Development of a 3D Silicon Coincidence Avalanche Detector (3D-SiCAD) for charged particle tracking

Matteo Maria Vignetti

N° d'ordre NNT : 2017LYSEI017

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
INSA de Lyon

Ecole Doctorale N° ED 160
EEA (Électronique, Électrotechnique et Automatique)

Spécialité/discipline de doctorat :
Electronique, micro et nanoélectronique, optique et laser

Soutenue publiquement le 09/03/2017, par :
Matteo Maria VIGNETTI

Development of a 3D Silicon Coincidence Avalanche Detector (3D-SiCAD) for charged particle tracking

---

Devant le jury composé de :

DAUVERGNE Denis	Directeur de Recherche CNRS	*Université Grenoble Alpes*	Président
PANCHERI Lucio	Assistant Professor	*Université de Trente*	Rapporteur
UHRING Wilfried	Professeur des Universités	*Université de Strasbourg*	Rapporteur
PITTET Patrick	Ingénieur de Recherche	*Université Claude Bernard*	Examinateur
SAVOY-NAVARRO Aurore	Directeur de Recherche CNRS	*Université Paris Diderot*	Co-directrice de thèse
CALMON Francis	Professeur des Universités	*INSA-LYON*	Directeur de thèse
GOLANSKI Dominique	Ingénieur	*STMicroelectronics*	Invité
PARES Gabriel	Docteur - Ingénieur	*CEA - LETI*	Invité
ROCHAS Alexis	Docteur - Ingénieur	*CEA - LETI*	Invité

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec : Renée EL MELHEM Bat Blaise Pascal 3e etage secretariat@edchimie-lyon.fr Insa : R. GOURDON	M. Stéphane DANIELE Institut de Recherches sur la Catalyse et l'Environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 avenue Albert Einstein 69626 Villeurbanne cedex directeur@edchimie-lyon.fr
E.E.A.	**ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec : M.C. HAVGOUDOUKIAN Ecole-Doctorale.eea@ec-lyon.fr	M. Gérard SCORLETTI Ecole Centrale de Lyon 36 avenue Guy de Collongue 69134 ECULLY Tél : 04.72.18 60.97 Fax : 04 78 43 37 17 Gerard.scorletti@ec-lyon.fr
E2M2	**EVOLUTION, ECOSYSTEME, MICROBIOLOGIE, MODELISATION** **http://e2m2.universite-lyon.fr** Sec : Sylvie ROBERJOT Bât Atrium - UCB Lyon 1 04.72.44.83.62 Insa : H. CHARLES secretariat.e2m2@univ-lyon1.fr	M. Fabrice CORDEY CNRS UMR 5276 Lab. de géologie de Lyon Université Claude Bernard Lyon 1 Bât Géode 2 rue Raphaël Dubois 69622 VILLEURBANNE Cédex Tél : 06.07.53.89.13 cordey@univ-lyon1.fr
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTE** **http://www.ediss-lyon.fr** Sec : Sylvie ROBERJOT Bât Atrium - UCB Lyon 1 04.72.44.83.62 Insa : M. LAGARDE secretariat.ediss@univ-lyon1.fr	Mme Emmanuelle CANET-SOULAS INSERM U1060, CarMeN lab, Univ. Lyon 1 Bâtiment IMBL 11 avenue Jean Capelle INSA de Lyon 696621 Villeurbanne Tél : 04.72.68.49.09 Fax :04 72 68 49 16 Emmanuelle.canet@univ-lyon1.fr
INFOMATHS	**INFORMATIQUE ET MATHEMATIQUES** **http://infomaths.univ-lyon1.fr** Sec :Renée EL MELHEM Bat Blaise Pascal 3e etage infomaths@univ-lyon1.fr	Mme Sylvie CALABRETTO LIRIS – INSA de Lyon Bat Blaise Pascal 7 avenue Jean Capelle 69622 VILLEURBANNE Cedex Tél : 04.72. 43. 80. 46 Fax 04 72 43 16 87 Sylvie.calabretto@insa-lyon.fr
Matériaux	**MATERIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec : M. LABOUNE PM : 71.70 –Fax : 87.12 Bat. Direction Ed.materiaux@insa-lyon.fr	M. Jean-Yves BUFFIERE INSA de Lyon MATEIS Bâtiment Saint Exupéry 7 avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél : 04.72.43 71.70 Fax 04 72 43 85 28 jean-yves.buffiere@insa-lyon.fr
MEGA	**MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE** **http://mega.universite-lyon.fr** Sec : M. LABOUNE PM : 71.70 –Fax : 87.12 Bat. Direction mega@insa-lyon.fr	M. Philippe BOISSE INSA de Lyon Laboratoire LAMCOS Bâtiment Jacquard 25 bis avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél : 04.72 .43.71.70 Fax : 04 72 43 72 37 Philippe.boisse@insa-lyon.fr
ScSo	**ScSo*** **http://recherche.univ-lyon2.fr/scso/** Sec : Viviane POLSINELLI Brigitte DUBOIS Insa : J.Y. TOUSSAINT Tél : 04 78 69 72 76 viviane.polsinelli@univ-lyon2.fr	M. Christian MONTES Université Lyon 2 86 rue Pasteur 69365 LYON Cedex 07 Christian.montes@univ-lyon2.fr

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Acknowledgements

This work has been carried out in the *Electronic Devices* group at the *Institut des Nanotechnologies de Lyon* (INL), UMR 5270 – CNRS, and has been funded by the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program (FP7 2007-2013) under grant agreement N° 317446 *"INFIERI"* (*Intelligent Fast Interconnected and Efficient Devices for Frontier Exploitation in Research and Industry*).

I would like to express my sincere gratitude to my supervisor prof. Francis Calmon for his wise and continuous guidance and encouragement throughout the period of my Ph.D work. I wish to thank my co-supervisor Dr. Aurore Savoy-Navarro, for her precious advices and, most importantly, for her incredible work in leading the *INFIERI* network.

I would like to thank Dr. Patrick Pittet, for his insightful comments and challenging questions which incented me to widen my research from various perspectives. My sincere thanks also go to Dr. Laurent Quiquerez and Dr. Remy Cellier.

The realization of our 3D prototype was made possible thanks to Irène Pheng, from CIME-Nanotech in Grenoble, and Dr. Gabriel Parès, from the DCOS group of CEA-LETI in Grenoble, who gave me the opportunity to join his team as a visiting student. Last but not least, I wish to thank the ICube Laboratory in Strasbourg and the *Institut de Physique Nucléaire de Lyon* for their kind support on the tape-out and characterization of our 3D prototype.

Table of Contents

Introduction

The development and optimization of position sensitive detectors for charged particle tracking systems is particularly challenging in the field of the High Energy Physics (HEP) experiments as well as in emerging Medical Physics applications. Design and complexity have been increasingly evolving over the years in order to attain more aggressive performances for particle tracking systems in terms of low material budget, low noise, very high spatial resolution, radiation hardness, low power consumption and cost-effectiveness.

High Energy Physics (HEP) aims at studying the nature of elementary particles that constitute the matter. In HEP experiments, particle accelerators boost beams of particles up to very high energies before they collide with each other to produce other particles. The resulting by-products of these collisions are observed and recorded by means of dedicated and complex detection systems consisting of multiple layers of pixelated detectors which track and identify the particles passing through them. In the most powerful particle accelerator in the world, the Large Hadron Collider (LHC) at CERN in Geneva (Switzerland), two beams of hadrons are accelerated in opposite directions in a $27\ km$ ring up to a record energy of $13\ TeV$ and very high luminosity, on the order of $10^{34}\ cm^{-2}s^{-1}$ [1]. The collisions occur in four particle detectors that are placed in different locations over the ring and in charge of detecting and recording the resulting particles track: A Toroidal LHC Apparatus (ATLAS), Compact Muon Solenoid (CMS), A Large Ion Collider Experiment (ALICE) and Large Hadron Collider beauty (LHCb) [1]. Even if these detectors have been realized by implementing different technical solutions, they all share the same scientific goal: studying the Standard Model to search for extra dimensions and particles that could make up dark matter. The planned luminosity upgrade of the LHC to the High Luminosity LHC (HL-LHC) is proposed to operate at an instantaneous luminosity of $5 \cdot 10^{34}\ cm^{-2}s^{-1}$ for a further 10 years of operation. Proton-proton collisions are expected to occur synchronously with a period of $25\ ns$, i.e. $40\ MHz$ bunch crossing rate. Particle detectors are thus expected to be upgraded in order to operate in this extremely challenging environment. More specifically, the extraction of precise vertex information for each track is of crucial importance for a correct track reconstruction and thus, particle identification. The primary vertex is extracted from high-resolution position measurements near the interaction point by a vertex detector. For this upgrade the inner layer of the tracker detector is required to feature high granularity ($50 \times 50\ \mu m^2$ or $25 \times 100\ \mu m^2$) and a hit rate per pixel approximately between $25 - 50\ kHz$ with a hit loss probability of 10^{-3}, and a rate of fake hits (noise hit rate) significantly lower than the particle hit rate by at least a factor of three. Moreover the material budget and power dissipation have to be kept sufficiently low to values close to $1\%\ X_0$ (X_0 is the radiation length [2]) and $0{,}4\ W/cm^2$, respectively, and the radiation hardness should cope with $1\ Grad$ total dose over 10 years [3][4].

In emerging fields of medical physics such as hadron-therapy and proton computed tomography (pCT), tracking systems are used for the identification of the position and direction of protons delivered to a patient. Hadron-therapy is a medical technique that uses a beam of charged particles (protons, carbon ions) to irradiate diseased tissue, for the treatment of cancer. The main advantage of proton therapy over other types of external X-ray radiotherapy is that most of the dose of the charged particle is deposited over a narrow spatial range (referred to as "Bragg peak"), leaving a minimal exit dose which spares the healthy tissues all along the particle path [5]. One of the major challenges in this field is the realization of a compact and versatile detection system which is suitable for the various applications needed in hadron-therapy such as proton beam monitoring for quality assurance (beam cartography) and real-time control and monitoring of the ion range to ensure the correct dose delivery to the patient for an in-vivo verification during the irradiation time [6][7][8]. Improved image quality and timing resolution are required to this end. The ideal detector would thus require similar features to the ones from HEP experiments in term of granularity ($50 \times 50\,\mu m^2$). The hit rate will conversely depend on the application: from very low rate for pCT, i.e. $< 1\,Hz$ for $50 \times 50\,\mu m^2$ pixel size, i.e. particle flux of about $10^4\,s^{-1}cm^{-2}$ (where the noise can thus represent an important issue), to high rate ($> 10\,MHz$) for beam monitoring.

The aim of this work is to develop a CMOS demonstrator of a novel detector suitable for charged particle tracking systems. This novel detector has been referred to as *3D Silicon Coincidence Avalanche Detector* (3D-SiCAD) or *Avalanche Pixel Sensor* [9][4] (APiX), and consists of an array of 3D pixel cells based on a pair of vertically aligned Geiger-mode avalanche diodes operating in coincidence mode which are electrically connected by means of 3D integration techniques. As represented schematically in Figure 1, the detection of a charged particle hitting the detector at a given location consists of revealing the occurrence of two coincident avalanche events which are simultaneously produced in each sensing level of the 3D-SiCAD pixel cell. With respect to simple Geiger-mode avalanche diodes (or Single Photon Avalanche Diodes, SPAD), this concept allows discriminating true events, i.e. a charged particle hit over the pixel, from false and random avalanche counts occurring independently in each SPAD device of the two sensing levels due to the undesired detection of background photons and dark counts. On the one hand, a background photon can be detected by only one of the two sensing levels. On the other hand, dark counts occurring in a sensing level are statistically uncorrelated to the dark counts occurring in the other one. Nevertheless, false coincidences may be detected when two random dark counts from the two sensing levels occur within the finite time window required by the electronics for the coincidence check. Such a time-window is actually limited by the shortest time that the electronics is capable to resolve. The rate at which these false coincidences are expected to occur in a 3D-SiCAD is referred to as *Fake Coincidence Rate (FCR)*.

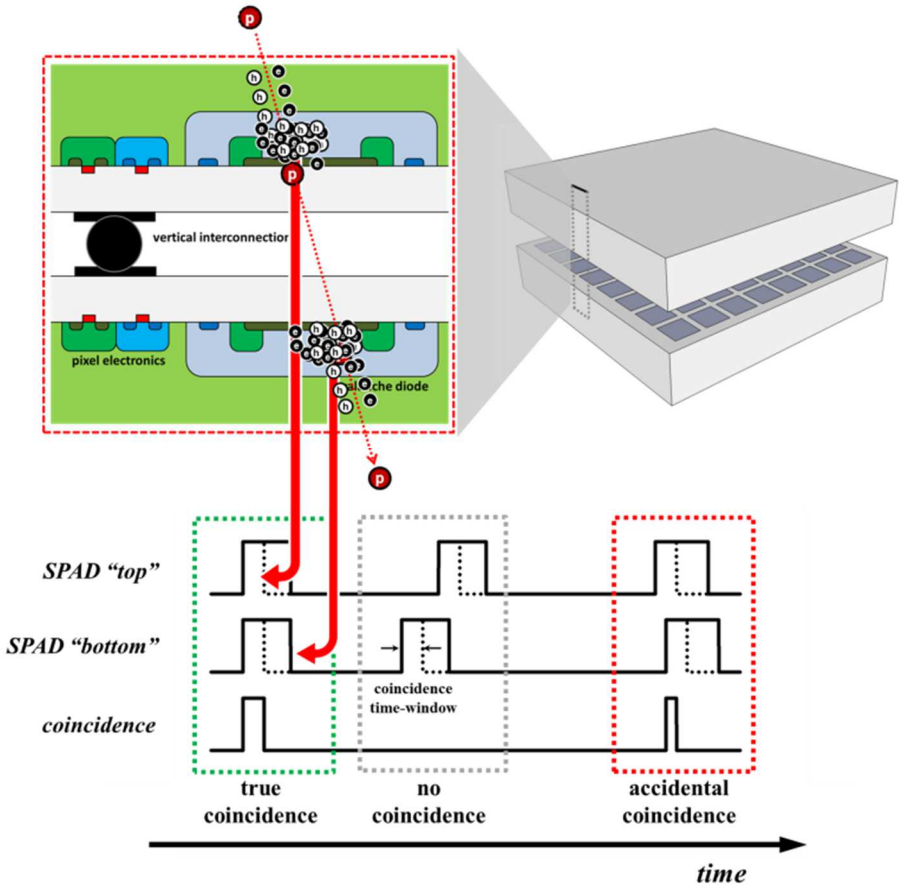

Figure 1: *Schematic representation of a 3D-SiCAD pixel considering a (qualitative) possible implementation in a CMOS process by adopting a stud-bump vertical interconnection. A qualitative time-diagram representing the main waveforms in a 3D-SiCAD pixel is reported in the bottom part of the picture. Solid line waveforms associated to the "top" and "bottom" SPADs represent the output signals produced by the quenching electronics in each sensing level. The dotted lines represent the ultra-short pulses that are synchronous to the leading edge of an avalanche event. The width of the synchronous pulses represent the coincidence time-window. Observe that even a partial overlap between these pulses produces a coincidence count. The adoption of an ultra-short coincidence time-window is the only way to improve the noise rejection of the detector.*

Nevertheless, the coincidence operating mode of this novel detector is expected to enable excellent noise rejection capability with respect to conventional SPAD based detectors, such as *Silicon Photo-Multipliers* (SiPM). According to Figure 2 (more details in Chapter 1), the expected *FCR* achieved with a low cost standard technology is indeed much lower than the *Dark Count Rate* (*DCR*) achievable with expensive dedicated detector technologies. Moreover, it is worth observing that one of the most important advantages provided by Geiger-mode avalanche diodes is the practically "infinite" internal charge multiplication gain [10], inherent of the avalanche multiplication process by impact ionization.

