, and H. Zimmermann, "Fast cascoded quenching circuit for decreasing afterpulsing effects in 0.35 μm CMOS", *IEEE Solid-State Circuits Letters*, vol. 1, no. 3, pp. 62–65, 2018. DOI: 10.1109/LSSC.2018.2827881.

[224] F. Nolet, S. Parent, N. Roy, *et al.*, "Quenching circuit and SPAD integrated in CMOS 65 nm with 7.8 ps FWHM single photon timing resolution", *Instruments*, vol. 2, p. 19, Sep. 2018. DOI: 10.3390/instruments2040019.

[225] Y. Yu, C. Wang, H. Shi, B. Yang, and X. Liu, "A review of quenching circuit design based on Geiger-Mode APD", *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 28–33, 2018.

[226] G. Acconcia, J. Smith, R. Hare, *et al.*, "Fully integrated high-speed electronics for remote sensing with a large array of single photon avalanche diodes", in *Remote Sensing*, 2019.

[227] J. Rhim, X. Zeng, Z. Huang, *et al.*, "Monolithically-integrated single-photon avalanche diode in a zero-change standard CMOS process for low-cost and low-voltage LiDAR application", *Instruments*, vol. 3, no. 2,33, 2019. DOI: 10.3390/instruments3020033.

[228] M. D. Lakeh, J.-B. Kammerer, W. Uhring, J.-B. Schell, and F. Calmon, "An ultrafast active quenching circuit for SPAD in CMOS 28nm FDSOI technology", in *2020 IEEE SENSORS*, 2020, pp. 1–4. DOI: 10.1109/SENSORS47125.2020.9278902.

[229] Y.-Q. Fang, K. Luo, X. Gao, *et al.*, "High detection efficiency silicon single-photon detector with a monolithic integrated circuit of active quenching and active reset.", *The Review of scientific instruments*, vol. 91 12, p. 123 106, 2020.

[230] Y. Xu, J. Lu, and Z. Wu, "A compact high-speed active quenching and recharging circuit for SPAD detectors", *IEEE Photonics Journal*, vol. 12, pp. 1–8, 2020.

[231] M. Liu, C. Hu, J. C. Campbell, Z. Pan, and M. M. Tashima, "Reduce afterpulsing of single photon avalanche diodes using passive quenching with active reset", *IEEE Journal of Quantum Electronics*, vol. 44, no. 5, pp. 430–434, 2008. DOI: 10.1109/JQE.2007.916688.

[232] 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*, pp. 1–1, 2021. DOI: 10.1109/JSTQE.2021.3088216.

[233] C. Hu, M. Liu, X. Zheng, and J. C. Campbell, "Dynamic range of passive quenching active reset circuit for single photon avalanche diodes", *IEEE Journal of Quantum Electronics*, vol. 46, no. 1, pp. 35–39, 2010. DOI: 10.1109/JQE.2009.2024086.

[234] M. Gronholm, J. Poikonen, and M. Laiho, "A ring-oscillator-based active quenching and active recharge circuit for single photon avalanche diodes", in *2009 European Conference on Circuit Theory and Design*, 2009, pp. 5–8. DOI: 10.1109/ECCTD.2009.5275143.

[235] 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, no. 12, pp. 2977–2989, 2008. DOI: 10.1109/JSSC.2008.2006445.

[236] L. Pancheri and D. Stoppa, "A SPAD-based pixel linear array for high-speed time-gated fluorescence lifetime imaging", in *2009 Proceedings of ESSCIRC*, 2009, pp. 428–431. DOI: 10.1109/ESSCIRC.2009.5325948.

[237] R. G. W. Brown, R. Jones, J. G. Rarity, and K. D. Ridley, "Characterization of silicon avalanche photodiodes for photon correlation measurements. 2: active quenching", *Appl. Opt.*, vol. 26, no. 12, pp. 2383–2389, 1987. DOI: 10.1364/AO.26.002383. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-26-12-2383.

[238] R. G. W. Brown and M. Daniels, "Characterization of silicon avalanche photodiodes for photon correlation measurements. 3: Sub-Geiger operation", *Appl. Opt.*, vol. 28, no. 21, pp. 4616–4621, 1989. DOI: 10.1364/AO.28.004616. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-28-21-4616.

[239] A. Gallivanoni, I. Rech, D. Resnati, M. Ghioni, and S. Cova, "Monolithic active quenching and picosecond timing circuit suitable for large-area single-photon avalanche diodes", *Opt. Express*, vol. 14, no. 12, pp. 5021–5030, 2006. DOI: 10.1364/OE.14.005021. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-14-12-5021.

[240] A. Eisele, R. Henderson, B. Schmidtke, *et al.*, "185 MHz Count Rate, 139 dB dynamic range single-photon avalanche diode with active quenching circuit in 130 nm CMOS technology", in *IISW 2011, Hokkaido, Japan*, Jun. 2011.

[241] S. Isaak, S. Bull, M. C. Pitter, and I. Harrison, "Fully integrated linear single photon avalanche diode (SPAD) array with parallel readout circuit in a standard 180 nm CMOS process", in *2010 International Conference on Enabling Science and Nanotechnology (ESciNano)*, 2010, pp. 1–2. DOI: 10.1109/ESCINANO.2010.5700935.

[242] F. Villa, R. Lussana, D. Tamborini, A. Tosi, and F. Zappa, "High-fill-factor 60×1 SPAD array with 60 subnanosecond integrated TDCs", *IEEE Photonics Technology Letters*, vol. 27, no. 12, pp. 1261–1264, 2015.

[243] S. Burri, C. Bruschini, and E. Charbon, "LinoSPAD: a compact linear SPAD camera system with 64 FPGA-Based TDC modules for versatile 50 ps resolution time-resolved imaging", *Instruments*, vol. 1, no. 1, 2017, ISSN: 2410-390X. DOI: 10.3390/instruments1010006. [Online]. Available: https://www.mdpi.com/2410-390X/1/1/6.

[244] A. Incoronato, F. Severini, F. Madonini, F. Villa, and F. Zappa, "Linear SPAD array for quantum communication", in *Quantum Optics and Photon Counting 2021*, International Society for Optics and Photonics, vol. 11771, SPIE, 2021, pp. 9 –13. [Online]. Available: https://doi.org/10.1117/12.2592091.

[245] R. M. Field, S. Realov, and K. L. Shepard, "A 100 fps, time-correlated single-photon-counting-based fluorescence-lifetime imager in 130 nm CMOS", *IEEE Journal of Solid-State Circuits*, vol. 49, no. 4, pp. 867–880, 2014. DOI: 10.1109/JSSC.2013.2293777.

[246] P. Buzhan, B. Dolgoshein, L. Filatov, *et al.*, "Silicon photomultiplier and its possible applications", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 504, no. 1, pp. 48–52, 2003, Proceedings of the 3rd International Conference on New Developments in Photodetection, ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)00749-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900203007496.

[247] V. Golovin and V. Saveliev, "Novel type of avalanche photodetector with Geiger mode operation", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 518, no. 1, pp. 560–564, 2004, Frontier Detectors for Frontier Physics: Proceedin, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2003.11.085. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900203029528.