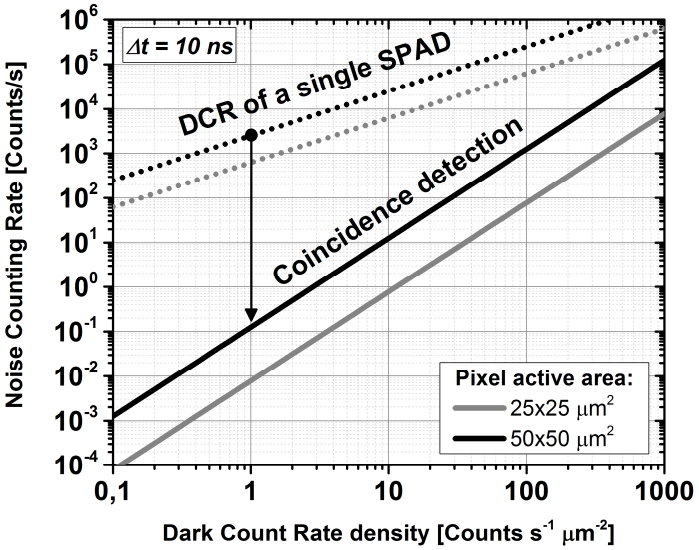

Figure 2: *Comparison of the expected noise counting rate levels between a 3D-SiCAD pixel and a conventional SPAD pixel, as a function of the Dark Count Rate density (number of dark counts per seconds per unit surface affecting a SPAD device), i.e. a technology dependent figure of merit. For noise density values typically achievable in standard CMOS process, i.e. $DCR' \sim$ a few counts $s^{-1}\mu m^{-2}$, the expected FCR is four orders of magnitude lower than the DCR of a conventional SPAD cell (in case of a device active area of $50\,\mu m \times 50\,\mu m$ and a coincidence time window $\Delta t = 10\,ns$). Noise rejection improves for smaller pixel active areas, and shorter coincidence-time windows.*

This intrinsic feature of SPAD devices allows thinning down the silicon substrate of 3D-SiCADs down to unprecedented figures for silicon based detectors, enabling tremendous improvement on the material budget. Moreover a SPAD pixel provides a direct conversion of a single particle event to a well-defined voltage pulse with sub-nanoseconds resolution which is compatible with typical CMOS standards and, more importantly, removes the need of preamplifiers or pulse shapers in the read-out electronics. The CMOS compatibility is probably the most attractive feature for this novel device, as the realization of a 3D-SiCAD detector system can benefit of the continuous progresses in CMOS technology developments and the nowadays available 3D integration techniques. Therefore a 3D-SiCAD pixel is expected to provide single charged particle detection capability despite of the small thickness of the SPAD active region (i.e. a few microns) and the inherent fluctuations of the ionization yield.

The objective of this work is to develop a 3D-SiCAD demonstrator based on a commercial CMOS technology and common 3D integration techniques, consisting of simple single-pixel and small matrix cells. The aim is to demonstrate the feasibility of this novel detector and to validate the expected performances in terms of excellent particle detection efficiency and noise rejection capability with respect to background counts.
This manuscript is organized as follows.

Chapter 1 reviews the physics, the working principle, the figures of merit and the state-of-the art of Single Photon Avalanche Diodes (SPADs), i.e. the building block of the 3D-SiCAD detector developed in the present work. Moreover the working principle of this latter device is here discussed in more details, and the main figures of merit and the current state-of-the art are presented.

Chapter 2 describes the development steps that have been faced for the design of a first 3D-SiCAD demonstrator: the design of a SPAD pixel cell in a commercial high voltage 0,35 μm CMOS process, consisting of a proper avalanche diode architecture with associated quenching electronics to ensure correct Geiger-mode operation; the design of the 3D-level pixel electronics to provide a proper interfacing between the two sensing levels in a 3D-SiCAD pixel and, more importantly, assessing the occurrence of coincidence hits; the study of a 3D integration strategy impacting the layout of the test-chip and finally the choice of the assembling technique for the realization of a first 3D prototype.

Chapter 3 shows the results of the characterization of the SPAD cells constituting the building block for the 3D-SiCAD pixels. The characterization includes: the current-voltage curves of the avalanche diode and the breakdown voltage as a function of the temperature; the electro-luminescence test to assess the quality of the avalanche diode architecture for Geiger-mode operation; the validation of the quenching electronics; the study of the SPAD noise performance in terms of dark counts and after-pulsing; the photon detection efficiency.

Chapter 4 shows the characterization results of a single 3D-SiCAD pixel from a first 3D prototype. The results include: the noise performance in coincidence-mode operation to validate the device noise rejection capability; the particle detection efficiency by means of inverse square-law measurements by adopting a Strontium-90 radioactive source.

The research leading to these results has received funding from the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program (FP7 2007-2013) under grant agreement n° 317446 *INFIERI (Intelligent, Fast, Interconnected and Efficient devices for Frontier Exploitation in Research and Industry)*. *INFIERI* is an inter-disciplinary and multi-national network of research centers and industries aimed at training young physicists and engineers in developing, designing and managing intelligent devices and tools for cutting-edge applications in fundamental physics research and its technological spin-offs such as Astrophysics, Particle Physics, Medical Physics and Telecommunications.
This work has been carried out within the framework of the Work Package 2 of the *INFIERI* network, wherein the main objectives have been to disseminate, implement and study the benefits provided by the most recent advances in very deep submicron technologies and 3D integration techniques in the fields of high energy physics and astrophysics experiments, and medical applications.

References

[1] "The Large Hadron Collider at CERN." [Online]. Available: http://home.cern/topics/large-hadron-collider.

[2] J. Beringer et al, "Review of Particle Physics*," *Phys. Rev. D*, vol. 86, no. 1, p. 10001, 2012.

[3] J. Christiansen, "Outline and requirements of Phase 2 Pixel system and Read-Out Chip," 2014.

[4] N. D'Ascenzo, P. S. Marrocchesi, C. S. Moon, F. Morsani, L. Ratti, V. Saveliev, A. S. Navarro, and Q. Xie, "Silicon avalanche pixel sensor for high precision tracking," *J. Instrum.*, vol. 9, no. 3, p. C03027, 2014.

[5] D. Schardt, T. Elsässer, and D. Schulz-Ertner, "Heavy-ion tumor therapy: Physical and radiobiological benefits," *Reviews of Modern Physics*, vol. 82, no. 1. pp. 383–425, 2010.

[6] R. Schulte, V. Bashkirov, T. Li, Z. Liang, K. Mueller, J. Heimann, L. R. Johnson, B. Keeney, H. F. W. Sadrozinski, A. Seiden, D. C. Williams, L. Zhang, Z. Li, S. Peggs, T. Satogata, and C. Woody, "Conceptual design of a proton computed tomography system for applications in proton radiation therapy," in *IEEE Transactions on Nuclear Science*, 2004, vol. 51, no. 3 III, pp. 866–872.

[7] C. Golnik, F. Hueso-González, A. Müller, P. Dendooven, W. Enghardt, F. Fiedler, T. Kormoll, K. Roemer, J. Petzoldt, A. Wagner, and G. Pausch, "Range assessment in particle therapy based on prompt γ -ray timing measurements," *Phys. Med. Biol.*, vol. 59, no. 18, p. 5399, 2014.

[8] J. Krimmer, L. Balleyguier, D. Dauvergne, N. Freud, J. Hérault, J. M. Létang, H. Mathez, M. Pinto, E. Testa, and Y. Zoccarato, "Prompt-gamma detection towards absorbed energy monitoring during hadrontherapy," *ANNIMA Conf.*, 2015.

[9] V. Saveliev, "Avalanche Pixel Sensor and Related Methods," US patent 8269181, 2012.

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

Chapter 1: 3D Silicon Coincidence Avalanche Detector

This chapter reviews the physics, the working principle, the figures of merit and the state-of-the art of Single Photon Avalanche Diodes (SPADs), i.e. the building block of the 3D-SiCAD detector developed in the present work. Moreover the working principle of this latter device is here discussed in more details, and the main figures of merit and the current state-of-the art are presented.

1.1 Geiger-mode Avalanche Diodes

Geiger-mode Avalanche Diodes, also referred to as Single Photon Avalanche Diodes (SPADs), are p-n junctions working at a reverse bias well above the breakdown voltage V_{bd}. These devices achieve single-photon detection capability by exploiting the avalanche multiplication process responsible for the breakdown current in an avalanche diode [1]. Under this bias condition, both free electrons and holes in the space charge region (SCR) are indeed accelerated up to a point where they can each produce an electron-hole pair (EHP) by impact ionization, i.e. by breaking a covalent bond when colliding with the semiconductor lattice atoms, thus ionizing a valence electron from the valence band to the conduction band [2]. The initial carriers together with the just-generated pairs are then subject to the same process, which leads to a positive feed-back loop of ionizations, and eventually sets-up a self-sustained multiplication of the carriers. The charge multiplication can in principle diverge to infinity but in reality the resulting breakdown current – as it is observed in the classic I-V curve of a diode - is finite because of the finite internal resistance of the SCR, responsible for lowering the voltage drop across the junction [3].

1.1.1 SPAD working principle

Single-photon detection capability arises from the fact that even a single EHP collected into the multiplication region of the diode, i.e. the SCR, can initiate an avalanche multiplication process providing a macroscopic electrical signal. Under this scenario, the leading edge of the avalanche current would mark the occurrence of the physical phenomenon, i.e. the event, related to EHP creation or injection in the multiplication region, which can be a single-photon absorption by photo-electric effect [1], but also ionization effect due to the passage of a charged particle [4], generation – recombination phenomena in the SCR, etc. In order to allow the detection of other events occurring after the one that initiated

the first avalanche, a SPAD pixel needs suitable driving electronics. This is responsible for interrupting (quenching) the multiplication process right after the avalanche build-up by promptly lowering the reverse bias of the junction at or below the breakdown voltage. Then, after a certain dead-time (hold-off time) during which the pixel cannot allow any multiplication process, the electronics restores the p-n junction to the initial bias (reset phase), and the device is finally ready to detect another event (Figure 1). This mode of operation is referred to as "Geiger-mode" in analogy with Geiger-Muller counters used for the detection of ionizing radiation in nuclear physics. It is important to point out that a SPAD pixel can only provide the information that a certain physical event has occurred, since the output pulses have the same amplitude regardless of the amount of charge injected into the diode SCR. If, for instance, two or more photons are absorbed simultaneously in the diode, the amplitude of the avalanche pulse is the same as in the case of single-photon detection. Therefore the information about the energy released by the physical source generating free carriers in the device (e.g. the amount of absorbed photons) is not available in SPAD detectors. The Geiger-mode condition can be finally defined quantitatively by the following condition [5]:

$$1 \leq \int_0^W \alpha_e \cdot exp\left(\int_x^W (\alpha_h - \alpha_e)dx'\right)dx \qquad (1)$$

where W is the SCR width, and α_e, α_h are the ionization coefficients (or ionization rate) for the electrons and holes, respectively, defined as the number of electron-hole pairs generated by impact ionization by a carrier per unit distance traveled.

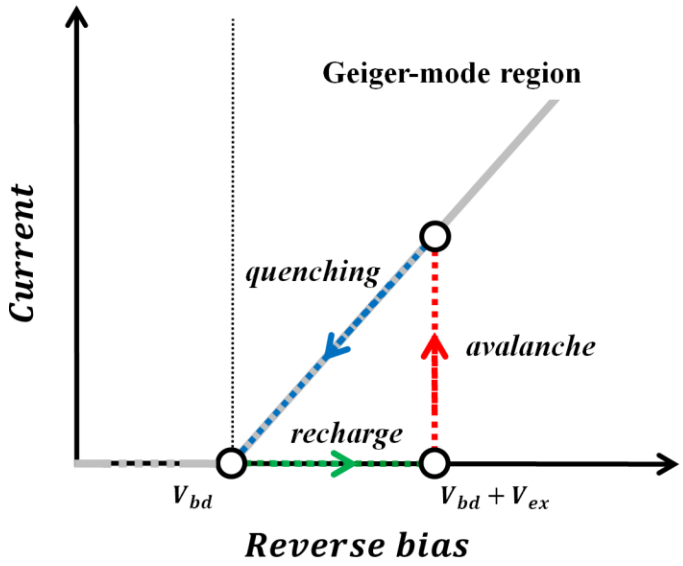

Figure 1: *Qualitative representation of the "Geiger-mode cycle", showing the main phases characterizing the dynamic behavior of a SPAD.*

The bias providing an equality in the above condition defines the diode breakdown voltage V_{bd}. The ionization coefficients increase with increasing electric field over the junction[1], and decrease with increasing temperature [2]. In silicon, if the diode is biased slightly below the breakdown voltage, the electric field over the SCR is only sufficient for electrons to cause significant ionization but not holes (due to the lower ionization coefficient for holes with respect to electrons in Silicon) and the diode operating mode is referred to as "linear mode" [5]. Under this scenario, the charge resulting from the multiplication process is indeed proportional to the amount of charge injected initially, according to a modest multiplication gain (around a few hundred) which is however affected by strong statistical fluctuations [3]. The multiplication process quenches itself as soon as there is no more carrier injection in the depleted region of the diode.

1.1.2 Detection Efficiency

Due to the statistical nature of the avalanche multiplication process, an EHP injected in the SCR of an avalanche diode has a certain probability to successfully initiate a breakdown process. Not every injected (or generated) carrier can indeed initiate a self-sustained multiplication process: some carrier will not produce any ionization before leaving the device active region, while others will initiate an ionization process resulting, however, in a short chain of carriers that eventually fade-out [6]. The probability for an EHP to successfully initiate an avalanche multiplication process, referred to as "Avalanche Triggering Probability" (ATP), can be evaluated with the help of [6], by solving the following couple of differential equations [7]:

$$\begin{cases} \dfrac{dP_e}{dx} = \alpha_e(1-P_e)(P_e + P_h - P_eP_h) \\ \dfrac{dP_h}{dx} = -\alpha_h(1-P_h)(P_e + P_h - P_eP_h) \end{cases} \quad (2)$$

where $\alpha_e(x)$ and $\alpha_h(x)$ are the electrons and holes ionization coefficients, respectively, while $P_e(x)$, $P_h(x)$ are the probabilities for initiation an avalanche by an electron or a hole, respectively, generated at the position x within the space charge region of the avalanche diode.

[1] Concerning this point, it is important to point out that conventional model for impact ionization in semiconductors assumes that the ionization coefficients are function only of the local value of the electric field. However, it has been recognized that this is a poor approximation for diodes having very thin SCR [43]. A carrier starting with near zero energy, relative to the band edge, will have indeed an ionizing collision probability close to zero until the energy it has gained from the electric field will be sufficiently high to allow ionization. Therefore a more accurate model proposed by McIntyre for the characterization of the ionization properties of semiconductors, relies on history-dependent ionization coefficients, based on a set of field-profile-independent parameters [43].

These equations can be integrated by taking into account the fact that carriers exiting the SCR cannot trigger any multiplication process, i.e. by adopting the boundary condition $P_h(x = x_p) = P_e(x = x_n) = 0$ (where x_p and x_n are the boundaries of the diode SCR at the p-side and n-side, respectively) [6]. The ATP at a position x can be finally obtained as the joint probability of P_e and P_h, i.e. as the sum of the probability for a primary electron to initiate an avalanche multiplication process and the probability for a primary hole to do it if the electron does not succeed [6]:

$$P_{tr}(x) = P_e(x) + P_h(x)\big(1 - P_e(x)\big) \tag{3}$$

Due to the dependency of ATP on the carrier ionization rates, the initiation probability will thus depend on the semiconductor material, temperature of operation and the electric field (magnitude and spatial distribution). As shown in Figure 2 (obtained from post-processing calculations after TCAD simulation of a SPAD conceived in a 28nm CMOS FDSOI technology [8]) the calculated ATP increases for higher reverse bias voltages beyond the breakdown threshold, since the higher electric field magnitude raises the probability for the carriers to ionize the lattice atoms in the SCR of the SPAD. The figure reports also the plot of the reverse bias I-V curve of the avalanche diode, showing that the ATP is approximately zero in proximity of the breakdown voltage.