[248] D. Herbert, V. Saveliev, N. Belcari, N. D'Ascenzo, A. Del Guerra, and A. Golovin, "First results of scintillator readout with silicon photomultiplier", *IEEE Transactions on Nuclear Science*, vol. 53, no. 1, pp. 389–394, 2006. DOI: 10.1109/TNS.2006.869848.

[249] D Renker and E Lorenz, "Advances in solid state photon detectors", *Journal of Instrumentation*, vol. 4, no. 04, P04004–P04004, 2009. DOI: 10.1088/1748-0221/4/04/p04004. [Online]. Available: https://doi.org/10.1088/1748-0221/4/04/p04004.

[250] N. D'Ascenzo and Q. Xie, "Silicon photomultiplier for the plug & imaging PET system: physics, technological challenges and application to modern nuclear medicine", in *Photon Counting*, N. Britun and A. Nikiforov, Eds., Rijeka: IntechOpen, 2018, ch. 8. DOI: 10.5772/intechopen.73007. [Online]. Available: https://doi.org/10.5772/intechopen.73007.

[251] A Gasanov, V Golovin, Z. Y. Sadygov, and N. Y. Yusipov, "Avalanche photodetector based on metal-resistive-layer semiconductor structures", *Pisma V Zhurnal Tekh. Fiz*, vol. 14, pp. 706–709, 1988.

[252] D. A. Shushakov and V. E. Shubin, "New solid state photomultiplier", in *Optoelectronic Integrated Circuit Materials, Physics, and Devices*, M. Razeghi, Y.-S. Park, and G. L. Witt, Eds., International Society for Optics and Photonics, vol. 2397, SPIE, 1995, pp. 544 –554. [Online]. Available: https://doi.org/10.1117/12.206900.

[253] A. Akindinov, A. Martemianov, P. Polozov, V. Golovin, and E. Grigoriev, "New results on MRS APDs", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 387, no. 1, pp. 231–234, 1997, New Developments in Photodetection, ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(96)01201-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900296012016.

[254] Z Sadygov, *Patent no*, RU Patent 2,102,820, 1998.

[255] V Saveliev and V Golovin, "Silicon avalanche photodiodes on the base of metal-resistor-semiconductor (MRS) structures", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 442, no. 1, pp. 223–229, 2000, ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(99)01225-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900299012255.

[256] C. Piemonte, R. Battiston, M. Boscardin, *et al.*, "Characterization of the first prototypes of silicon photomultiplier fabricated at ITC-irst", *IEEE Transactions on Nuclear Science*, vol. 54, no. 1, pp. 236–244, 2007. DOI: 10.1109/TNS.2006.887115.

[257] A. G. Stewart, V. Saveliev, S. J. Bellis, D. J. Herbert, P. J. Hughes, and J. C. Jackson, "Performance of 1mm^2 silicon photomultiplier", *IEEE Journal of Quantum Electronics*, vol. 44, no. 2, pp. 157–164, 2008. DOI: 10.1109/JQE.2007.910940.

[258] A. Gola, F. Acerbi, M. Capasso, *et al.*, "NUV-sensitive silicon photomultiplier technologies developed at Fondazione Bruno Kessler", *Sensors*, vol. 19, no. 2, 2019, ISSN: 1424-8220. DOI: 10.3390/s19020308. [Online]. Available: https://www.mdpi.com/1424-8220/19/2/308.

[259] A. A. Wagadarikar, S. Katz, A. Ivan, and S. Dolinsky, "Performance of low afterpulsing probability multi-pixel photon counters for time-of-flight positron emission tomography", in *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC)*, 2013, pp. 1–5. DOI: 10.1109/NSSMIC.2013.6829595.

[260] F. Acerbi, A. Gola, V. Regazzoni, *et al.*, "High efficiency, ultra-high-density silicon photomultipliers", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 2, pp. 1–8, 2018. DOI: 10.1109/JSTQE.2017.2748927.

[261] F. Villa, Y. Zou, A. Dalla Mora, A. Tosi, and F. Zappa, "SPICE Electrical Models and Simulations of Silicon Photomultipliers", *IEEE Transactions on Nuclear Science*, vol. 62, no. 5, pp. 1950–1960, 2015. DOI: 10.1109/TNS.2015.2477716.

[262] F. Corsi, C. Marzocca, A. Perrotta, *et al.*, "Electrical characterization of silicon photomultiplier detectors for optimal front-end design", in *2006 IEEE Nuclear Science Symposium Conference Record*, vol. 2, 2006, pp. 1276–1280. DOI: 10.1109/NSSMIC.2006.356076.

[263] S. Seifert, H. T. van Dam, J. Huizenga, *et al.*, "Simulation of silicon photomultiplier signals", *IEEE Transactions on Nuclear Science*, vol. 56, no. 6, pp. 3726–3733, 2009. DOI: 10.1109/TNS.2009.2030728.

[264] D. Marano, M. Belluso, G. Bonanno, *et al.*, "Silicon photomultipliers electrical model extensive analytical analysis", *IEEE Transactions on Nuclear Science*, vol. 61, no. 1, pp. 23–34, 2014. DOI: 10.1109/TNS.2013.2283231.

[265] G Collazuol, "Review of silicon photo-multiplier physics and applications, including a study at low temperature", *Innovative Particle and Radiation*, 2008.

[266] L. Huf Campos Braga, "Modeling, design and characterization of a multi-pixel digital SiPM for PET applications", Ph.D. dissertation, University of Trento, 2014.

[267] S. Gundacker, E. Auffray, P. Jarron, T. Meyer, and P. Lecoq, "On the comparison of analog and digital SiPM readout in terms of expected timing performance", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 787, pp. 6–11, 2015, New Developments in Photodetection NDIP14, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2014.10.020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016890021401167X.

[268] S. Mandai and E. Charbon, "Multi-channel digital SiPMs: concept, analysis and implementation", in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, 2012, pp. 1840–1844. DOI: 10.1109/NSSMIC.2012.6551429.

[269] A. Carimatto, S. Mandai, E. Venialgo, *et al.*, "11.4 A 67,392-SPAD PVTB-compensated multi-channel digital SiPM with 432 column-parallel 48ps 17b TDCs for endoscopic time-of-flight PET", in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015, pp. 1–3. DOI: 10.1109/ISSCC.2015.7062996.

[270] S. Gnecchi, N. A. W. Dutton, L. Parmesan, *et al.*, "Digital silicon photomultipliers with or/xor pulse combining techniques", *IEEE Transactions on Electron Devices*, vol. 63, no. 3, pp. 1105–1110, 2016. DOI: 10.1109/TED.2016.2518301.

[271] C. Zhang, S. Lindner, I. M. Antolović, J. Mata Pavia, M. Wolf, and E. Charbon, "A 30-frames/s, 252 × 144 SPAD Flash LiDAR with 1728 dual-clock 48.8-ps TDCs, and pixel-wise integrated histogramming", *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 1137–1151, 2019. DOI: 10.1109/JSSC.2018.2883720.

[272] D. Gascón, *Integrated signal processing for a new generation of active hybrid single photon sensors with ps time resolution (fasticpix)*, 2019.

[273] N. Egidos, R. Ballabriga, F. Bandi, *et al.*, "20-ps resolution clock distribution network for a fast-timing single-photon detector", *IEEE Transactions on Nuclear Science*, vol. 68, no. 4, pp. 434–446, 2021. DOI: 10.1109/TNS.2021.3057581.

[274] J. Jourdon, S. Lhostis, S. Moreau, *et al.*, "Hybrid bonding for 3D stacked image sensors: impact of pitch shrinkage on interconnect robustness", in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 7.3.1–7.3.4. DOI: 10.1109/IEDM.2018.8614570.

[275] S. N. Farrens, "(Invited) Hybrid Wafer Bonding – the fusion of low temperature dielectric and metal bonding technologies", *ECS Transactions*, vol. 98, no. 4, pp. 145–155, 2020. DOI: 10.1149/09804.0145ecst. [Online]. Available: https://doi.org/10.1149/09804.0145ecst.

[276] Z. Wang, "Microsystems using three-dimensional integration and TSV technologies: fundamentals and applications", *Microelectronic Engineering*, vol. 210, pp. 35–64, 2019, ISSN: 0167-9317. DOI: https://doi.org/10.1016/j.mee.2019.03.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167931719300553.

[277] B. F. Aull, D. R. Schuette, D. J. Young, D. M. Craig, B. J. Felton, and K. Warner, "A study of crosstalk in a 256 × 256 photon counting imager based on silicon Geiger-Mode avalanche photodiodes", *IEEE Sensors Journal*, vol. 15, no. 4, pp. 2123–2132, 2015. DOI: 10.1109/JSEN.2014.2368456.

[278] F. Gramuglia, E. Ripiccini, C. A. Fenoglio, *et al.*, *Sub-10 ps minimum ionizing particle detection with Geiger-Mode APDs*, 2021. arXiv: 2111.09998 [physics.ins-det].

[279] F. Gramuglia, P. Keshavarzian, E. Kizilkan, *et al.*, "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 a publication of the IEEE Lasers and Electro-optics Society*, vol. 28, no. 2, 2022, ISSN: 1077-260X.

[280] C. Veerappan and E. Charbon, "A low dark count p-i-n diode based SPAD in CMOS technology", *IEEE Transactions on Electron Devices*, vol. 63, pp. 65–71, 2016.

[281] I. M. Antolovic, C. Bruschini, and E. Charbon, "Dynamic range extension for photon counting arrays", *Optics Express*, vol. 26, no. 17, pp. 22 234–22 248, 2018. DOI: 10.1364/OE.26.022234.

[282] P. Mroszczyk and P. Dudek, "Tunable CMOS delay gate with improved matching properties", *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 9, pp. 2586–2595, 2014. DOI: 10.1109/TCSI.2014.2312491.

[283] E. A. G. Webster, R. L. Nicol, L. Grant, and D. Renshaw, "Per-Pixel Dark Current Spectroscopy Measurement and Analysis in CMOS Image Sensors", *IEEE Transactions on Electron Devices*, vol. 57, no. 9, pp. 2176–2182, 2010. DOI: 10.1109/TED.2010.2052399.

[284] E. A. G. Webster and R. K. Henderson, "A TCAD and spectroscopy study of dark count mechanisms in single-photon avalanche diodes", *IEEE Transactions on Electron Devices*, vol. 60, no. 12, pp. 4014–4019, 2013. DOI: 10.1109/TED.2013.2285163.

[285] N. Loukianova, H. Folkerts, J. Maas, *et al.*, "Leakage current modeling of test structures for characterization of dark current in CMOS image sensors", *IEEE Transactions on Electron Devices*, vol. 50, no. 1, pp. 77–83, 2003. DOI: 10.1109/TED.2002.807249.

[286] F. Villa, D. Bronzi, Y. Zou, *et al.*, "CMOS SPADs with up to 500 μm diameter and 55% detection efficiency at 420 nm", *Journal of Modern Optics*, vol. 61, no. 2, pp. 102–115, 2014.

[287] N Cartiglia, M Baselga, G Dellacasa, *et al.*, "Performance of ultra-fast silicon detectors", *Journal of instrumentation*, vol. 9, no. 02, p. C02001, 2014.

[288] M. Assanelli, A. Ingargiola, I. Rech, A. Gulinatti, and M. Ghioni, "Photon-timing jitter dependence on injection position in single-photon avalanche diodes", *IEEE Journal of Quantum Electronics*, vol. 47, no. 2, pp. 151–159, 2011. DOI: 10.1109/JQE.2010.2068038.

[289] A. Schenk, "An improved approach to the Shockley–Read–Hall recombination in inhomogeneous fields of space-charge regions", *Journ. Appl. Phys.*, vol. 71, no. 7, pp. 3339–3349, 1992.

[290] G. Zappalà, F. Acerbi, A. Ferri, *et al.*, "Set-up and methods for SiPM photo-detection efficiency measurements", *Journal of Instrumentation*, vol. 11, no. 08, P08014, 2016.

[291] 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*, IEEE, 2008, pp. 270–273.

[292] J. A. Richardson, L. A. Grant, and R. K. Henderson, "Low dark count single-photon avalanche diode structure compatible with standard nanometer scale CMOS technology", *IEEE Photonics Technology Letters*, vol. 21, no. 14, pp. 1020–1022, 2009. DOI: 10.1109/LPT.2009.2022059.

[293] E. A. G. Webster, J. A. Richardson, L. A. Grant, D. Renshaw, and R. K. Henderson, "A single-photon avalanche diode in 90-nm CMOS imaging technology with 44% photon detection efficiency at 690 nm", *IEEE Electron Device Letters*, vol. 33, no. 5, pp. 694–696, 2012. DOI: 10.1109/LED.2012.2187420.

[294] E. A. G. Webster, L. A. Grant, and R. K. Henderson, "A high-performance single-photon avalanche diode in 130-nm cmos imaging technology", *IEEE Electron Device Letters*, vol. 33, no. 11, pp. 1589–1591, 2012. DOI: 10.1109/LED.2012.2214760.

[295] T. Leitner, A. Feiningstein, R. Turchetta, *et al.*, "Measurements and simulations of low dark count rate single photon avalanche diode device in a low voltage 180-nm CMOS image sensor technology", *IEEE Transactions on Electron Devices*, vol. 60, no. 6, pp. 1982–1988, 2013. DOI: 10.1109/TED.2013.2259172.

[296] E. Charbon, H. Yoon, and Y. Maruyama, "A Geiger mode APD fabricated in standard 65nm CMOS technology", in *2013 IEEE International Electron Devices Meeting*, 2013, pp. 27.5.1–27.5.4. DOI: 10.1109/IEDM.2013.6724705.

[297] M.-J. Lee, P. Sun, and E. Charbon, "A first single-photon avalanche diode fabricated in standard SOI CMOS technology with a full characterization of the device", *Optics Express*, vol. 23, no. 10, pp. 13 200–13 209, 2015.

[298] 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, pp. 2445–2448, 2015.