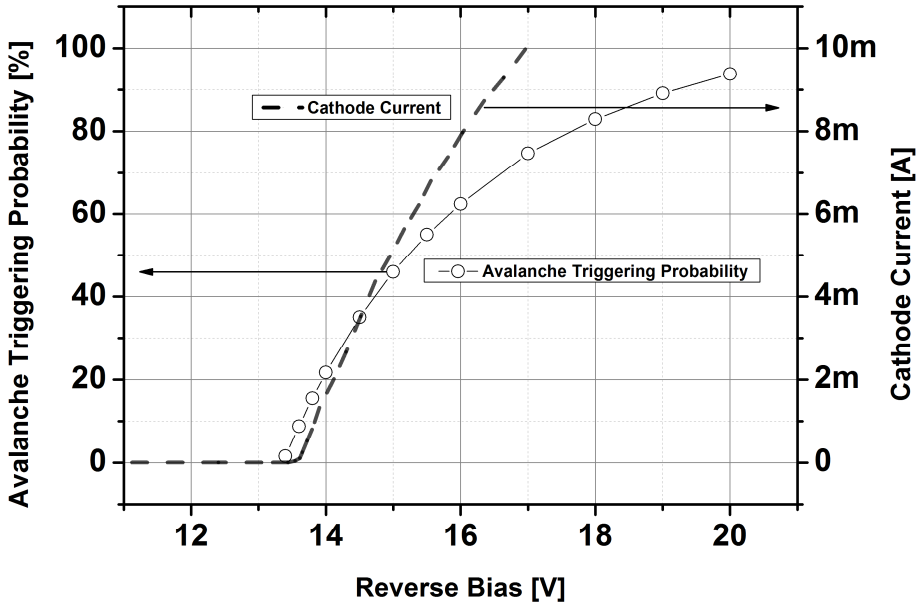

Figure 2: *Simulation results* [8] *of a SPAD conceived in a 28nm CMOS FDSOI technology. Symbols, left y-axes: average avalanche triggering probability; Dashes, right y-axes: reverse bias I-V curve of the avalanche diode. Observe that the SPAD active area diameter is D=7μm).*

1.1.2.2 Photon Detection Efficiency

The Photon Detection Efficiency (PDE), also referred to as Photon Detection Probability (PDP), is defined as the probability that a photon of a certain wavelength hitting a SPAD pixel, successfully fires a self-sustained avalanche multiplication process [1]. In order for this to happen, the following conditions must be satisfied:

- the incoming photon must produce a carrier injection into the SPAD multiplication region.
- the injected carrier must successfully fire an avalanche multiplication process.

A simple expression of the PDE can be therefore obtained as the product of the diode quantum efficiency (QE) and the Avalanche Triggering Probability (ATP)[3], as follows:

$$PDE(\lambda) = FF\big(1 - R(\lambda)\big)\overline{QE(\lambda)} \cdot \overline{P_{tr}} \tag{4}$$

The expression takes into account the reflection probability $R(\lambda)$ of the incoming photon along its optical path, as well as the optical losses due to the "blind" region surrounding the active area of the SPAD pixel, expressed in terms of Fill Factor, i.e. $FF = Surface_{active-region}/Surface_{pixel}$.
This formula represents however a simple approximation for the PDP, since the avalanche triggering probability is dependent on the position where the carrier has been generated. A photon with wavelength λ hitting a SPAD pixel has indeed a chance to be successfully detected only if this one is absorbed in the SPAD active area without being back-reflected along its optical path. In common CMOS avalanche diode architectures, the active area includes the space charge region (where the carriers are accelerated and multiplied) but also the nearby p / n neutral regions, up to an extension of around the minority carrier diffusion length (Figure 3). If the photon is absorbed in the multiplication region, as depicted in the scenario (a) in Figure 3, the local high electric field promptly accelerates the generated EHP towards the SCR ends, which might lead to a self-sustained multiplication process. Similarly, if the photon is absorbed in one of the two nearby neutral regions, i.e. scenarios (b) and (c) in Figure 3, an avalanche process may be initiated provided that a minority carrier belonging to the generated EHP manages to diffuse towards the SPAD depleted region, where it can be accelerated by the electric field and can eventually fire a breakdown process. Under this latter scenario the avalanche can be initiated only by the minority carrier diffusing in the considered neutral region, i.e. n or p. A majority carrier diffusing towards the multiplication region, e.g. an electron coming from the n-type side, would be immediately ejected by the electric field of the SCR, due to the field direction. If, conversely, the photon is absorbed outside the active region shown in Figure 3, the generated EHP has no chance to reach the multiplying region, meaning that the photon is irremediably lost.

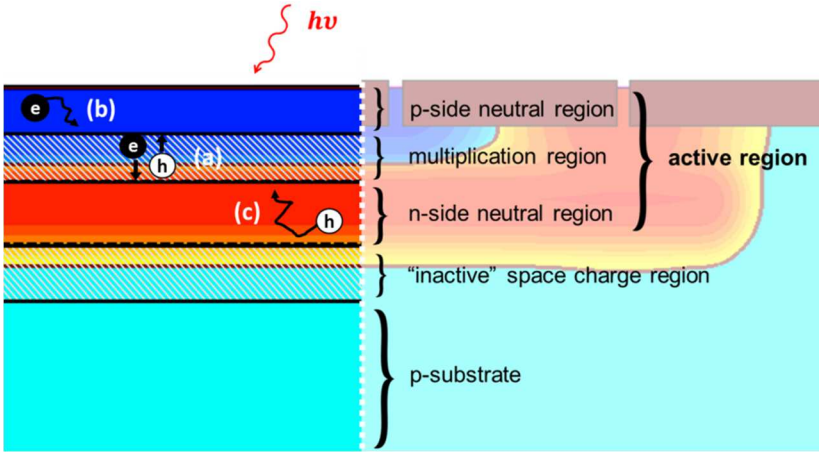

Figure 3: *Schematic representation of the photon absorption in the active region of a SPAD pixel. Three main scenarios have been depicted: (a) the photon is absorbed in the multiplication region and the generated EHP are swift away by the high electric field. (b) and (c) the photon is absorbed in the upper p-type / lower n-type neutral region and the generated minority electron / hole randomly diffuses towards the multiplication region where it can eventually fire an avalanche multiplication process.*

The generated carriers may indeed recombine along the random diffusive walk beyond the diffusion length boundary, or can be simply collected by other "secondary" SCR depending on the diode architecture (e.g. in the SCR referred to as "inactive" in Figure 3). In order to provide a more complete analytical formula for the PDE in a SPAD, it is convenient to simplify the study in a one-dimensional case as in practice it is really unlikely that minority carriers generated in device sensitive region can escape laterally. Only some of those generated in proximity to the device periphery will be probably lost, which justifies the one-dimensional approximation [9]. According to the previous discussion, the photon detection efficiency can be thus calculated as an integral sum over the SPAD geometry of the product between [9]:

- the probability $p_{abs}(x,\lambda)dx$ that an EHP is generated within x and $x+dx$ in the SPAD, after a photon absorption
- the probability $\eta_c(x)$ that the EHP generated at that location is collected in the diode SCR, referred to as "carrier collection efficiency"
- the probability $P_{tr}(x)$ that the collected carrier successfully fires an avalanche multiplication process

PDP can be finally expressed as follows:

$$PDE(\lambda) = FF\big(1 - R(\lambda)\big) \int p_{abs}(x,\lambda) n_c(x) P_{tr}(x) dx \qquad (5)$$

Observe that the product $n_c(x)P_{tr}(x)$ can be interpreted as an EHP detection probability such that:

$$n_c(x)P_{tr}(x) = \begin{cases} P_{tr}(x), & in\ the\ depletion\ region \\ \eta^*_{c,n}(x)P_{tr}(x_n), & in\ the\ n-type\ region \\ \eta^*_{c,p}(x)P_{tr}(x_p), & in\ the\ p-type\ region \\ 0, & elsewhere \end{cases} \quad (6)$$

The "carrier collection efficiency" $\eta_c(x)$ is indeed zero outside the SPAD active region, and is 100% in the diode SCR, while ATP for carriers diffusing from the neutral regions is given by the value of $P_{tr}(x)$ at the boundaries of the SCR. It is important to observe that the carrier collection efficiency is a parameter that cannot be estimated in a simple way since it is a result of the drift-diffusion carrier transport within the neutral region, which is affected by the semiconductor transport properties, doping profiles and boundary condition at the neutral region edges [9]. Naturally this parameter has an important effect on the PDE, especially for SPAD having neutral regions much wider than the SCR.
The PDE dependence on the photon wavelength is conversely related to the light absorption process in silicon, which can take place only if the photon has enough energy to excite an electron from the valence band to the conduction band by photo-electric effect, i.e. the photon energy E_{ph} has to be larger than the Silicon bandgap E_{GAP}. This condition translates into a cut-off wavelength beyond which the semiconductor is fully transparent to the incoming photon:

$$\lambda_C = \frac{hc}{E_{GAP}}$$

where h and c are the Planck constant and the speed of light in vacuum, respectively. In the case of Silicon, $\lambda_{C-Si} \approx 1.1\mu m$ ($E_{GAP} = 1.12eV$ at room temperature) meaning that only photons having wavelength shorter than this one are absorbed while they travel in the semiconductor. In general, the photon absorption process in a semiconductor can be described in terms of a wavelength-dependent parameter known as photon absorption coefficient $\alpha(\lambda)$, related to the probability that a photon is absorbed while traveling within an infinitesimal path dx. The probability that a photon traveling through a semiconductor is absorbed within x and $x + dx$, i.e. $p_{abs}(x,\lambda)dx$, is thus given by the probability that this one is absorbed within such an infinitesimal segment, i.e. $\alpha(\lambda)dx$, conditioned by the probability that the photon has not been absorbed before, i.e. $\big(1 - P_{abs}(x,\lambda)\big)$:

$$p_{abs}(x,\lambda)dx = \big(1 - P_{abs}(x,\lambda)\big)\alpha(\lambda)dx$$

By solving this differential equation ($p_{abs}(x,\lambda) = \partial P_{abs}(x,\lambda)/\partial x$) the photon absorption probability density is eventually obtained:

$$p_{abs}(x,\lambda) = e^{-\alpha(\lambda)x}\alpha(\lambda) \quad (7)$$

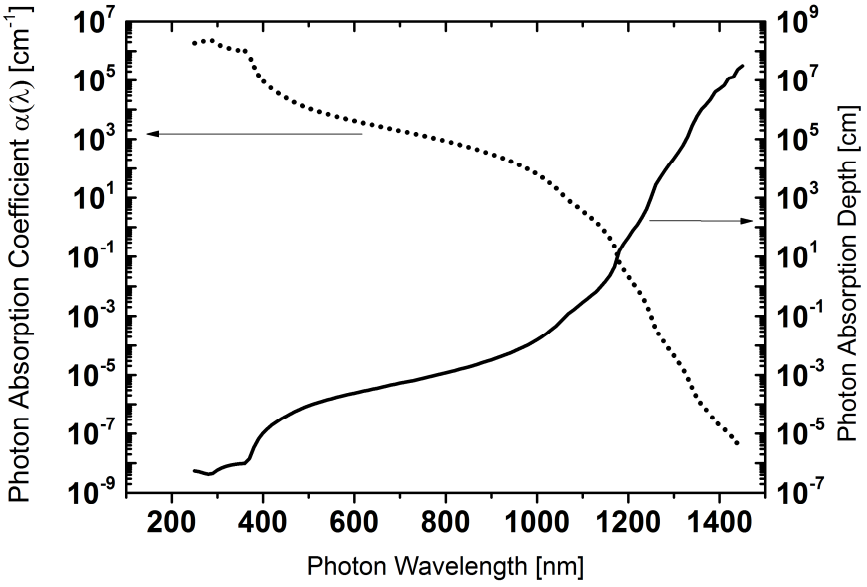

Figure 4: *Absorption coefficient and absorption depth in Silicon, as a function of the photon wavelength.*

Figure 4 [10], shows the photon absorption coefficient in silicon together with its reciprocal value, which is known as photon penetration depth. This latter parameter is probably more intuitive for the understanding of the absorption process of a photon traveling in a semiconductor, since it represents the depth at which a photon is absorbed, on average, with respect to the semiconductor surface. According to this Figure, at short wavelengths, the penetration depth is very small, meaning that a photon entering the semiconductor is absorbed within the first few tens of nanometers. Conversely, at longer wavelengths, the penetration depth becomes larger, meaning that the semiconductor behaves more and more as a transparent optical medium for the incoming radiation.

1.1.2.3 Detection Efficiency for ionizing particles

Moderately relativistic charged particles other than electrons lose energy in matter primarily by ionization and atomic excitation [11]. In many practical cases, most relativistic particles are referred to as minimum ionizing particles (MIP), as they feature a mean energy loss rate close to the minimum value, as described by the Bethe-Bloch equation [11]. It is important to observe that this mathematical description is not really accurate for describing the energy loss by single particles, as it provides a mean energy loss that is weighted by very rare events with large single-collision energy deposits and it is thus subject to large fluctuations [11]. Nevertheless, the mean energy loss of a MIP can still provide a fair estimation of the order of magnitude of the expected amount of electron-hole pairs produced by ionization. In case of silicon, the estimated number of electron-hole

pairs (EHP) produced per unit distance traveled by a relativistic particle is approximately of 80 $EHP/\mu m$ [12]. Therefore, despite of the inherent large fluctuations of the ionization yield, a SPAD device is expected to provide single charged particle detection capability thanks to the extremely high charge multiplication gain provided by the Geiger-mode operation.
As for the PDE calculation discussed in Section 1.1.2.2, it is useful to define a detection-efficiency for a minimum ionizing particle passing through a SPAD pixel. The passage of a MIP through the device active region produces a certain amount of EHPs:

$$N_{EHP} = R_i W_{SPAD}$$

where R_i is the average number of EHPs produced by ionization per unit distance, and W_{SPAD} is the width of the SPAD active region. Each of the N_{EHP} pairs concurs to fire an avalanche multiplication process. However, in order for this to happen, a pair has to be collected towards the space-charge region and eventually has to fire an avalanche process. This occurs with a probability that varies depending on the position where the pair has been generated. For this reason, it is convenient to define, for every EHP, an average avalanche probability P_{av} which is given by the following equation:

$$P_{av} = \frac{1}{W_{SPAD}} \int_0^{W_{SPAD}} n_c(x) P_{tr}(x) dx \tag{8}$$

where $\eta_c(x)$ and $P_{tr}(x)$ are the carrier collection efficiency and avalanche triggering probability, respectively, at a given position x in the SPAD active region. An EHP can successfully fire an avalanche only if no other pair has succeeded before. The whole process can be thus seen as a sequence of attempts where each generated EHP has only one try to successfully trigger an avalanche multiplication process, as follows:

$$P_{MIP} = P_{av} + P_{av}(1 - P_{av}) + .. + P_{av}(1 - P_{av})^{N_{EHP}-1} = P_{av} \sum_0^{N_{EHP}-1} (1 - P_{av})^i$$

$$\begin{aligned} P_{MIP} &= 1 - (1 - P_{av})^{N_{EHP}} \\ &= 1 - \left(1 - \frac{1}{W_{SPAD}} \int_0^{W_{SPAD}} n_c(x) P_{tr}(x) dx \right)^{R_i W_{SPAD}} \end{aligned} \tag{9}$$

An interesting numerical application of this formula is provided in Chapter 4.

1.1.3 Noise: the Dark Count Rate

Geiger-mode avalanche diodes are affected by dark counts, i.e. undesired avalanche pulses occurring in absence of any external radiation stimulus (i.e. light, charges particles, etc.) [1]. In order to allow an efficient detection of the incoming radiation, especially in low intensity condition (i.e. very low photon rate per pixel), it is essential that the quiescent time interval in between two consecutive dark pulses is sufficiently long. For this reason the frequency at which these undesired counts are generated, referred to as Dark Count Rate (DCR), is a very important figure of merit for SPAD devices. DCR is strictly related to the adopted CMOS process, i.e. density of defects and doping doses, layout (i.e. active area dimension and shape), and on external operating conditions (i.e. temperature, hold-off time and excess bias).
Dark Counts can be divided into primary and secondary dark counts [1] based on their physical nature.