[299] I. Takai, H. Matsubara, M. Soga, M. Ohta, M. Ogawa, and T. Yamashita, "Single-photon avalanche diode with enhanced NIR-sensitivity for automotive LIDAR systems", *Sensors*, vol. 16, no. 4, Mar. 2016.

[300] S. Pellegrini and B. Rae, "Fully industrialised single photon avalanche diodes", in *Proc. Adv. Photon Count. Techniq. XI*, vol. 10212, 2017, pp. 16.5.1–16.5.4.

[301] 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", *Optics Express*, vol. 25, no. 11, pp. 12 765–12 778, 2017. DOI: 10.1364/OE.25.012765.

[302] 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)*, IEEE, 2018, pp. 94–97.

[303] I. Vornicu, F. Bandi, R. Carmona-Galán, and Á. Rodríguez-Vázquez, "Low-noise and high-efficiency near-IR SPADs in 110nm CIS technology", in *ESSDERC 2019-49th European Solid-State Device Research Conference (ESSDERC)*, IEEE, 2019, pp. 250–253.

[304] I. Vornicu, J. M. López-Martínez, F. N. Bandi, R. C. Galán, and Á. Rodríguez-Vázquez, "Design of high-efficiency SPADs for LiDAR applications in 110nm CIS technology", *IEEE Sensors Journal*, vol. 21, no. 4, pp. 4776–4785, 2020.

[305] R. F. Post and L. I. Schiff, "Statistical limitations on the resolving time of a scintillation counter", *Phys. Rev.*, vol. 80, pp. 1113–1113, 6 1950. DOI: 10.1103/PhysRev.80.1113. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.80.1113.

[306] E. Gatti and V. Svelto, "Review of theories and experiments of resolving time with scintillation counters", *Nuclear Instruments and Methods*, vol. 43, pp. 248–268, 1966.

[307] G. Ranucci, "Time statistics of the photoelectron emission process in scintillation counters", *Nuclear Instruments & Methods in Physics Research Section A-accelerators Spectrometers Detectors and Associated Equipment*, vol. 335, pp. 121–128, 1993.

[308] M. W. Fishburn and E. Charbon, "System tradeoffs in gamma-ray detection utilizing spad arrays and scintillators", *IEEE Transactions on Nuclear Science*, vol. 57, no. 5, pp. 2549–2557, 2010. DOI: 10.1109/TNS.2010.2064788.

[309] S. Seifert, H. T. van Dam, and D. R. Schaart, "The lower bound on the timing resolution of scintillation detectors", *Physics in Medicine & Biology*, vol. 57, no. 7, p. 1797, 2012.

[310] E. Venialgo, S. Mandai, and E. Charbon, "Time mark estimators for md-sipm and impact of system parameters", in *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC)*, 2013, pp. 1–2. DOI: 10.1109/NSSMIC.2013.6829083.

[311] E. Venialgo, S. Mandai, T. Gong, D. Schaart, and E. Charbon, "Practical time mark estimators for multichannel digital silicon photomultipliers", in *2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2015, pp. 1–3. DOI: 10.1109/NSSMIC.2015.7582138.

[312] J. S. Karp, S. Surti, M. E. Daube-Witherspoon, and G. Muehllehner, "Benefit of time-of-flight in pet: experimental and clinical results", *Journal of Nuclear Medicine*, vol. 49, no. 3, pp. 462–470, 2008.

[313] F. Gramuglia, A. A. Muntean, E. Venialgo Araujo, *et al.*, "CMOS 3D-Stacked FSI multi-channel digital SiPM for time-of-flight PET applications", 2020. DOI: 10.1109/NSS/MIC42677.2020.9507833. [Online]. Available: http://infoscience.epfl.ch/record/285321.

[314] F. Gramuglia, A. A. Muntean, C. A. Fenoglio, *et al.*, "CMOS 3D-Stacked FSI multi-channel digital SiPM for time-of-flight vision applications", 4. 69–72, R19, 2021. [Online]. Available: http://infoscience.epfl.ch/record/288691.

[315] ——, "Architecture and characterization of a CMOS 3D-Stacked FSI multi-channel digital SiPM for time-of-flight PET applications", p. 2, 2021. [Online]. Available: http://infoscience.epfl.ch/record/288698.

[316] I. Jani, "Test and characterization of 3d high-density interconnects", Ph.D. dissertation, Nov. 2019.

[317] Y. Xie and Q. Zou, "3D integration technology", in *More than Moore Technologies for Next Generation Computer Design*, R. O. Topaloglu, Ed. New York, NY: Springer New York, 2015, pp. 23–48, ISBN: 978-1-4939-2163-8. DOI: 10.1007/978-1-4939-2163-8_2. [Online]. Available: https://doi.org/10.1007/978-1-4939-2163-8_2.

[318] M. Z. Straayer and M. H. Perrott, "A multi-path gated ring oscillator TDC with first-order noise shaping", *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1089–1098, 2009. DOI: 10.1109/JSSC.2009.2014709.

[319] A. Muntean, E. Venialgo, S. Gnecchi, C. Jackson, and E. Charbon, "Towards a fully digital state-of-the-art analog SiPM", in *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2017, pp. 1–4. DOI: 10.1109/NSSMIC.2017.8533036.

[320] A. Muntean, E. Venialgo, A. Ardelean, *et al.*, "Blumino: the first fully integrated analog SiPM with on-chip time conversion", *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 5, pp. 671–678, 2021. DOI: 10.1109/TRPMS.2020.3045081.

[321] S. E. Brunner and D. R. Schaart, "BGO as a hybrid scintillator / Cherenkov radiator for cost-effective time-of-flight PET", *Physics in Medicine and Biology*, vol. 62, no. 11, pp. 4421–4439, 2017. DOI: 10.1088/1361-6560/aa6a49. [Online]. Available: https://doi.org/10.1088\%2F1361-6560\%2Faa6a49.

[322] G. Lerondel, L. Le Cunff, R. S. Montiel, P. Lecoq, D. Turover, and D. Grosso, *Method for optimizing the collection of photons in scintillator crystals, and related crystal and uses. patent application number: WIPO (PCT) WO2015136165A1*, 2015.

[323] M. Janecek, "Reflectivity spectra for commonly used reflectors", *IEEE Transactions on Nuclear Science*, vol. 59, no. 3, pp. 490–497, 2012. DOI: 10.1109/TNS.2012.2183385.

[324] S. Ghosh, J. Haefner, J. Martín-Albo, *et al.*, "Dependence of polytetrafluoroethylene reflectance on thickness at visible and ultraviolet wavelengths in air", *JINST*, Jul. 2020.

[325] D. N. ter Weele, D. R. Schaart, and P. Dorenbos, "Picosecond time resolved studies of photon transport inside scintillators", *IEEE Transactions on Nuclear Science*, vol. 62, no. 5, pp. 1961–1971, 2015. DOI: 10.1109/TNS.2015.2460267.

[326] C. Kuntner and D. Stout, "Quantitative preclinical PET imaging: opportunities and challenges", *Frontiers in Physics*, vol. 2, Nov. 2013. DOI: 10.3389/fphy.2014.00012.

[327] 3M, *Enhanced Specular Reflector (ESR)*, online, accessed 17/10/2019, 2019. [Online]. Available: http://multimedia.3m.com/mws/media/374730O/vikuiti-tm-esr-sales-literature.pdf?fn=ESR\%20ss2.pdf.

[328] F. Loignon-Houle, C. M. Pepin, S. A. Charlebois, and R. Lecomte, "Reflectivity quenching of ESR multilayer polymer film reflector in optically bonded scintillator arrays", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 851, pp. 62 –67, 2017, ISSN: 0168-9002. DOI: 10.1016/j.nima.2017.01.051. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900217301146.

[329] A. R. Knapitsch, "Photonic Crystals: Enhancing the Light Output of Scintillation Based Detectors", Ph.D. dissertation, Vienna, Tech. U, 2012-10-21.

[330] A. Knapitsch and P. Lecoq, "Review on photonic crystal coatings for scintillators", *International Journal of Modern Physics A*, vol. 29, no. 30, p. 1 430 070, 2014. DOI: 10.1142/S0217751X14300701.

[331] M. Salomoni, R. Pots, E. Auffray, and P. Lecoq, "Enhancing light extraction of inorganic scintillators using photonic crystals", *Crystals*, vol. 8, no. 2, 2018, ISSN: 2073-4352. DOI: 10.3390/cryst8020078. [Online]. Available: https://www.mdpi.com/2073-4352/8/2/78.

[332] C. Chen, Z. Zhu, B. Liu, *et al.*, "Effect of a conformal layer on the photonic crystal for light extraction of scintillator", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 950, p. 162 953, 2020, ISSN: 0168-9002. DOI: 10.1016/j.nima.2019.162953. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900219313464.

[333] F. Liu, Y. Yang, Y. Liu, *et al.*, "Large energy resolution improvement of LYSO scintillator by electron beam lithography method", *AIP Advances*, vol. 10, 2020. DOI: 10.1063/1.5136077.

[334] R. Pots, M. Salomoni, S. Gundacker, *et al.*, "Improving light output and coincidence time resolution of scintillating crystals using nanoimprinted photonic crystal slabs", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 940, Jun. 2019. DOI: 10.1016/j.nima.2019.06.026.

[335] Z. Zhu, B. Liu, H. Zhang, *et al.*, "Improvement of light extraction of LYSO scintillator by using a combination of self-assembly of nanospheres and atomic layer deposition", *Opt. Express*, vol. 23, no. 6, pp. 7085–7093, 2015. DOI: 10.1364/OE.23.007085. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-23-6-7085.

[336] B. Singh, M. S. J. Marshall, S. Waterman, *et al.*, "Enhanced scintillation light extraction using nanoimprinted photonic crystals", *IEEE Transactions on Nuclear Science*, vol. 65, no. 4, pp. 1059–1065, 2018. DOI: 10.1109/TNS.2018.2811646.

[337] X. Ouyang, B. Liu, X. Xiang, *et al.*, "Enhanced light output of CsI(Na) scintillators by photonic crystals", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 969, p. 164 007, 2020, ISSN: 0168-9002. DOI: 10.1016/j.nima.2020.164007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900220304599.

[338] X. Ouyang, B. Liu, X. Xiang, *et al.*, "Enhancement of the energy resolution of CsI(Na) scintillators by photonic crystals prepared with dry-transfer technique", *Opt. Express*, vol. 28, no. 22, pp. 33 077–33 083, 2020. DOI: 10.1364/OE.404815. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-28-22-33077.

[339] Y. Liu, F. Liu, W. Tang, *et al.*, "Light extraction enhancement of BGO scintillator by monolayers of SiO2 periodic array", *AIP Advances*, vol. 9, 2019. DOI: 10.1063/1.5124361.

[340] F. Gramuglia, M. Lee, E. Venialgo, C. Bruschini, and E. Charbon, "Towards 10ps 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)*, 2017, pp. 1–3. DOI: 10.1109/NSSMIC.2017.8532951.

[341] Q. Sun, Q. Peng, Z. Wu, Q. Huang, and J. Xu, "Ultra-thin high-reflector film designed for LYSO scintillators", in *2016 IEEE Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop (NSS/MIC/RTSD)*, 2016, pp. 1–2. DOI: 10.1109/NSSMIC.2016.8069726.

[342] Q. Sun, Q. Peng, Z. Wu, S. Xie, Q. Huang, and J. Xu, "Design of ultra-thin anti-reflection films directly coated on lyso scintillators", in *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2017, pp. 1–3. DOI: 10.1109/NSSMIC.2017.8532600.

[343] F. Gramuglia, N. Descharmes, E. Venialgo, H. P. Herzig, E. Charbon, and C. Bruschini, "Light extraction enhancement in scintillation crystals using thin film coatings", in *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, 2018, pp. 1–2. DOI: 10.1109/NSSMIC.2018.8824270.

[344] M. Aliofkhazraei, *Advances in Nanostructured Composites: Volume 1: Carbon Nanotube and Graphene Composites*. CRC Press, 2019, ISBN: 9781482236644. [Online]. Available: https://books.google.ch/books?id=i02WDwAAQBAJ.

[345] Y. Fink, J. N. Winn, S. Fan, *et al.*, "A dielectric omnidirectional reflector", *Science*, vol. 282, no. 5394, pp. 1679–1682, 1998, ISSN: 0036-8075. DOI: 10.1126/science.282.5394.1679.

[346] S Zanettini, V Gâté, E Usureau, *et al.*, "Improved light extraction efficiency on 2 inches LYSO with nanopatterned TiO2 photonic crystals", *IEEE Strasbourg*, 2016.

[347] H. Kadiri, S. Kostcheev, D. Turover, *et al.*, "Topology assisted self-organization of colloidal nanoparticles: application to 2D large-scale nanomastering", *Beilstein journal of nanotechnology*, vol. 5, no. 1, pp. 1203–1209, 2014.

[348] E. Zanchetta, V. Auzelyte, J. Brugger, A. Savegnago, G. Della Giustina, and G. Brusatin, "Highly inorganic titania based sol–gel as directly patternable resist for micro- and nano- structured surfaces", *Microelectronic Engineering*, vol. 98, pp. 176–179, 2012, Special issue MNE 2011 - Part II, ISSN: 0167-9317. DOI: 10.1016/j.mee.2012.07.043.

[349] R. Ganesan, J. Dumond, M. S. M. Saifullah, S. H. Lim, H. Hussain, and H. Y. Low, "Direct patterning of TiO_2 using step-and-flash imprint lithography", *ACS Nano*, vol. 6, no. 2, pp. 1494–1502, 2012, PMID: 22229254. DOI: 10.1021/nn204405k.

[350] P. Ma, Z. Xu, M. Wang, *et al.*, "Fast fabrication of TiO_2 hard stamps for nanoimprint lithography", *Materials Research Bulletin*, vol. 90, pp. 253–259, 2017, ISSN: 0025-5408. DOI: 10.1016/j.materresbull.2017.03.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0025540817306001.