1.1.3.1 Primary dark counts: intrinsic dark counts

Primary dark counts are caused by the statistical processes responsible for the random generation of carriers in a semiconductor. An EHP generated randomly in the SPAD active region can indeed undergo the same scenario leading to a detection of a single-photon, as described in Section 1.1.2.2. Due to the statistical nature of the EHP generation mechanism, the resulting sequence of avalanche pulses is randomly distributed over the time, and well described by a Poisson process with rate λ_0 [1].
The physical nature of random EHP generation is the same that is responsible for the reverse current in p-n junctions, and thus mainly caused by the generation phenomena involved in the depletion layer (the contribution of the minority carriers diffusing from the neutral regions to the SCR is negligible [3]). In general, whenever the thermal-equilibrium condition in a region of the semiconductor is disturbed (i.e. $pn \neq n_i^2$, where p, n and n_i are the hole, electron and intrinsic concentration, respectively) there exist generation-recombination mechanisms aiming at restoring the system to equilibrium [13]. In the SCR of a p-n junction, the non-equilibrium condition is such that $pn < n_i^2$, which results in a net carrier generation mechanism [2][13]. In order for an EHP to be generated, a valence electron has to be "promoted" up in the conduction band, leaving a "hole" in the valence band. In Silicon, and more generally in indirect-gap semiconductors, this can occur with the help of localized energy states in the forbidden energy gap, i.e. generation–recombination (G–R) centers. These states are related to the presence of defects and impurities in the semiconductor, and act as "stepping stones" in the transitions between the valence band and the conduction band [13]. In absence of a strong electric field over the junction, the EHP generation process can only rely on the thermal energy of the carriers, as described by the well-known Shockley-Read-Hall (SRH) theory [2]. At higher electric field (above $3 \times 10^5\ V/cm$ [14]), the thermal generation of carriers is combined with electric field assisted generation phenomena, resulting in an important enhance-

ment of the overall generation rate. The electric field can indeed increase the emissivity of G-R centers (i.e. the release probability of a carrier sitting on a localized state) via trap-assisted-tunneling (TAT), or even allow direct tunneling from the valence band to the conduction band (band-to-band tunneling, BBT) [1], [3]. The overall generation process is therefore affected by the amount of impurities and defects in the diode depletion region, but also by the working temperature and the electric field over the junction. On the one hand, it is clear that the resulting DCR is strictly related to the CMOS process quality and its associated technological features, such as the density of defects and impurities but also the doping profiles, respectively. These latter, in particular, have an important influence on the electrostatics of the junction, and play a crucial role in determining whether field-assisted mechanisms give or not an important contribution to the overall carrier generation in Geiger-mode operation. On the other hand, design choices play an important role too and must be carefully defined with respect to the desired performance. The number of dark counts can indeed be reduced by using lower excess bias, smaller active area or by cooling down the detector.
As an useful example, Figure 5 reports the DCR in terms of counts per unit surface as a function of the applied excess bias, obtained from post-processing calculations after TCAD simulation of a SPAD conceived in a 28nm CMOS FDSOI technology [8]. It is possible to distinguish a dominating SRH region for excess bias lower than 4V and a dominating tunneling region for voltages higher than 5V, depending on the strongest EHP generation mechanism.

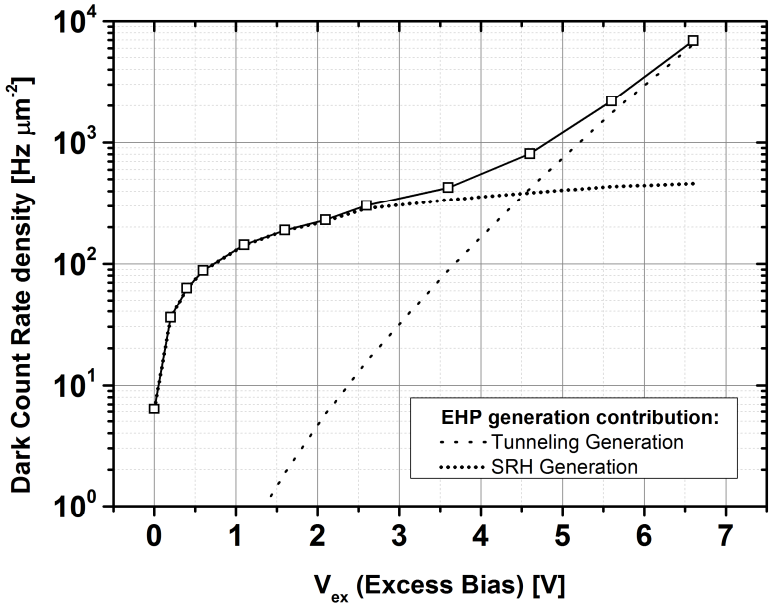

Figure 5: *Total DCR per unit surface resulting from post-processing calculations after TCAD simulation of a SPAD conceived in a 28nm CMOS FDSOI technology [8]. The DCR produced individually by the two EHP generation mechanisms have been plotted too in order to highlight the contribution provided by each of them on the total counts.*

In the SRH region, the DCR is indeed relatively moderate and increases with the excess bias according to the ATP enhancement shown in Figure 3 (symbols, left y-axes). On the other hand, this latter is not really influent in the tunneling region (ATP is larger than 80% and slowly saturates to 100%) where the DCR rises really fast with the excess bias mainly because of field-enhanced carrier generation.

1.1.3.2 Secondary dark counts: after-pulsing

Secondary dark counts (generally referred to as after-pulses) are avalanche events that are correlated to a previously occurred avalanche count. During a breakdown event in a SPAD, part of the large amount of carrier flowing through the p-n junction can be captured by some of the trapping centers in the deep energy level range of the depletion region (i.e. in between the mid-gap and the band edge). These captured carriers are then released after a statistically fluctuating delay which can be considerably long, up to several microseconds. During the hold-off phase, a de-trapped carrier cannot fire an avalanche pulse since the SPAD cannot provide any multiplying capability. If conversely a de-trapping event occurs after the SPAD voltage has been restored above breakdown, the released carrier can retrigger an avalanche process, thus causing a correlated spurious ignition [1]. It is therefore important to stress that this phenomenon is not just a boost of primary dark pulses, since an after-pulse can be correlated to a precedent after-pulse, to a primary dark count, or even to an avalanche pulse related to the detection of a photon. The main consequence of after-pulsing is thus an important enhancement of the observed amount of noise counts, especially in high counting rate applications, where short hold-off times are required. The shorter this time is, the larger is the amount of carriers released when the diode multiplication capability is restored, resulting in a higher probability to observe an after-pulse and thus a noise count. Cooling down the device to reduce primary counts might even increase the overall noise counting rate since the de-trapping process becomes slower at lower temperature. If conversely the counting rate is not crucial, after-pulsing can be mitigated or even totally suppressed by using a long enough hold-off time that covers most of the carrier de-trapping transient after the primary avalanche pulse [1][3]. Reducing the amount of charge produced during an avalanche pulse may be an effective way to mitigate after-pulsing in a SPAD. The probability to observe an after-pulse P_{ap} is indeed proportional to the amount of carriers that are captured during the avalanche cycle N_T^*, and hence to the amount of charge Q_{av} flowing through the junction:

$$P_{ap} \propto N_T^* \propto Q_{av} = C_{out} V_{ex} \tag{10}$$

The avalanche charge can be thus reduced by minimizing as much as possible the overall capacitance on the avalanche diode output node C_{out} (by adopting suitable quenching electronics as discussed in Chapter 2) and by lowering the excess bias V_{ex} above the breakdown voltage of the SPAD. This latter action can reduce after-pulsing probability at the expense of lower detection efficiency.

1.1.3.3 DCR analytical modeling

As discussed in Section 1.1.3.1, the primary dark counts of a SPAD device are randomly distributed over the time but a proper statistical description of the phenomenon can be provided. The time elapsed between the leading edge of two consecutive avalanche pulses can be examined in terms of two temporal components. There is indeed a deterministic time during which the SPAD is arbitrarily kept in the OFF state by the quenching circuit, i.e. the hold-off time t_h (In reality, this latter parameter is well-defined only if an "active recharge" approach is adopted, as in the present work. See Chapter 2 for more details. A more complex mathematical description would be required in case of a passive recharge approach [1]). The second component is, conversely, a stochastic time which is strictly related to the generation mechanisms in the diode SCR and the avalanche triggering probability of a generated EHP. This latter component is well statistically described by a Poisson distribution $p_d(t)$ with rate λ_0:

$$p_d(t) = \lambda_0 e^{-\lambda_0 t} \tag{11}$$

The average waiting time before two consecutive events is thus given by:

$$< t_0 >= t_h + \int_0^\infty t\, p_d(t) dt = t_h + 1/\lambda_0$$

The observed primary dark count rate, i.e. the DCR in absence of after-pulsing effect, is finally obtained as:

$$DCR_0^{\text{obs}} = \frac{1}{< t_0 >} = \frac{\lambda_0}{1 + \lambda_0 t_h} \tag{12}$$

It is important to stress that the above formula describes the dark count rate effectively observed at the output of the device, as (12) is dependent on the adopted hold-off time t_h. That's why, for a correct SPAD characterization, it is always necessary to extract the intrinsic rate λ_0, which is only dependent on the device parameters such as the EHP generation mechanisms, the avalanche triggering probability, geometry etc. Moreover this formula shows that SPAD devices have in practice a maximum achievable counting rate, which is limited by the minimum attainable hold-off time, i.e. $f_{max} = 1/t_h$.

A different analysis is necessary for the study of after-pulsing. The after-pulsing probability can be defined as follows:

$$p_{ap}(t)dt = f(t)dt\left(1 - P_{ap}(t)\right)$$

The probability to have a correlated pulse within t and $t+dt$, defined as $p_{ap}(t)dt$, is given by the probability that a carrier is released by one of the traps within t and $t+dt$ and successfully initiate an avalanche process, $f(t)dt$, conditioned by the probability that no after-pulse has occurred before, i.e. $\left(1-P_{ap}(t)\right)$ where $p_{ap}(t)=\partial P_{ap}(t)/\partial t$. The solution of the above differential equation gives the after-pulsing probability for a SPAD:

$$P_{ap}(t)=1-e^{-\int_0^t f(t')dt'} \tag{13}$$

It is therefore possible to define an overall avalanche probability, accounting for the fact that an observed avalanche event can be either due to a primary count or to an after-pulse, but not to both:

$$P_{av}(t)=P_{ap}(t)+P_d(t)-P_{ap}(t)P_d(t)$$

where $P_d(t)=\int_0^t p_d(t)dt=1-e^{-\lambda_0 t}$ is the primary counts probability. The resulting statistical process is a Poisson distribution with a combined time-dependent rate parameter $\lambda(t)=\lambda_0+f(t)$ (in agreement with [15]):

$$p_{av}(t)=\big(\lambda_0+f(t)\big)e^{-\int_0^t\left(\lambda_0+f(t')\right)dt'}=\lambda(t)e^{-\int_0^t\lambda(t')dt'} \tag{14}$$

Similarly to the discussion made initially for the primary counts, the average waiting time before two consecutive events in presence of after-pulsing is given by (obtained by means of integration by parts):

$$<t>\triangleq t_h+\int_0^\infty t\lambda(t)e^{-\int_0^t\lambda(t')dt'}dt=t_h+\frac{1-\int_0^\infty e^{-\lambda_0 t}p_{ap}(t)dt}{\lambda_0}$$

The intrinsic DCR (i.e. not dependent on the hold-off time) is thus obtained from the second term in the equation above, as reported here below:

$$\lambda^*=\frac{\lambda_0}{1-\int_0^\infty e^{-\lambda_0 t}p_{ap}(t)dt}$$

In practical cases, the probability that a primary count occur within the time interval where after-pulsing effect is intense, is absolutely negligible. The above formula can be thus re-written in a simpler and more convenient form:

$$\lambda^*=\frac{\lambda_0}{1-\int_0^\infty e^{-\lambda_0 t}p_{ap}(t)dt}\approx\frac{\lambda_0}{1-\int_0^\infty p_{ap}(t)dt}=\frac{\lambda_0}{1-P_{ap}} \tag{15}$$

where P_{ap} is defined as the SPAD overall after-pulsing probability, i.e. $P_{ap} = \int_0^\infty p_{ap}(t)dt$. This effect can thus dramatically increase the overall noise count rate of a SPAD. Interestingly, according to what it can be found in literature, equation (15) is typically derived in a different way. The after-pulsing effect is indeed modeled as a chain of secondary events with the same probability P_{ap}, produced by a primary event, resulting in a geometric series [16][17]:

$$N_{noise} = N_{primary} + N_{primary}\left(P_{ap} + P_{ap}^2 + \cdots\right) = \sum_{i=0}^{\infty} N_{primary} P_{ap}^i = \frac{N_{primary}}{1 - P_{ap}}$$

According to the discussion made in section 1.1.3.2, an after-pulse can be correlated to any kind of previous avalanche event, such as an after-pulse, a primary dark count, or even to an avalanche pulse related to the detection of a photon. Therefore if the SPAD is shined with light, the overall amount of noise counting rate might be dramatically enhanced, depending on the intensity of the after-pulsing effect. The photon arrival time on the pixel can be modeled with a Poisson process with rate λ_{ph}, similarly as for the primary dark counts, but the rate of avalanches produced after a photon detection has to account for the SPAD Photon Detection Efficiency (PDE), i.e. the photon detection rate is $\lambda_{ph}PDE$. This gives rise to the following overall counting rate:

$$\lambda_{tot}^* = \frac{\lambda_0 + \lambda_{ph}PDE}{1 - P_{ap}}$$

The noise counting rate is, in this case, given by the overall measured counting rate λ_{tot}^*, minus the detected photon counting rate, as follows:

$$\lambda_{tot}^* - \lambda_{ph}PDE = \frac{\lambda_0 + PDE\ \lambda_{ph}P_{ap}}{1 - P_{ap}} = \lambda^* + \frac{P_{ap}}{1 - P_{ap}}\lambda_{ph}PDE \qquad (16)$$

The overall amount of noise counting rate is thus enhanced by the additional term on the right-hand side of equation (16). The enhancement is as strong as the after-pulsing probability is large.

1.1.4 Cross-talk

In case of devices consisting of arrays of SPAD pixels, cross-talk represents an additional correlated source of spurious counts affecting the DCR of a SPAD. An avalanche pulse fired in a given pixel may indeed be correlated to a breakdown process occurred, right before, in one of the surrounding neighboring pixels. Depending on the physical mechanism leading to the correlated pulse, it is possible to distinguish between electrical and optical cross-talk (Figure 6).

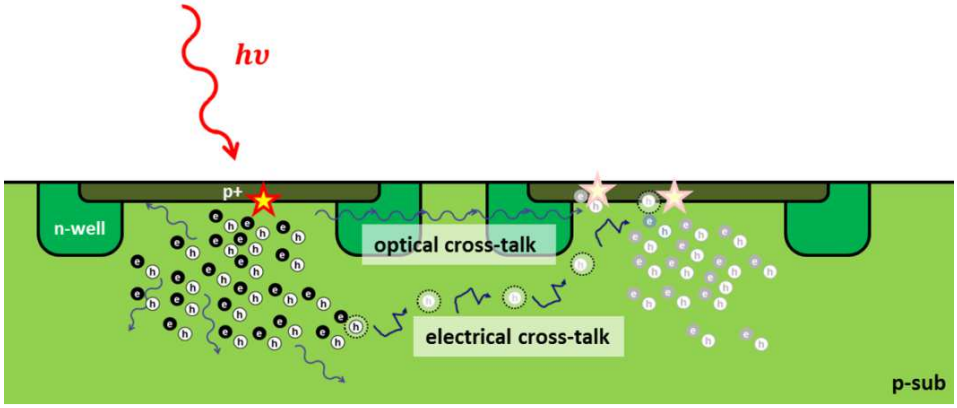

Figure 6: *Qualitative representation of the cross-talk mechanisms occurring I SPAD arrays.*

1.1.4.1 Optical cross-talk

When a diode is biased above the breakdown voltage, an intense emission of photons in the visible spectrum range is observed[2]. If one (or more) of them is re-absorbed within the active area of a neighboring pixel, an avalanche process can be triggered there too, producing a spurious pulse correlated to the primary avalanche. This undesired effect is strictly related to the distance between neighboring pixels and to the amount of charge produced during an avalanche process. Optical crosstalk can be thus mitigated by properly reducing the avalanche current (with the effect of mitigating after-pulsing too) and, more effectively, by surrounding every pixel with suitable absorption media such as highly doped isolation diffusions [3] or deep trenches [18].

1.1.4.2 Electrical cross-talk

The EHP generated during an avalanche multiplication process, are not exclusively subject to a drift transport within the SPAD multiplying region. Due to the large amount of carriers produced by impact ionization within a very small volume (i.e. the SCR of the SPAD) a strong diffusive transport component in all directions can set up. In case of an array implementation of SPADs, if the pixels share the same substrate, part of the generated carrier might diffuse up to a neighboring pixel and fire an avalanche process, producing in this way a correlated spurious pulse. For this reason SPAD pixels are typically electrically insulated by placing them into deep diffusions of opposite doping type with respect to the substrate. However this improvement in terms of noise performance is obtained at the expense of a lower fill factor. [19]

[2] Since the discovery of the phenomenon in 1955, the physical nature responsible for this effect is still subject to controversy [44][45].