[351] K. Yoon, K.-Y. Yang, and H. Lee, "Fabrication of polycrystalline TiO_2 nanopatterns by TiO_2 sol base imprint lithography", *Thin Solid Films*, vol. 518, no. 1, pp. 126–129, 2009, ISSN: 0040-6090. DOI: 10.1016/j.tsf.2009.07.056.

[352] K.-m. Yoon, K.-Y. Yang, H. Lee, and H.-S. Kim, "Formation of TiO_2 nanopattern using reverse imprinting and sol-gel method", *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, vol. 27, no. 6, pp. 2810–2813, 2009. DOI: 10.1116/1.3246394. eprint: https://avs.scitation.org/doi/pdf/10.1116/1.3246394.

[353] I. Valais, C. Michail, S. David, C. Nomicos, G. Panayiotakis, and I. Kandarakis, "A comparative study of the luminescence properties of LYSO:Ce, LSO:Ce, GSO:Ce and BGO single crystal scintillators for use in medical X-ray imaging", *Physica Medica*, vol. 24, no. 2, pp. 122–125, 2008, Highlights of the "X EFOMP Congress - First European Conference on Medical Physics", ISSN: 1120-1797. DOI: j.ejmp.2008.01.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1120179708000276.

[354] P. Bienstman, "Rigorous and efficient modelling of wavelenght scale photonic components", Ph.D. dissertation, Ghent University, 2001.

[355] G. Yang, H. Huang, and H. Wang, "High-reflectance of hybrid reflector with distributed Bragg reflector and metal mirror for flip-chip ultra-violet light-emitting diodes", in *International Symposium on Optoelectronic Technology and Application 2014: Laser Materials Processing; and Micro/Nano Technologies*, G. Jin, S. Zhuang, and J. Liu, Eds., International Society for Optics and Photonics, vol. 9295, SPIE, 2014, pp. 54 –59. DOI: 10.1117/12.2073098. [Online]. Available: https://doi.org/10.1117/12.2073098.

[356] P. Lecoq, E. Auffray, and A. Knapitsch, "How photonic crystals can improve the timing resolution of scintillators", *IEEE Transactions on Nuclear Science*, vol. 60, no. 3, pp. 1653–1657, 2013. DOI: 10.1109/TNS.2013.2260768.

[357] A. Iltis, S. Zanettini, L. de Magalhaes, *et al.*, "Impact on timing and light extraction of a photonic crystal as measured on a half patterned LYSO crystal: implications for time of flight PET imaging", *Journal of Instrumentation*, vol. 14, no. 06, P06036–

P06036, 2019. DOI: 10.1088/1748-0221/14/06/p06036. [Online]. Available: https://doi.org/10.1088\%2F1748-0221\%2F14\%2F06\%2Fp06036.

[358] C. S. Castro, L. L. Cunff, A. Vial, and G. Lerondel, "Light management in scintillator crystals: a multi-scale computational study", in *UV and Higher Energy Photonics: From Materials to Applications 2018*, G. Lérondel, S. Kawata, and Y.-H. Cho, Eds., International Society for Optics and Photonics, vol. 10727, SPIE, 2018. DOI: 10.1117/12.2324143. [Online]. Available: https://doi.org/10.1117/12.2324143.

[359] P Valerio, R Cardarelli, G Iacobucci, *et al.*, "A monolithic ASIC demonstrator for the Thin Time-of-Flight PET scanner", *Journal of instrumentation : an IOP and SISSA journal.*, vol. 14, no. 07, 2019-07-16, ISSN: 1748-0221.

[360] L Paolozzi, Y Bandi, R Cardarelli, *et al.*, "Characterization of the demonstrator of the fast silicon monolithic ASIC for the TT-PET project", *Journal of instrumentation : an IOP and SISSA journal.*, vol. 14, no. 02, 2019-02-01, ISSN: 1748-0221.

[361] L Paolozzi, Y Bandi, M Benoit, *et al.*, "Test beam measurement of the first prototype of the fast silicon pixel monolithic detector for the TT-PET project", *Journal of instrumentation : an IOP and SISSA journal.*, vol. 13, no. 04, 2018-04-01, ISSN: 1748-0221.

[362] E. Ripiccini, D. Hayakawa, G. Iacobucci, *et al.*, "Expected performance of the TT-PET scanner", *arXiv preprint arXiv:1811.12381*, 2018.

[363] D. E. Groom and S. Klein, "Passage of particles through matter", *The European Physical Journal C-Particles and Fields*, vol. 15, no. 1, pp. 163–173, 2000.

[364] W. Shockley, "Currents to Conductors Induced by a Moving Point Charge", *Journal of Applied Physics*, vol. 9, no. 10, pp. 635–636, 1938. DOI: 10.1063/1.1710367.

[365] S. Ramo, "Currents Induced by Electron Motion", *Proceedings of the IRE*, vol. 27, no. 9, pp. 584–585, 1939. DOI: 10.1109/JRPROC.1939.228757.

[366] N. Cartiglia, R. Arcidiacono, M. Baselga, *et al.*, "Design optimization of ultra-fast silicon detectors", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 796, pp. 141–148, 2015, Proceedings of the 10th International Conference on Radiation Effects on Semiconductor Materials Detectors and Devices, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2015.04.025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900215004982.

[367] N. Cartiglia, R. Arcidiacono, M. Ferrero, *et al.*, "Timing layers, 4- and 5-dimension tracking", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 924, pp. 350–354, 2019, 11th International Hiroshima Symposium on Development and Application of Semiconductor Tracking Detectors, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2018.09.157. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900218313147.

[368] L. Paolozzi, "Development of particle detectors and related front end electronics for sub-nanosecond time measurement in high radiation environment.", Ph.D. dissertation, Roma Tor Vergata, 2015. [Online]. Available: https://cds.cern.ch/record/2318760.

[369] M Jadhav, W Armstrong, I Cloet, *et al.*, "Picosecond timing resolution measurements of low gain avalanche detectors with a 120 GeV proton beam for the TOPSiDE detector concept", *Journal of instrumentation*, vol. 16, no. 06, P06008, 2021, ISSN: 1748-0221.

[370] A. Lai, L. Anderlini, M. Aresti, *et al.*, "First results of the TIMESPOT project on developments on fast sensors for future vertex detectors", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 981, p. 164 491, 2020, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2020.164491. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900220308883.

[371] L Anderlini, M Aresti, A Bizzeti, *et al.*, "Intrinsic time resolution of 3D-trench silicon pixels for charged particle detection", *Journal of instrumentation*, vol. 15, no. 09, P09029, 2020, ISSN: 1748-0221.

[372] N D'Ascenzo, P. S. Marrocchesi, C. S. Moon, *et al.*, "Silicon avalanche pixel sensor for high precision tracking", *Journal of instrumentation*, vol. 9, no. 03, p. C03027, 2014, ISSN: 1748-0221.

[373] M. M. Vignetti, "Development of a 3D Silicon Coincidence Avalanche Detector (3D-SiCAD) for charged particle tracking", Theses, Université de Lyon, Mar. 2017. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01920836.