1.1.5 Time-jitter

There are applications where it is important to determine with high accuracy the arrival time of the impinging radiation, for instance in Time Correlated Single-Photon Counting (TCSPC) [20]. As discussed in Section 1.1.1, the avalanche leading edge is indeed synchronous with the photon arrival time but the detection delay is unavoidably affected by a certain statistical fluctuation (jitter), which may seriously limit the temporal resolution of the SPAD. The jitter is typically measured in terms of the full-width at half maximum (FWHM) of the time distribution of arrival times obtained through a repetitive collection of fast laser pulses [3] (Figure 7). If the time-resolution is not limited by the electronics, the lowest attainable jitter limit in a SPAD device is set by the physical phenomena involved in the detection process of a single-photon. If a photon is absorbed in the SCR of a SPAD, the generated EHP would be promptly accelerated by the local electric field, producing an avalanche current almost immediately. The detection process would be anyway affected by some time uncertainty responsible for the peak in the curve shown in Figure 7. The time-jitter in this case is due to the statistical nature of the avalanche build-up, especially due to the physics related to the lateral propagation of the multiplication process. The lateral propagation can indeed occur via a multiplication-assisted diffusion [21] or via photon-assisted multiplication [22]. However, depending on the specific device architecture, this contribution can be as small as a few tens of picoseconds. Completely different is the case where a photon is absorbed into one of the two neutral regions. Due to the absence of a strong electric field, a photo-generated minority carrier diffuses randomly through the neutral region for a relatively long time before it possibly reaches the multiplying region and finally fires an avalanche. These photons are therefore responsible for the right-hand side of Figure 7, commonly referred to as “diffusion tail”.

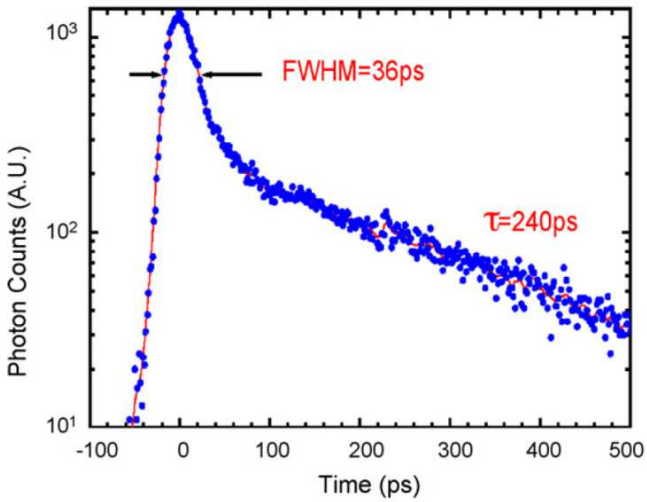

Figure 7: *Example of a typical time-response of a SPAD. The histogram has been obtained by irradiating the device with a short laser pulse at 820-nm wavelength. Picture is taken from reference* [3].

1.1.6 SPAD State of the Art

The development of SPADs has been driven, since their early stages, by the requirements demanded in the manifold applications where single-photon sensitivity is necessary to study a certain physical phenomenon. SPADs have been indeed widely employed in photon-counting and photon-timing applications, such as laser (LIDAR/LADAR) [23] and 3-D optical ranging [24], but also in fluorescence lifetime imaging (FLIM) [25], positron emission tomography (PET) [26], and Time-Correlated Single-Photon Counting (TCSPC). Since their early implementations up to about fifteen years ago, SPADs were exclusively fabricated by means of fully custom silicon processes, achieving best-in-class performance in terms of noise, time-jitter and detection efficiency [17]. Custom process provides indeed full control over the fabrication steps which made possible to achieve excellent noise performance and good uniformity thanks to dedicated annealing and gettering steps that have the beneficial effects of minimizing the density of lattice defects and impurities. The first CMOS SPAD implementation appeared only recently, in the early 2000s [27], opening the way for a large scale production of low-cost and versatile photon counting systems such as monolithic detector arrays incorporating sensors and associated read-out electronics.

SPAD devices can be generally regrouped into two main categories, depending on the topology of the depletion region in the p-n junction which can be thin or thick [1]. The main difference arises from the extension of the drift region, as the multiplication region (i.e. the region where avalanche multiplication takes place) tends to be the same in both cases. The two SPAD categories are discussed in more details in the following sections.

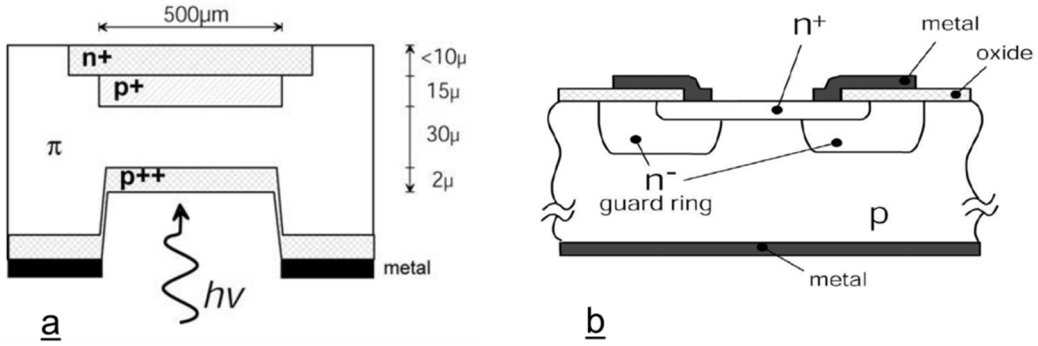

Figure 8: *Schematic cross-section of (a) an avalanche photodiode with reach-through structure developed by McIntyre and Webb; (b) a planar p-n diode developed by Haitz et al. to investigate avalanche behavior above breakdown. Figures are from reference* [28].

1.1.6.1 Reach-through SPAD

The first "reach-through SPAD" (Figure 8a), which is also known as "thick SPAD" because of the wide extension of the active region, was developed by McIntyre and Webb developed in the 70's [28]. These devices have indeed a deep vertical structure consisting of a p-n junction defining the high-field region, and a fully depleted thick drift region of intrinsic silicon responsible for collecting all the charge injected into the device. Thanks to the wide depletion region, thick SPADs show very high photon detection efficiency (PDE) in the visible region, which can be larger than 50% within the range from around 550 nm up to 850 nm wavelength. The efficiency decreases rapidly in the near-infrared, even if it can be on the order of a few % at 1064 nm [1]. Thanks to dedicated annealing and gettering steps provided by the custom processes, it is possible to achieve excellent noise performance at room temperature, allowing very large area detectors of several hundred of μm in diameter [1]. However the deep drift intrinsic region limits the time-resolution of thick SPADs to values on the order of 300-800 ps [1]. As stated before, thick SPADs are fabricated by means of low yield and costly silicon processes not compatible with technologies employed for the fabrication of integrated circuits. Moreover they are affected by high power dissipation that may require strong cooling. Pulse-peak power can indeed reach values up to 10W, due to the high voltage bias required for Geiger-mode operation (300 – 500 V) and the high avalanche current (few tens of mA) [3]. For all these reasons, "reach-though SPADs" are certainly unsuitable for low-cost and versatile photon counting devices as there is no feasible perspective of integrating them in compact full detection systems incorporating both detectors and associated read-out circuits [3].

2.1.6.2 Planar SPAD

In parallel to the development of the "reach-through" SPAD, a different SPAD architecture referred to as "planar SPAD" or "thin SPAD" was proposed in the 60's by Haitz [28]. These two different names actually indicate, respectively, the compatibility with planar silicon processes (custom or not) and the thin extension of the depletion layer of the p-n junction, ranging from a few hundreds of nanometers up to a few micrometers (Figure 8b) [28]. A planar SPAD is commonly defined by a p-n junction consisting of a doped diffusion implant over a substrate having an opposite doping type. In order to prevent premature breakdown at the junction edge (PEB), the diffusion is normally surrounded by a guard-ring, i.e. a low-doped well of the same doping type of the diffusion with the aim of lowering the electric field all around the device periphery. There are of course other possible solutions that can be adopted in order to realize a thin SPAD device topology. Since the early stages of SPADs, researchers have been looking for architectures providing PEB prevention as well as a planar and uniform electric field distribution all over the active region. Moreover, depending on the specific application, a special effort has been addressed in order to ensure a fairly good PDP as well as optimal SPAD timing performance.

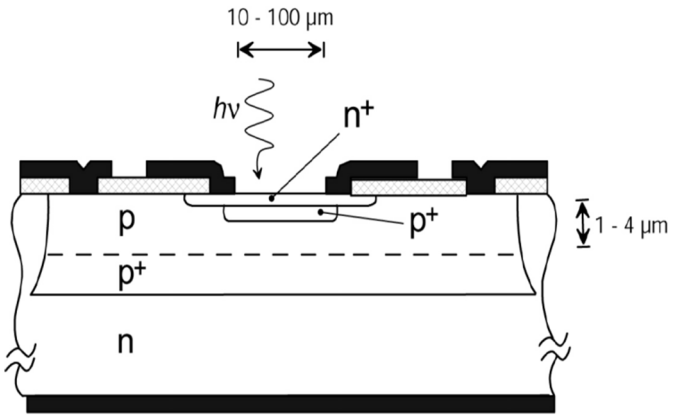

Figure 9: *Schematic cross-section of a double epitaxial silicon SPAD introduced by Lacaita et al. The figure is taken from* [28]**.**

For instance, in the late 80's, Lacaita et al. [29] introduced a novel planar SPAD architecture, referred to as "double epitaxial silicon SPAD", with the aim of reducing the diffusion tail degrading the time-resolution of the device (Figure 9). In this approach, an active n+p junction is built in a low-doped p-epi layer grown over a buried highly doped one that provides a low resistance path to the anode contact. This latter is also responsible for reducing the diffusion tail, since it forms, together with the n-type substrate, an additional p-n junction preventing carriers photo-generated in substrate from reaching the avalanche region. Interestingly PEB is avoided thanks to a "virtual" guard-ring strategy, by means of a p-type enrichment in the central part of the device, to locally increase the electric field. In general, "planar SPAD" devices allow a wide range of active area diameters, from about 10 µm, up to 500 µm [30], [31] and time-resolution can be very good, i.e. lower than 100 ps (provided that the SPAD diameter is not larger than 50 µm). In particular, devices with a small active area (around 10 µm diameter) can attain better than 30 ps at room temperature and better than 20 ps when cooled to 265 K [1], [30]. In this case, the photon detection efficiency is not as good as for a thick-SPAD device. The maximal efficiency value can be larger than 40% only around a photon wavelength of $500\,nm$ [1]. The efficiency drops indeed towards lower values at higher wavelength, even if it is still on the order of 35%, at $650\,nm$, 15% at $750\,nm$ and about 10% at $850\,nm$ [1]. More importantly planar SPADs feature a breakdown voltage on the order of a few tens of volts, which, together with their compatibility with planar processes, facilitate the integration of the device in CMOS technologies.

1.1.6.3 CMOS SPAD

Standard CMOS processes enable SPAD monolithic integration with readout electronics for the design of low-cost and high-performance fully integrated imaging systems with single-photon detection capability. The first SPAD pixel in CMOS technology has been demonstrated only in 2002 by A. Rochas et al. [27] but, since then, further developments rapidly followed towards submicron

CMOS technologies with minimum feature size down to 65nm [32]. The main issue with standard CMOS processes is that they are optimized for the fabrication of integrated circuits (ICs) where the primary objective is to pursue the Moore's law, by reducing the transistor feature size. This naturally translates into severe design constraints as well as performance limitations for SPAD detectors. There are indeed some major design requirements that have to be addressed in the development of CMOS compatible SPADs. Among them, it is worth mentioning: the choice of a sufficiently clean CMOS fabrication process that minimizes the density of defects and impurities in order to obtain very low DCR and after-pulsing probability; the design of a "PEB-safe" SPAD architecture ensuring an uniform electric field distribution, i.e. a uniform breakdown probability, all over the active region; a moderate electric field that avoid field-enhanced generation phenomena in order to reduce DCR as much as possible. In order to pursue these objective, within the last fifteen years, many groups worldwide developed different SPAD architectures conceived specifically for the adopted CMOS technology [17]. The state of the art of CMOS SPAD architectures can be grouped based on the guard ring implementation, as shown in Figure 10.

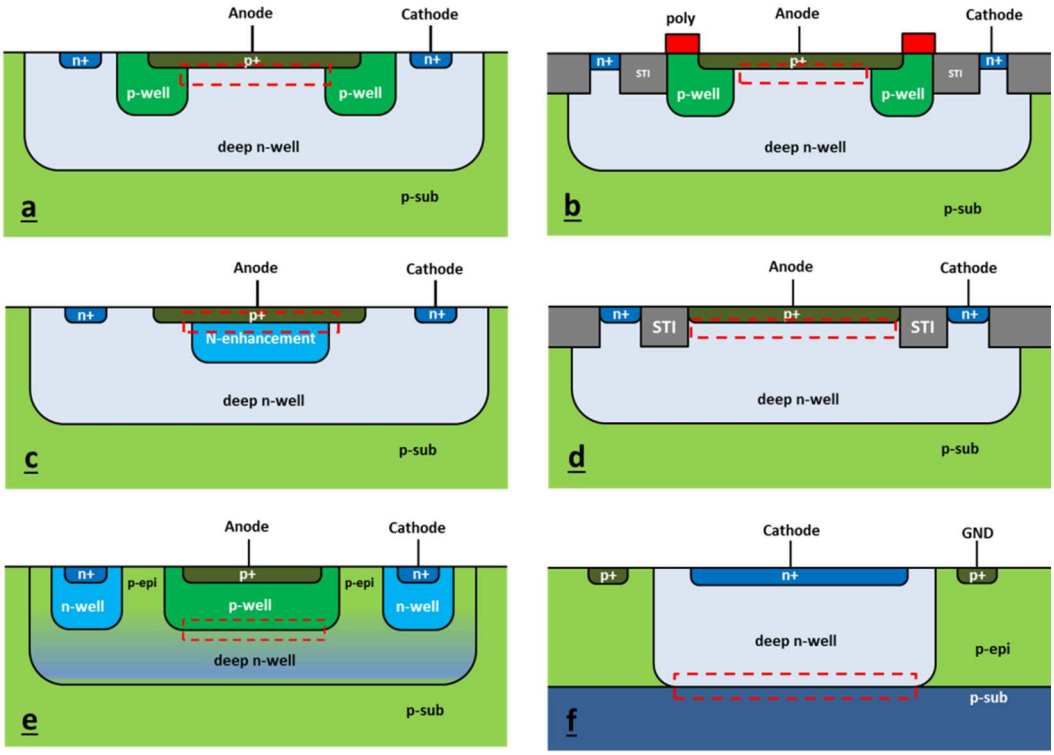

Figure 10: *Cross sections of different CMOS SPAD architectures among those found in literature. The multiplication region is highlighted with a dashed rectangle. (a) Diffused Guard Ring; (b) STI-free diffused Guard Ring; (c)"Virtual" Guard Ring; (d) Shallow Trench Isolation (STI) Guard Ring; (e) retrograde-doping guard ring; (f) buried multiplying region.*