[374] M. M. Vignetti, F. Calmon, P. Pittet, *et al.*, "Development of a 3D silicon coincidence avalanche detector for charged particle tracking in medical applications", *2016 IEEE Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop (NSS/MIC/RTSD)*, pp. 1–7, 2016. DOI: 10.1109/NSSMIC.2016.8069553.

[375] V. Saveliev, *Avalanche pixel sensors and related methods*, US Patent 8,269,181, 2012.

[376] L. Ratti, P. Brogi, G. Collazuol, *et al.*, "Layered CMOS SPADs for Low Noise Detection of Charged Particles", *Frontiers in Physics*, vol. 8, p. 625, 2021, ISSN: 2296-424X. DOI: 10.3389/fphy.2020.607319. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphy.2020.607319.

[377] M. Vignetti, F. Calmon, P. Pittet, *et al.*, "3D silicon coincidence avalanche detector (3D-SiCAD) for charged particle detection", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 881, pp. 53–59, 2018, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2017.10.089. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900217311786.

[378] S Terzo, M Benoit, E Cavallaro, *et al.*, "Characterisation of AMS H35 HV-CMOS monolithic active pixel sensor prototypes for HEP applications", *Journal of instrumentation*, vol. 14, no. 02, P02016, 2019, ISSN: 1748-0221.

[379] F. Moscatelli, M. Marisaldi, P. Maccagnani, *et al.*, "Radiation tests of single photon avalanche diode for space applications", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 711, pp. 65–72, 2013, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2013.01.056.

[380] V. Malherbe, S. De Paoli, B. Mamdy, G. Gasiot, and P. Roche, "Displacement Damage Characterization of CMOS Single-Photon Avalanche Diodes: Alpha-Particle and Fast-Neutron Measurements", *IEEE Transactions on Nuclear Science*, vol. 68, no. 5, pp. 777–784, 2021. DOI: 10.1109/TNS.2021.3071171.

[381] M. Campajola, F. D. Capua, D. Fiore, E. Sarnelli, and A. Aloisio, "Proton induced dark count rate degradation in 150-nm CMOS single-photon avalanche diodes", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 947, p. 162 722, 2019, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2019.162722.

[382] C. R. Crowell and S. M. Sze, "Temperature dependence of avalanche multiplication in semiconductors", *Applied Physics Letters*, vol. 9, no. 6, pp. 242–244, 1966. DOI: 10.1063/1.1754731. eprint: https://doi.org/10.1063/1.1754731. [Online]. Available: https://doi.org/10.1063/1.1754731.

[383] 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. DOI: 10.1109/T-ED.1971.17210.

[384] P. Mars, "Temperature dependence of avalanche breakdown voltage temperature dependence of avalanche breakdown voltage in p-n junctions", *International Journal of Electronics*, vol. 32, no. 1, pp. 23–37, 1972. DOI: 10.1080/00207217208938266. eprint: https://doi.org/10.1080/00207217208938266. [Online]. Available: https://doi.org/10.1080/00207217208938266.

[385] K. Kajiyama and H. Kanbe, "Temperature dependence of avalanche breakdown voltage in Si p-n junctions", *Journal of Applied Physics*, vol. 47, no. 6, pp. 2744–2745, 1976. DOI: 10.1063/1.322999. eprint: https://doi.org/10.1063/1.322999. [Online]. Available: https://doi.org/10.1063/1.322999.

[386] L. K. Nanver, M. Krakers, T. Knezevic, *et al.*, "Investigation of light-emission and avalanche-current mechanisms in PureB SPAD devices", in *Fifth Conference on Sensors, MEMS, and Electro-Optic Systems*, M. du Plessis, Ed., International Society for Optics and Photonics, vol. 11043, SPIE, 2019, pp. 40 –53. [Online]. Available: https://doi.org/10.1117/12.2501598.

[387] M. A. Karami, M. H. Ghormishi, and A. Amiri-Sani, "Light emission enhancement in light emitting single-photon avalanche diodes (LESPADs)", *Optik*, vol. 126, no. 23, pp. 4146–4148, 2015, ISSN: 0030-4026. DOI: https://doi.org/10.1016/j.ijleo.2015.07.205. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0030402615007585.

[388] E Gros d'Aillon, L Verger, D. Bonifacio, *et al.*, "Microscale mapping of the photon detection probability of SPADs", in *IEEE Nuclear Science Symposium and Medical Imaging Conference*, 2016.

[389] D. Bonifacio, E Gros-Daillon, and L. Verger, "Microscale mapping of SPAD photon detection probability", *Journal of Instrumentation*, vol. 12, no. 05, P05010, 2017.

[390] M. Nemallapudi, S. Gundacker, P. Lecoq, and E. Auffray, "Single photon time resolution of state of the art SiPMs", *Journal of Instrumentation*, vol. 11, no. 10, P10016–P10016, 2016. DOI: 10.1088/1748-0221/11/10/p10016. [Online]. Available: https://doi.org/10.1088/1748-0221/11/10/p10016.

[391] K. Jensen, P. Hopman, E. Duerr, *et al.*, "Afterpulsing in Geiger-mode avalanche photodiodes for 1.06 μ m wavelength", *Applied Physics Letters*, vol. 88, no. 13, p. 133 503, 2006.

[392] M. Stipčević, D. Wang, and R. Ursin, "Characterization of a commercially available large area, high detection efficiency single-photon avalanche diode", *Journal of Lightwave Technology*, vol. 31, no. 23, pp. 3591–3596, 2013. DOI: 10.1109/JLT.2013.2286422.

[393] S. K. Yang, J. Lee, S.-W. Kim, *et al.*, "Precision measurement of the photon detection efficiency of silicon photomultipliers using two integrating spheres", *Optics Express*, vol. 22, no. 1, pp. 716–726, Jan. 2014. DOI: 10.1364/OE.22.000716.

[394] I. Rech, I. Labanca, G. Armellini, A. Gulinatti, M. Ghioni, and S. Cova, "Operation of silicon single photon avalanche diodes at cryogenic temperature", *Review of Scientific Instruments*, vol. 78, no. 6, p. 063 105, 2007. DOI: 10.1063/1.2743167. eprint: https://doi.org/10.1063/1.2743167. [Online]. Available: https://doi.org/10.1063/1.2743167.

[395] C. Hu, X. Zheng, J. C. Campbell, B. M. Onat, X. Jiang, and M. A. Itzler, "Characterization of an InGaAs/InP-based single-photon avalanche diode with gated-passive quenching with active reset circuit", *Journal of Modern Optics*, vol. 58, no. 3-4, pp. 201–209, 2011. DOI: 10.1080/09500340.2010.515749. eprint: https://doi.org/10.1080/09500340.2010.515749. [Online]. Available: https://doi.org/10.1080/09500340.2010.515749.

[396] F. Acerbi, M. Cazzanelli, A. Ferri, *et al.*, "High detection efficiency and time resolution integrated-passive-quenched single-photon avalanche diodes", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 268–275, 2014. DOI: 10.1109/JSTQE.2014.2341580.