The most straightforward implementation of a SPAD in a standard CMOS process reproduces the planar architecture proposed by Haitz (Figure 8b) in the 60's. The junction is in this case obtained with an n+ diffusion layer (commonly used for the source/drain regions of n-type channel MOS transistors) over the common p-type substrate, while the guard-ring is provided by surrounding the junction with a standard n-well (normally used as bulk region for the formation of p-type channel MOS transistors). Further improvements can however be achieved if the CMOS process features a deep n-well, i.e. a deep n-type tub that is available in advanced CMOS technologies in order to insulate both p-channel and n-channel MOS transistors from the substrate. As depicted in Figure 10a, in this case the active region consists of a p+/n junction, obtained by means of a p+ diffusion over the deep n-well, while a low doped p-well surrounding the junction forms the guard ring. Apart from insulating the SPAD from substrate noise, the deep n-well prevents free carriers generated in the substrate to diffuse into the junction (improving the SPAD timing-performance, similarly to the "double-epitaxial SPAD" of Figure 9) and allow biasing the anode and cathode independently from the substrate. This architecture was implemented for the first time in 2002 by Rochas et al. [27] using a 0.8μm high-voltage standard CMOS process provided by AustriaMicroSystem, and afterwards also in more advanced sub-micron CMOS processes [32]. However, from the 250nm node, standard CMOS processes feature shallow trenches filled with SiO_2, commonly referred to as Shallow Trench Isolations (STI), etched all around p+ and n+ implantation areas in order to effectively prevent punch through and latch-up effects in CMOS circuits. Therefore STI would sit right next to the SPAD active region which may increase the device DCR to very prohibitive values (up to a few MHz [33]) if a large amount of defects is present at the Silicon/Oxide interface. A possible way to deal with this technological drawback consists of drawing "dummy" poly-silicon gate of a standard transistor in the region surrounding the p+ diffusion, i.e. where the STI has to be moved away, as proposed by Niclass et al. [34]. In order to prevent a high-electric field in the thin oxide layer below the poly-silicon gate, this one is kept at the same potential as the p+ anode. Moreover, it is interesting to observe that the p-well guard-ring would act, in this case, as a passivation layer for the STI interface, which can probably mitigate the DCR enhancement. This observation was exploited in the work of Gersbach et al. [35], by surrounding the STI with several passivation implants, in a glove-like p-type structure. At the STI interface, the doping level is high, which results in a very short mean free path for the minority carriers generated at the Si/SiO_2 interface. As the distance from this surface increases, the doping concentration is reduced in order to lower the electric field at the p+ edge, thus preventing PEB. Furthermore, an important limitation of the diffused guard-ring approach arises in case of a SPAD array implementation. The scalability of the active region would be indeed limited by the guard-ring depletion region whose inner edge would merge when scaling down the SPAD active area diameter. A CMOS implementation of the "virtual" guard-ring approach (Figure 10c) adopted for the "double epitaxial SPAD" (see Figure 9) has been implemented for the first time in a standard CMOS process by L. Pancheri et al. [36] and can be a possible solution for an improved SPAD scalability. This architecture does not suffer

from depletion region merging effect as there is no actual physical guard-ring surrounding the junction anymore. As discussed previously, the n-type enrichment over the p+ diffusion is indeed enough to provide an electric field enhancement in the desired location. Moreover, for deep sub-micrometer CMOS processes, i.e. from the 250 nm node, the STI can be easily kept away from the high field region by adopting a sufficiently large diameter for the p+ diffusion while keeping the same active region extension [37] (at the expense of a lower fill factor). A further improvement in the SPAD scaling can be achieved by exploiting the isolation trenches in a deep sub-micrometer CMOS processes as a guard-ring, as shown in Figure 10d, assuming a good interface quality in order to avoid any DCR increase. This solution can provide great scalability thanks to an effective physical electric field confinement due to the isolation trench surrounding the multiplying region. Moreover an excellent fill-factor can be obtained since the dielectric strength of SiO_2 is 30 times higher than the breakdown field of silicon which allows a 30 times narrower STI guard-ring with respect to a diffused one. However all the architectures described so far present an active region made of a p+ diffusion / n-well junction whose doping levels become unfortunately excessively high when adopting modern CMOS processes. The high doping concentration levels are indeed required by the scaling constraints for MOS transistors in deep sub-micrometer CMOS technologies. However they cause a very narrow depletion region and a very high electric field in the junctions, which translates into a significant field-assisted carrier generation and thus a dramatic DCR enhancement. More advanced SPAD architectures have been therefore conceived in order to mitigate field-assisted generation in deep sub-micrometer technologies, by building the multiplying region with lighter doping layers. Figures 10e and 10f show two SPAD architectures implementing a “virtual” guard-ring approach, where the active region has been displaced at a deeper location with respect to the previous structures, i.e. where doping concentrations are lower with respect to p+ or n+ diffusion in the surface. In solution (e) the active region is defined by a p-well / buried n-well junction. If no n-well is drawn together with deep n-well (and provided that p-well formation is explicitly prevented) the result is indeed a “buried n-well” at a certain depth, with n-type doping concentration progressively diminishing towards the upper surface of the SPAD. Therefore the p-well can be seen as a p-type doping enrichment in the retrograde n-type doping, which enhances the electric field in a well localized region, as for the “virtual” guard-ring case. The buried n-well is naturally connected to the cathode contacts by means of n+/n-well layers drawn at the outer edge. This architecture has been demonstrated by Richardson et al. [37] in a 130nm CMOS image sensor technology, achieving a DCR of only 40 Hz for a 8 μm diameter SPAD. Another solution, depicted schematically in Figure 10f, has been implemented for the first time by Webster et al. [38] in a 90 nm and a 130nm [39] CMOS imaging technologies, achieving respectively, a DCR of only 100 Hz and 18 Hz, and an excellent scalability allowing an active region diameter as small as 6.4 μm and 8 μm. A well localized and PEB-free high-field region is, in this case, defined by the junction between a deep n-well and a thin p-epitaxy on a low resistivity p-substrate, resulting in a “virtual” guard-ring solution as for architecture (e).

1.2 A novel detector: the 3D Silicon Coincidence Avalanche Detector

The "3D Silicon Coincidence Avalanche Detector" (3D-SiCAD), also referred to as Avalanche Pixel Sensor (APiX) [40] [41], is a novel device suitable for the detection of high energy charged particles. A 3D-SiCAD pixel consists of a pair of vertically aligned Geiger-mode avalanche diodes, which are electrically connected by means of 3D integration techniques. As represented schematically in Figure 11, the passage of a charged particle through a 3D-SiCAD pixel fires a breakdown process in both avalanche diodes, producing two coincident avalanche pulses.

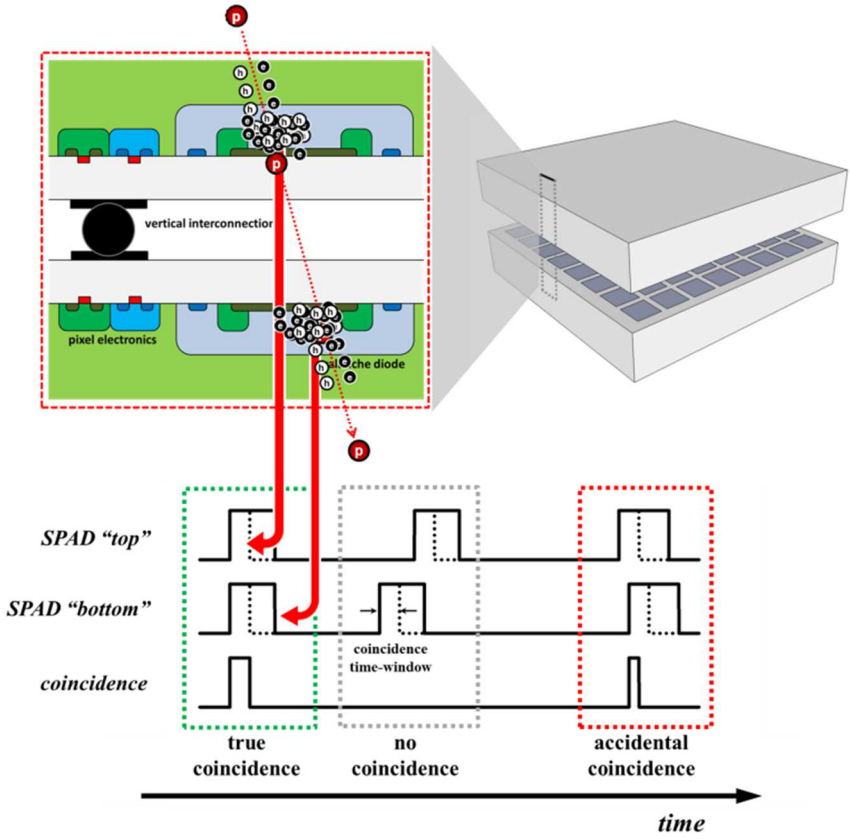

Figure 11: Schematic representation of a 3D-SiCAD pixel considering a (qualitative) possible implementation in a CMOS process by adopting a stud-bump vertical interconnection. A qualitative time-diagram representing the main waveforms in a 3D-SiCAD pixel is reported in the bottom part of the picture. Solid line waveforms associated to the "top" and "bottom" SPADs represent the output signals produced by the quenching electronics in each sensing level. The dotted lines represent the ultra-short pulses that are synchronous to the leading edge of an avalanche event. The width of the synchronous pulses represents the coincidence time-window.

Dedicated coincidence electronics thus determine whether top and bottom layers are simultaneously activated, allowing to distinguish the detection of a charged particle from false and random avalanche events occurring in either the first or second avalanche diode due to background photons (in the Ultra-Violet, Visible and Near-Infrared ranges) and dark counts (see Section 1.1). A 3D-SiCAD pixel can thus provide an excellent noise rejection capability with respect to a simple SPAD device. A background photon cannot indeed provide any coincidence event since an avalanche pulse can only occur in one sensing level, i.e. only where the photon is absorbed. Similarly, dark counts occurring in a sensing level are statistically uncorrelated to the dark counts occurring in the other one. Moreover, despite of the small thickness of the SPAD active region and the inherent fluctuations of the ionization yield, a 3D-SiCAD pixel is able to provide single charged particle detection capability thanks to the high intrinsic gain, inherent of the Geiger-mode operation. Since the building blocks of this novel device are SPAD pixels, a 3D-SiCAD detector can be certainly realized in a standard CMOS process. The electrical interconnection between the two dies can be obtained by means of commercially available 3D integration techniques. The main figures of merit for this novel device can be deduced from those of SPADs, as these latter represent an important building block for a 3D-SiCAD pixel.

1.2.1 Noise in 3D-SiCAD devices

In a similar way as for SPAD detectors, the noise in 3D-SiCADs is defined in terms of spurious coincidence counts that are not related to any particle hit. False counts in dark condition may be produced when two random dark pulses from the two sensing levels occur within the time window required for the coincidence check. Such a time window is actually limited by the shortest time that the electronics is capable to resolve. Therefore, an important figure for a 3D-SiCAD is the width of the coincidence time-window, as it is responsible for suppressing the probability to have a fake count during the coincidence check. However, the resolving time cannot be smaller than the time jitter intrinsically affecting the detector response to an incoming particle, since this would inevitably lead to true coincidences losses. The rate at which these false coincidences are expected to occur in a 3D-SiCAD is referred to as Fake Coincidence Rate (FCR) and is given by equation (17):

$$FCR = 2 \cdot DCR_{top} \cdot DCR_{bottom} \cdot \Delta t \tag{17}$$

where DCR_{top}, DCR_{bottom} refers to the "observed" dark count rate (see Section 1.1) in the top and bottom SPAD device, respectively, while Δt is the coincidence time window. For every dark count occurring in the top level there is indeed a probability $P_{dark,bottom} = DCR_{bottom} \cdot \Delta t$ to have a "dark" avalanche in the bottom one, and vice-versa.

1.2.2 Minimum Ionizing Particles (MIP) Detection Probability

Minimum Ionizing Particles (MIP) detection probability can be defined in a similar way as for SPAD detectors, as the probability that a single minimum ionizing particle hitting a 3D-SiCAD pixel is successfully detected. This can occur only if:

- the particle crosses the two active areas of the aligned SPAD pixels;
- the charged particle effectively ionizes silicon atoms producing EHP along the portion of the particle path falling within the sensitive volumes;
- in both SPAD pixels, the generated EHPs successfully fire an avalanche multiplication process.

These considerations can be translated into equation (18):

$$P_{MIP} = P_{MIP,SPAD_{top}} P_{MIP,SPAD_{bottom}} \eta_\alpha FF \tag{18}$$

where $P_{MIP,SPAD_{top}}$, $P_{MIP,SPAD_{bottom}}$ are the MIP detection efficiencies for the two SPAD sensing levels, FF is the fill factor for each SPAD pixel and η_α is the angular efficiency. This latter parameter accounts for the fact that only a portion of the pixel surface is sensitive to the incoming particle, depending on the angle at which the particle hits the 3D-SiCAD pixel as depicted in Figure 12.
Assuming, for instance, a circular geometry for the avalanche diode, the sensitive area as a function of the incident radiation angle is given by (19):

$$A_\alpha = 2R_0^2 \left(\frac{\pi}{2} - arcsin\left(\frac{d}{2R_0 \tan\alpha}\right) - \frac{d}{2R_0 \tan\alpha} \sqrt{1 - \left(\frac{d}{2R_0 \tan\alpha}\right)^2} \right) \tag{19}$$

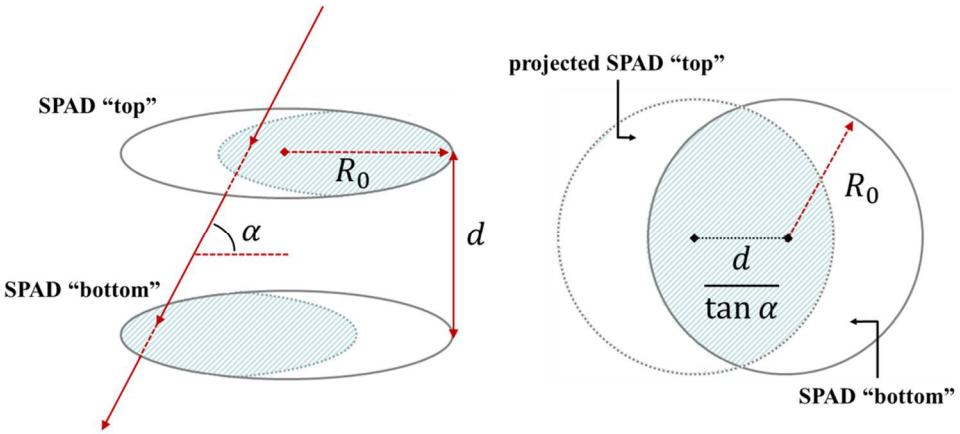

Figure 12: *Geometrical model of a 3D-SiCAD pixel, assuming a circular geometry for the avalanche diodes. Only a portion of the pixel surface is sensitive to the incoming particle, depending on the angle of incidence of the particle.*

where R_0 is the radius of the active area in each SPAD, d is the distance between the two sensing levels of the 3D-SiCAD pixel, and α is the incidence angle of the radiation, with respect to the pixel surface. The angular efficiency can be defined according to equation (20), as the ratio between (19) and the SPAD active area $A_{SPAD} = \pi R_0^2$:

$$\eta_\alpha = 1 - \frac{2}{\pi}\left(arcsin\left[\left(\frac{d}{2R_0}\right)\frac{1}{tan\,\alpha}\right] - \left(\frac{d}{2R_0}\right)\frac{1}{tan\,\alpha}\sqrt{1-\left[\left(\frac{d}{2R_0}\right)\frac{1}{tan\,\alpha}\right]^2}\right) \tag{20}$$

Figure 13 plots the angular efficiency as a function of the incidence angle α, for different $d/2R_0$ ratios. Observe that if the radiation hits the pixel perpendicularly on its surface, the angular efficiency is one, as expected. Conversely, there exists a cut-off angle defined as $\alpha_{cut-off}$, below which the efficiency suddenly drops to zero:

$$\alpha_{cut-off} = \arctan\left(\frac{d}{2R_0}\right) \tag{21}$$

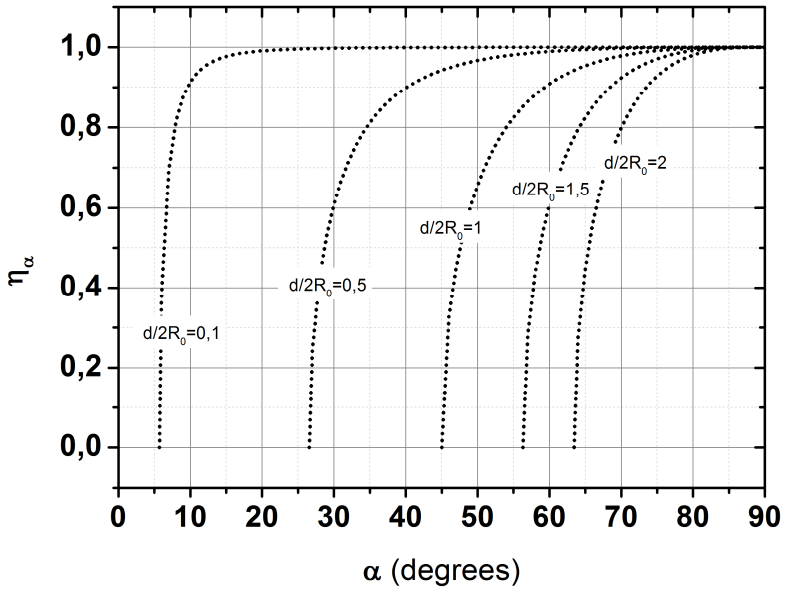

Figure 13: *Angular efficiency as a function of the radiation incidence angle, for several values of the* $d/2R_0$ *ratio.*

1.2.3 3D-SiCAD State of the Art

The 3D-SiCAD concept has been proposed for the first time by V. Saveliev in 2012 (US patent [40]) where the device has actually been referred to as "Avalanche Pixel Sensor" (APiX) detector. This concept has been further studied by N. D'Ascenzo et al. [41]) where a first APiX proof of concept was provided by coupling together in a "face-to-face" configuration a pair of Silicon Photo-Multiplier (SiPM) sensors. The APiX prototype was placed between two scintillator counters and the whole architecture was exposed to a 120 GeV proton test beam in the CERN North Hall area as depicted in Figure 14. In this way, a detected count in both sensing devices could be interpreted as a correct detection of an ionizing particle. Therefore the experiment consisted of comparing the coincidence rate of the two SiPMs stack with the triggering rate of the monitor scintillators counters.

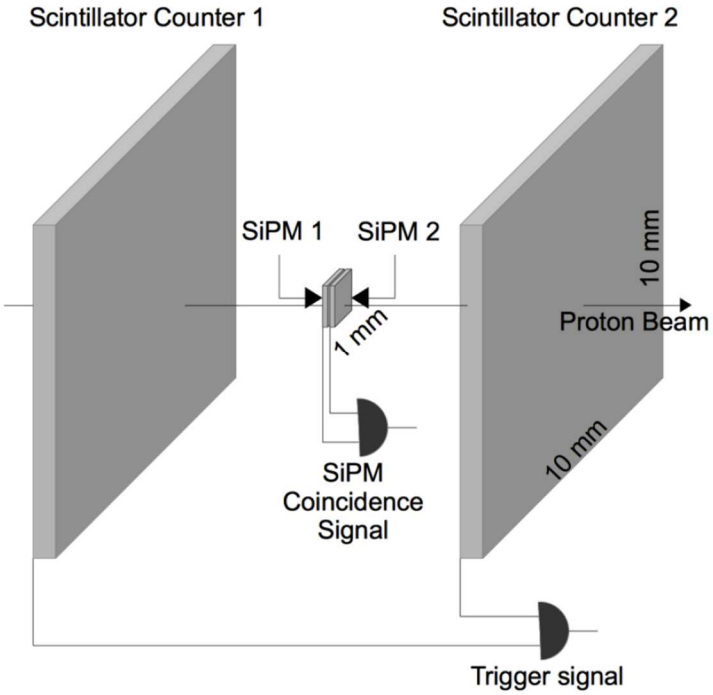

Figure 14: *Schematic representation of the test-beam setup for the proof of concept of the 3D-SiCAD detector (from* [41] *)*

Even if the alignment of the two SiPMs was not well controlled in the setup (which certainly affected the measurement accuracy) the authors obtained a rather reasonable result in agreement with expectations. The ratio between the coincidence rate of the SiPM stack and the monitor scintillator counters, was found to be proportional to the ratio between their respective sensitive areas.

Quite recently, we discovered that a similar work is being developed in Italy by L. Pancheri et al. [42] within a project funded by the Istituto Nazionale di Fisica Nucleare (INFN) involving different groups from Trento, Pisa, Siena and Pavia.

In this manuscript, we will show that we have obtained very promising results even if this research work was carried out with quite limited resources.

Conclusions

A complete overview on Single Photon Avalanche Diodes (SPAD), i.e. the building block of a 3D-SiCAD device, has been addressed in the first part of this chapter. The physics and the corresponding mathematical description of the processes behind the detection of a single photon and a single charged particle have been discussed in detail. The physical processes responsible for the noise counts in SPAD devices have been then discussed together with the related mathematical modeling required for a correct data analysis of the experimental results. Other important figures of merits such as the time-resolution (or time-jitter) and cross-talk between pixels in case of matrix implementation have been further discussed. A complete overview about the state-of-the art for the SPAD devices has been finally proposed, with a special focus on CMOS implementations.
The 3D Silicon Coincidence Avalanche Detector concept has been introduced and discussed in the second part of the chapter. The noise performance and the charged particle detection efficiency for this device have been defined in analogy with the discussion about SPAD detectors. Given the 3D structure for this novel device, a specific parameter referred to as “angular efficiency” has been introduced, in order to account for counting losses due to the incidence angle of the incoming radiation. Given the novelty of the device, a short state-of-the art, mostly concerning current developments, has been finally presented.

The following chapter will address the main development steps that have been faced for the design of a first demonstrator of a 3D Silicon Coincidence Avalanche Detector (3D-SiCAD).

References

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

[2] S. M. Sze and K. Ng, *Physics of Semiconductor Devices*, 3rd Editio. 2006.

[3] F. Zappa, S. Tisa, a. Tosi, and S. Cova, "Principles and features of single-photon avalanche diode arrays," *Sensors Actuators, A Phys.*, vol. 140, no. 1, pp. 103–112, 2007.

[4] E. Vilella and a. Diéguez, "A gated single-photon avalanche diode array fabricated in a conventional CMOS process for triggered systems," *Sensors Actuators, A Phys.*, vol. 186, pp. 163–168, 2012.

[5] E. Charbon, M. Fishburn, R. Walker, R. K. Henderson, and C. Niclass, "SPAD-based sensors," *TOF Range-Imaging Cameras*, pp. 11–38, 2013.

[6] R. J. McIntyre, "On the avalanche initiation probability of avalanche diodes above the breakdown voltage," *IEEE Trans. Electron Devices*, vol. 20, no. 7, 1973.

[7] W. G. Oldham, R. R. Samuelson, and P. Antognetti, "Triggering phenomena in avalanche photodiodes," *IEEE Trans. Electron Devices.*, vol. ED-19, pp. 1056–1060, 1972.

[8] M. M. Vignetti, F. Calmon, P. Lesieur, and A. Savoy-Navarro, "Simulation study of a novel 3D SPAD pixel in an advanced FD-SOI technology," *Solid. State. Electron.*, 2016.

[9] A. Gulinatti, A. Gulinatti, I. Rech, I. Rech, S. Fumagalli, S. Fumagalli, M. Assanelli, M. Assanelli, M. Ghioni, M. Ghioni, S. D, and S. D, "Modeling photon detection efficiency and temporal response of single photon avalanche diodes," *Processing*, vol. 7355, pp. 1–17, 2009.

[10] M. A. Green and M. Keevers, "Optical properties of intrinsic silicon at 300 K," *Prog. Photovoltaics*, vol. 3, no. 3, pp. 189–192, 1995.

[11] J. Beringer et al, "Review of Particle Physics*," *Phys. Rev. D*, vol. 86, no. 1, p. 10001, 2012.

[12] P. G. Rancoita, "Silicon detectors and elementary particle physics," *J. Phys. G Nucl. Phys.*, vol. 10, no. 3, p. 299, 1984.

[13] S. M. Sze, *Semiconductor Devices: Physics and Technology*. 2002.

[14] "Sentaurus Device User," no. March, 2013.

[15] K. E. Jensen, P. I. Hopman, E. K. Duerr, E. a. Dauler, J. P. Donnelly, S. H. Groves, L. J. Mahoney, K. a. McIntosh, K. M. Molvar, a. Napoleone, D. C. Oakley, S. Verghese, C. J. Vineis, and R. D. Younger, "Afterpulsing in Geiger-mode avalanche photodiodes for 1.06 μm wavelength," *Appl. Phys. Lett.*, vol. 88, no. 13, pp. 27–30, 2006.

[16] S. Vinogradov, "Analytical models of probability distribution and

excess noise factor of solid state photomultiplier signals with crosstalk," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 695, pp. 247–251, 2012.

[17] D. Bronzi, F. Villa, S. Bellisai, S. Tisa, G. Ripamonti, and A. Tosi, "Figures of merit for CMOS SPADs and arrays," *Proc. SPIE 8773, Phot. Count. Appl. IV; Quantum Opt. Quantum Inf. Transf. Process.*, vol. 8773, p. 877304, 2013.

[18] T. Nagano, K. Yamamoto, K. Sato, N. Hosokawa, A. Ishida, and T. Baba, "Improvement of Multi-Pixel Photon Counter (MPPC)," pp. 1657–1659, 2011.

[19] A. Vila, E. Vilella, O. Alonso, and A. Dieguez, "Crosstalk-free single photon avalanche photodiodes located in a shared well," *IEEE Electron Device Lett.*, vol. 35, no. 1, pp. 99–101, 2014.

[20] M. Wahl, "Time-correlated single photon counting," pp. 1–14, 2014.

[21] A. Lacaita, M. Mastrapasqua, M. Ghioni, and S. Vanoli, "Observation of avalanche propagation by multiplication assisted diffusion in p-n junctions," *Appl. Phys. Lett.*, vol. 57, no. 5, 1990.

[22] A. Lacaita, S. Cova, A. Spinelli, and F. Zappa, "Photon-assisted avalanche spreading in reach-through photodiodes," *Appl. Phys. Lett.*, vol. 62, no. 6, 1993.

[23] M. A. Albota, R. M. Heinrichs, D. G. Kocher, D. G. Fouche, B. E. Player, M. E. O'Brien, B. F. Aull, J. J. Zayhowski, J. Mooney, B. C. Willard, and R. R. Carlson, "Three-dimensional imaging laser radar with a photon-counting avalanche photodiode array and microchip laser," *Appl. Opt.*, vol. 41, no. 36, pp. 7671–7678, 2002.

[24] D. Bronzi, F. Villa, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, and W. Brockherde, "100 000 Frames/s 64 x 32 Single-Photon Detector Array for 2-D Imaging and 3-D Ranging," *IEEE J. Sel. Top. Quantum Electron.*, vol. 20, no. 6, pp. 354–363, Nov. 2014.

[25] M. Vitali, D. Bronzi, A. J. Krmpot, S. N. Nikolic, F. J. Schmitt, C. Junghans, S. Tisa, T. Friedrich, V. Vukojevic, L. Terenius, F. Zappa, and R. Rigler, "A Single-Photon Avalanche Camera for Fluorescence Lifetime Imaging Microscopy and Correlation Spectroscopy," *IEEE J. Sel. Top. Quantum Electron.*, vol. 20, no. 6, pp. 344–353, Nov. 2014.

[26] M. a. Tetrault, E. D. Lamy, a. Boisvert, C. Thibaudeau, M. Kanoun, F. Dubois, R. Fontaine, and J. F. Pratte, "Real-Time Discrete SPAD Array Readout Architecture for Time of Flight PET," *IEEE Trans. Nucl. Sci.*, pp. 1–6, 2015.

[27] A. Rochas, A. R. Pauchard, P. A. Besse, D. Pantic, Z. Prijic, and R. S. Popovic, "Low-noise silicon avalanche photodiodes fabricated in conventional CMOS technologies," *IEEE Trans. Electron Devices*, vol. 49, no. 3, pp. 387–394, Mar. 2002.

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

[29] A. Lacaita, M. Ghioni, and S. Cova, "Double epitaxy improves single-photon avalanche diode performance," *Electron. Lett.*, vol. 25, no. 13, pp. 841–843, 1989.
[30] F. Villa, D. Bronzi, Y. Zou, C. Scarcella, G. Boso, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, U. Paschen, and W. Brockherde, "CMOS SPADs with up to 500 μm diameter and 55% detection efficiency at 420 nm," *J. Mod. Opt.*, vol. 61, no. 2, pp. 102–115, 2014.
[31] D. Bronzi, F. Villa, S. Bellisai, S. Tisa, A. Tosi, G. Ripamonti, F. Zappa, S. Weyers, D. Durini, W. Brockherde, and U. Paschen, "Large-area CMOS SPADs with very low dark counting rate," *Proc. SPIE 8631, Quantum Sens. Nanophotonic Devices X*, vol. 8631, p. 86311B–86311B–8, 2013.
[32] M. a Karami, H. J. Yoon, and E. Charbon, "Single-photon Avalanche Diodes in sub-100nm Standard CMOS Technologies," *Intl. Image Sens. Work.*, 2011.
[33] H. Finkelstein, M. J. Hsu, and S. Esener, "An ultrafast Geiger-mode single photon avalanche diode in," vol. 6372, pp. 1–10, 2006.
[34] C. Niclass, M. Gersbach, R. K. Henderson, L. a. Grant, and E. Charbon, "A Single Photon Avalanche Diode Implemented in 130-nm CMOS Technology," *Sel. Top. Quantum Electron. IEEE J.*, vol. 13, no. 4, pp. 863–869, 2007.
[35] M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, and E. Charbon, "A low-noise single-photon detector implemented in a 130 nm CMOS imaging process," *Solid. State. Electron.*, vol. 53, no. 7, pp. 803–808, 2009.
[36] L. Pancheri and D. Stoppa, "Low-noise CMOS single-photon avalanche diodes with 32 ns dead time," *ESSDERC 2007 - Proc. 37th Eur. Solid-State Device Res. Conf.*, pp. 362–365, 2008.
[37] 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 Trans. Electron Devices*, vol. 58, no. 7, pp. 2028–2035, 2011.
[38] 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 Lett.*, vol. 33, no. 5, pp. 694–696, 2012.
[39] 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 Lett.*, vol. 33, no. 11, pp. 1589–1591, 2012.
[40] V. Saveliev, "Avalanche Pixel Sensor and Related Methods," US patent 8269181, 2012.
[41] N. D'Ascenzo, P. S. Marrocchesi, C. S. Moon, F. Morsani, L. Ratti, V. Saveliev, A. S. Navarro, and Q. Xie, "Silicon avalanche pixel sensor for high precision tracking," *J. Instrum.*, vol. 9, no. 3, p. C03027,

2014.
[42] L. Pancheri, P. Brogi, G. Collazuol, G. F. Dalla Betta, A. Ficorella, P. S. Marrocchesi, F. Morsani, L. Ratti, and A. Savoy-Navarro, "First prototypes of two-tier avalanche pixel sensors for particle detection," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, pp. 1–4, 2016.
[43] R. J. Mcintyre, "A New Look at Impact Ionization — Part I : A Theory of Gain , Noise , Breakdown Probability , and Frequency Response," vol. 46, no. 8, pp. 1623–1631, 1999.
[44] S. Aboujja, "Electroluminescence en avalanche des jonctions p-n a base de silicium et d'arseniure de gallium, et effet d'irradiation," University of Sherbrooke (Canada), 2000.
[45] A. L. Lacaita, F. Zappa, S. Bigliardi, and M. Manfredi, "On the bremsstrahlung origin of hot-carrier-induced photons in silicon devices," *IEEE Trans. Electron Devices*, vol. 40, no. 3, pp. 577–582, Mar. 1993.

Chapter 2: Design of a 3D Silicon Coincidence Avalanche Detector prototype

This chapter presents the main development steps that have been faced for the design of a first demonstrator of a 3D Silicon Coincidence Avalanche Detector (3D-SiCAD): the design of a SPAD pixel cell in a commercial High-Voltage 0,35 μm CMOS process, consisting of a proper avalanche diode architecture with associated quenching electronics to ensure correct Geiger-mode operation; the design of the 3D-level pixel electronics to provide a proper interfacing between the two sensing levels in a 3D-SiCAD pixel and, more importantly, assessing the occurrence of coincidence hits; the study of a 3D integration strategy impacting the layout of the final test-chip and the choice of the assembling technique for the realization of a first prototype.

2.1 Design of the avalanche diode

There are some major challenges that have to be addressed in the development of a CMOS compatible SPAD, as discussed in Section 1.1.6.3. First of all, the CMOS fabrication process should be sufficiently clean, i.e. characterized by a low density of defects and impurities, in order to obtain a low Dark Count Rate (DCR) and a negligible after-pulsing probability. Furthermore, the avalanche diode architecture should ensure a uniform and moderate electric field all over the active region in order to provide a uniformly distributed breakdown probability as well as to reduce field-enhanced generation phenomena in the space charge region of the SPAD. More importantly, the avalanche diode architecture has to prevent premature breakdown at the junction edges, i.e. Premature Edge Breakdown (PEB), by properly softening the electric field all around the device periphery. The accomplishment of all these design requirements is particularly challenging in a commercial standard CMOS technology which does not allow a tailored design but enables, conversely, a cost-effective production of a complete detector based system, integrating the sensor together with the read-out electronics. The choice of the avalanche diode architecture among those briefly presented in Chapter 1 and studied in detail in [1] (summarized in the Annex of this thesis) has been eventually driven by the adopted CMOS technology for the design of a 3D-SiCAD prototype. The selection of the CMOS process has been driven, in turn, by the manufacturing costs (e.g. multi-project wafer runs), the existence of previous SPAD implementations with this specific technology and expected noise performance.