[397] F. Shojaee, T. Haddadifam, and M. A. Karami, "Jitter modulation by photon wavelength variation in single-photon avalanche diodes (SPADs)", *Optical and Quantum Electronics*, vol. 53, no. 7, pp. 1–10, 2021.

[398] N. Da Dalt and A. Sheikholeslami, *Understanding Jitter and Phase Noise: A Circuits and Systems Perspective*. Cambridge University Press, 2018. DOI: 10.1017/9781316981238.

[399] N. Lusardi, F. Garzetti, S. Salgaro, *et al.*, "FPGA-based SiPM timestamp detection setup for high timing resolution TOF-PET application", p. 2, 2021. [Online]. Available: http://infoscience.epfl.ch/record/288700.

[400] G. Sportelli, M. G. Bisogni, C. Bruschini, *et al.*, "Towards the ideal pet detector: a scalable architecture with high intrinsic spatial resolution, doi and sub-200 ps tof capability", p. 2, 2021. [Online]. Available: http://infoscience.epfl.ch/record/288699.

[401] G. Sportelli, C. Bruschini, P. Carra, *et al.*, "Towards the ideal TOF-PET detector: a scalable architecture with uncompromised performance for clinical and total-body applications", 2021. [Online]. Available: http://infoscience.epfl.ch/record/288695.

[402] 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", p. 2, 2021. [Online]. Available: http://infoscience.epfl.ch/record/288697.

[403] E. Ciarrocchi, M. G. Bisogni, N. Camarlinghi, *et al.*, "Design of a highly scalable TOF-PET detector: the UTOFPET project", 2020. [Online]. Available: http://infoscience.epfl.ch/record/279892.

[404] N. Belcari, "UTOFPET: design of a highly scalable TOF-PET detector concept", 2019. [Online]. Available: http://infoscience.epfl.ch/record/279895.

[405] A. Muntean, F. Gramuglia, E. Venialgo, C. Bruschini, and E. Charbon, "Tradeoffs in cherenkov detection for positron emission tomography", in *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, 2018, pp. 1–2. DOI: 10.1109/NSSMIC.2018.8824430.

[406] C. Bruschini, C. Veerappan, F. Gramuglia, M. Bijwaard, Z. Papp, and E. Charbon, "A sensor network architecture for digital SiPM-Based PET systems", *Ieee Transactions On Radiation And Plasma Medical Sciences*, vol. 2, no. 6, pp. 574–587, 2018. DOI: 10.1109/TRPMS.2018.2866953. [Online]. Available: http://infoscience.epfl.ch/record/263674.

List of Publications

Conference Publications

1. F. Gramuglia, A. A. Muntean, C. A. Fenoglio, *et al.*, "CMOS 3D-Stacked FSI multi-channel digital SiPM for time-of-flight vision applications", 4. 69–72, R19, 2021. [Online]. Available: http://infoscience.epfl.ch/record/288691

2. N. Lusardi, F. Garzetti, S. Salgaro, *et al.*, "FPGA-based SiPM timestamp detection setup for high timing resolution TOF-PET application", p. 2, 2021. [Online]. Available: http://infoscience.epfl.ch/record/288700

3. G. Sportelli, M. G. Bisogni, C. Bruschini, *et al.*, "Towards the ideal pet detector: a scalable architecture with high intrinsic spatial resolution, doi and sub-200 ps tof capability", p. 2, 2021. [Online]. Available: http://infoscience.epfl.ch/record/288699

4. F. Gramuglia, A. A. Muntean, C. A. Fenoglio, *et al.*, "Architecture and characterization of a CMOS 3D-Stacked FSI multi-channel digital SiPM for time-of-flight PET applications", p. 2, 2021. [Online]. Available: http://infoscience.epfl.ch/record/288698

5. G. Sportelli, C. Bruschini, P. Carra, *et al.*, "Towards the ideal TOF-PET detector: a scalable architecture with uncompromised performance for clinical and total-body applications", 2021. [Online]. Available: http://infoscience.epfl.ch/record/288695

6. F. Gramuglia, M.-L. Wu, M. J. Lee, *et al.*, "SPAD microcells with 12.1 ps SPTR for SiPMs in TOF-PET applications", p. 2, 2021. [Online]. Available: http://infoscience.epfl.ch/record/288697

7. E. Ciarrocchi, M. G. Bisogni, N. Camarlinghi, *et al.*, "Design of a highly scalable TOF-PET detector: the UTOFPET project", 2020. [Online]. Available: http://infoscience.epfl.ch/record/279892

8. F. Gramuglia, A. A. Muntean, E. Venialgo Araujo, *et al.*, "CMOS 3D-Stacked FSI multi-channel digital SiPM for time-of-flight PET applications", 2020. DOI: 10.1109/NSS/MIC42677.2020.9507833. [Online]. Available: http://infoscience.epfl.ch/record/285321

9. N. Belcari, "UTOFPET: design of a highly scalable TOF-PET detector concept", 2019. [Online]. Available: http://infoscience.epfl.ch/record/279895

10. A. Muntean, F. Gramuglia, E. Venialgo, *et al.*, "Tradeoffs in cherenkov detection for positron emission tomography", in *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, 2018, pp. 1–2. DOI: 10.1109/NSSMIC.2018.8824430

11. F. Gramuglia, N. Descharmes, E. Venialgo, *et al.*, "Light extraction enhancement in scintillation crystals using thin film coatings", in *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, 2018, pp. 1–2. DOI: 10.1109/NSSMIC.2018.8824270

12. F. Gramuglia, M. Lee, E. Venialgo, *et al.*, "Towards 10ps 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)*, 2017, pp. 1–3. DOI: 10.1109/NSSMIC.2017.8532951

Journals Publications

1. F. Gramuglia, M.-L. Wu, C. Bruschini, *et al.*, "A low-noise CMOS SPAD pixel with 12.1 ps SPTR and 3 ns dead time", *IEEE Journal of Selected Topics in Quantum Electronics*, pp. 1–1, 2021. DOI: 10.1109/JSTQE.2021.3088216

2. F. Gramuglia, S. Frasca, E. Ripiccini, *et al.*, "Light extraction enhancement techniques for inorganic scintillators", *Crystals*, vol. 11, no. 4, 2021, ISSN: 2073-4352. DOI: 10.3390/cryst11040362. [Online]. Available: https://www.mdpi.com/2073-4352/11/4/362

3. F. Gramuglia, P. Keshavarzian, E. Kizilkan, *et al.*, "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 a publication of the IEEE Lasers and Electro-optics Society*, vol. 28, no. 2, 2022, ISSN: 1077-260X

4. F. Gramuglia, E. Ripiccini, C. A. Fenoglio, *et al.*, *Sub-10 ps minimum ionizing particle detection with Geiger-Mode APDs*, 2021. arXiv: 2111.09998 [physics.ins-det]

5. C. Bruschini, C. Veerappan, F. Gramuglia, *et al.*, "A sensor network architecture for digital SiPM-Based PET systems", *Ieee Transactions On Radiation And Plasma Medical Sciences*, vol. 2, no. 6, pp. 574–587, 2018. DOI: 10.1109/TRPMS.2018.2866953. [Online]. Available: http://infoscience.epfl.ch/record/263674