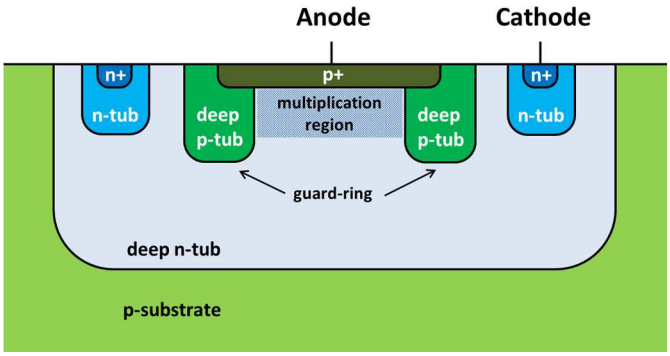

Figure 1: *Schematic cross-section of the avalanche diode architecture implemented in the HV-AMS* 0.35 µm *CMOS technology.*

In the end, the priority of this work was to demonstrate the feasibility of the 3D-SiCAD concept in a standard CMOS process by providing a first-ever made fully functioning prototype, avoiding any risk concerning the SPAD architecture. The High-Voltage Austria Micro-Systems 0.35 μm CMOS technology (HV-AMS 0.35) has been explored in this work for the realization of a 3D-SiCAD prototype. A conventional "diffused guard-ring" implementation [2] has been adopted for the avalanche diode as depicted in Figure 1. The active region consists of a p+/n junction, obtained by means of a p+ diffusion over the deep n-tub region (no n-tub enrichment has been drawn), while deep low doped p-tub surrounding the junction forms the guard ring. Apart from insulating the SPAD from substrate noise and allowing an independent bias for both the anode and the cathode, the deep n-tub prevents also free carriers generated in the substrate to diffuse into the junction, improving the SPAD timing-performance, as for the "double-epitaxial SPAD" [3] discussed in Chapter 1. As shown in Figure 2, the diode active area has a square-like geometry with a 50 μm side length, which is compliant with typical resolution requirements of nuclear physics, as discussed in the introduction of this thesis.

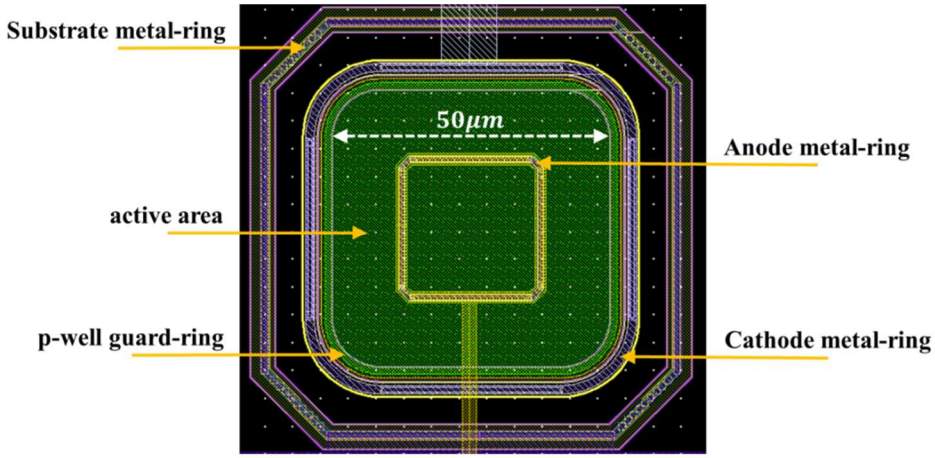

Figure 2: *Layout of the avalanche diode architecture implemented in the HV-AMS* 0.35 µm *CMOS technology.*

Since the sharp corners of square-like geometry can lead to localized high electric field, these have been round shaped.

2.2 Integrated Electronics for Geiger-mode operation

A SPAD pixel has to integrate suitable electronics capable to drive the avalanche diode in the *Geiger-mode*. As discussed in Chapter 1, the electronics is primarily responsible to *quench* the avalanche multiplication process right after the current build-up by promptly lowering the reverse bias of the junction at or below the breakdown voltage (quenching phase). Then, after a certain dead-time (hold-off time) during which the pixel cannot fire any avalanche, the electronics restores the p-n junction to the initial bias (reset or recharge phase), and the device is finally ready to detect another event. Sections 2.2.1 and 2.2.2 provide a brief description over the main concept of passive and active quenching approaches that are commonly adopted for realization of Geiger-mode avalanche diodes. A complete overview about the state-of-the-art of quenching circuits implementation until 2010 is provided by the work of Gallivanoni et al. [4].

2.2.1 Passive Quench

Since the early studies on avalanche breakdown in junctions, the avalanche current quenched itself simply by developing a voltage drop on a high resistive element placed in series with the sensor [5], as shown schematically in Figure 3. This simple solution is still widely employed and is commonly referred to as *passive quenching* approach [5]. In modern CMOS SPADs, the resistive element can be implemented by means of a simple integrated resistor or a MOS transistor with a suitable sizing and bias. As a rule of thumb, in order to guarantee a correct quenching, the resistance value should be at least $50\ k\Omega/V$ of applied excess bias voltage V_{ex} [5]. The reason of this sizing can be understood with the help of Figure 3. An avalanche triggering corresponds to closing the switch in the diode equivalent circuit. The quenching dynamics of the device is thus determined by the RC circuit defined by the resistive quenching element R_Q and the diode resistance R_d (the series of space–charge resistance of the avalanche junction and of the ohmic resistance of the neutral semiconductor crossed by the current), as well as the depletion capacitance of the sensor C_d together with the parasitic capacitance C_p. If the quenching resistor is correctly sized, the avalanche current discharges the capacitances so that the voltage across the diode exponentially falls from $V_{bd} + V_{ex}$ towards the breakdown voltage threshold of the avalanche diode V_{bd}. Similarly, the avalanche current falls towards an asymptotic value given by $I_Q = V_{ex}/R_Q$.

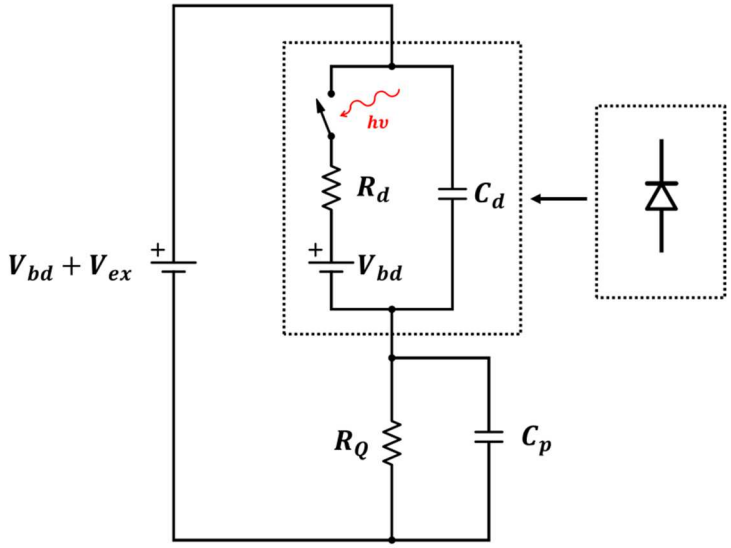

Figure 3: *Equivalent circuit of a passively quenched SPAD as discussed in [5].*

A correct quenching can thus happen only if this asymptotic value is lower than a certain latching current of roughly $100\ \mu A$ [5], below which the multiplication process cannot self-sustain anymore. The quenching time constant is thus given by:

$$\tau_Q = (C_d + C_p)R_Q \parallel R_d \approx (C_d + C_p)R_d \tag{1}$$

where it is reasonable to assume that $R_Q \gg R_d$, as in thin junction SPADs (or in general in the CMOS ones) the diode resistance is at most of the order of just a few $k\Omega$ [5]. Apart from representing an important parameter for the SPAD timing-performance, the quenching time is related to the overall charge flowing through the device during an avalanche pulse, and it is therefore an important figure with respect to after-pulsing phenomena. In order to mitigate this latter, it is indeed important to minimize the amount of avalanche charge, which is equivalent to requiring a very fast quenching time. Passive quench in this case does not look really suitable for SPAD devices featuring off-chip electronics as the external parasitic capacitance, introduced by the connection of the avalanche diode with the external quenching electronics, would lead to a long quenching time and a rather large amount of avalanche charge. Conversely, in fully integrated CMOS SPAD devices, the avalanche charge is mainly determined by the diode space charge capacitance. Therefore the quenching action is rather fast, i.e. on the order of a few hundreds of picosecond, and after-pulsing is certainly reduced [4]. Once the quenching of the avalanche has been accomplished (opening of the switch in the diode equivalent circuit), the capacitances are slowly recharged by the small current imposed by the quenching resistor R_Q and the diode voltage eventually recovers exponentially toward the initial bias voltage with a time constant given by:

$$\tau_R = (C_d + C_p)R_Q \tag{2}$$

Peculiarly, as the voltage across the avalanche diode recovers toward the steady state value (above the breakdown voltage), an electron-hole pair generated during the reset phase has a progressively higher probability to fire an avalanche process. It may therefore happen that an avalanche is ignited even if the recharge phase has not finished yet meaning that under this approach it is not possible to obtain a well-defined hold-off time. This latter is a desirable feature for a SPAD since it allows suppressing the noise count enhancement due to after-pulsing phenomena. Moreover the resulting output voltage pulses would have variable amplitude that depends on the instantaneous voltage across the diode. Therefore, the measured counting rate would be affected by the comparator threshold level, resulting into a loss of linearity and time-resolution at high counting rates [5].

2.2.2 Active Quench

Active quench approach aims at facing the main drawbacks of passive quenching and it has been introduced for the first time by Cova et al. [5] in 1975. Since then, a wide variety of circuit implementations has been proposed in literature. As depicted in Figure 4, active quench circuits have the capability to sense the rise of the avalanche pulse (commonly by means of a fast comparator) and promptly react back on the detector by forcing the reverse bias voltage of the diode at or below the breakdown threshold [5]. After a user-adjustable and well-defined hold-off time, the bias voltage is then restored back to the operating level. The main advantages provided by the active quench approach are manifold. First of all, by forcing a fast quenching action the amount of avalanche charge crossing the junction is definitely reduced, resulting in an important minimization of the after-pulsing probability. A fast recharge transition reduces the probability to miss rare avalanche events occurring during the reset phase.

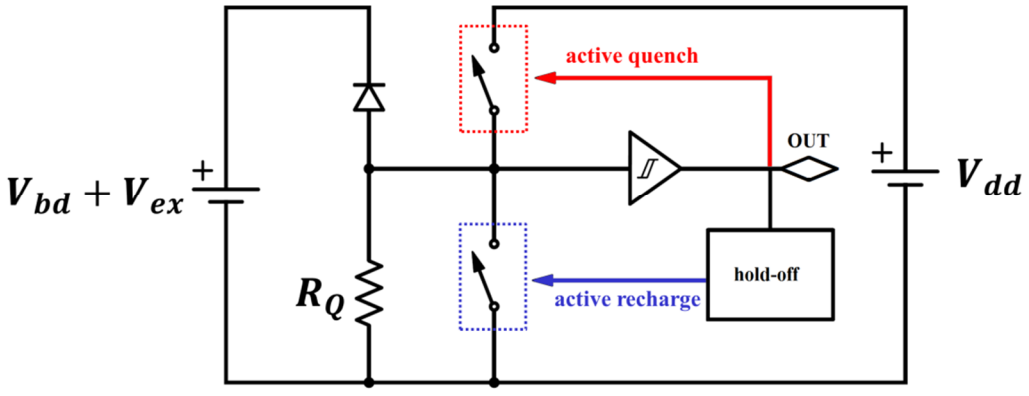

Figure 4: *Equivalent circuit of an active quench circuit. The avalanche leading edge is sensed by a fast comparator that reacts back on the avalanche diode interrupting the avalanche process. The reverse bias across the diode is then reset to the initial value after a well-defined hold-off time.*

A user-adjustable and well-defined hold-off time is provided, allowing the suppression of the noise count enhancement due to after-pulsing phenomena by waiting a sufficiently long enough time. Finally, a standard pulse synchronous to the avalanche leading edge is provided by the comparator output which can be easily processed by dedicated read-out circuits. It is interesting to observe that the active or passive quench approach may result in more or less effective way to minimize the trapped charge, depending on the specific features of the SPAD detector. Active quench might not be the best solution for a SPAD fully implemented in a standard CMOS technology, where the avalanche diode would feature small series resistance R_d as well as small parasitic capacitance C_p. The avalanche charge would indeed be mainly determined by the diode space charge capacitance C_d which results in quenching timings on the order of a few hundreds of picoseconds. The active quench delay depends indeed on the circulation time in the quenching loop from detector to quenching circuit and backwards, and on the speed of the quenching action. It turns out that it is rather difficult to implement active quenching circuits with response times shorter than the passive quench case under these circumstances. A mixed passive–active-quenching approach may be the most suitable trade-off for minimizing the avalanche charge and, at the same time, providing a well-defined and user-adjustable hold-off time. This can be accomplished by implementing a passive quench – active recharge approach as illustrated in Section 2.2.4.

2.2.3 Read-out mode

Geiger-Mode avalanche diodes can be biased in two different ways depending on whether the output node is the anode (diode TOP) or the cathode (diode BOTTOM) as shown in Figure 5. The “diode TOP” approach (Figure 5a) requires a diode whose anode is insulated from the grounded p-substrate. The anode is indeed the moving node and it cannot be stuck at ground. Therefore p+/n-well or p-well/deep n-well based avalanche diodes should generally be suitable for such a configuration. It is worth noticing that the “TOP bias” applied to such devices minimizes the total charge produced during an avalanche which consequently reduces the after-pulsing probability [1]. That is because the avalanche diodes architectures that are compatible with this configuration are made of two p-n junctions: the avalanche junction, i.e. the p+/n-well or p-well/deep n-well junctions, and a parasitic deep n-well/p-substrate junction (See Annex for more details). This means that there is a space charge capacitance C_{sub} between the cathode and the substrate and another one, i.e. C_d, between the anode and the cathode. In the “diode TOP” configuration the cathode-to-substrate parasitic capacitance C_{sub} does not contribute to the avalanche charge since the cathode and the substrate are not moving nodes. On the other hand, special care must be taken in order to make sure that the parasitic diode doesn’t experience any premature breakdown i.e. the parasitic diode breakdown voltage must be larger than the avalanche diode one.

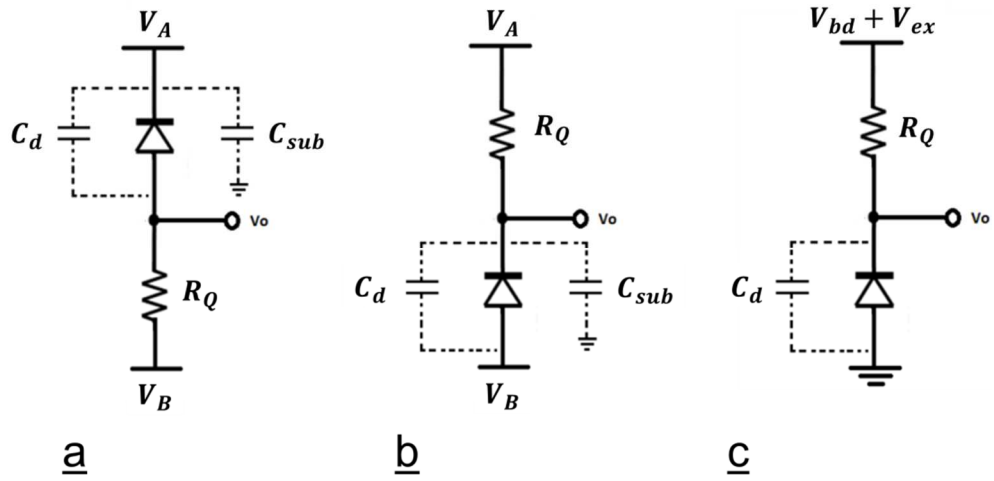

Figure 5: *Schematic circuit of a) "diode TOP" configuration b) "diode BOTTOM" configuration c) "diode BOTTOM" configuration with the grounded anode showing all the capacitive contributions (stray capacitance excluded). Observe that* $\mathrm{V_A} - \mathrm{V_B} = \mathrm{V_{bd}} + \mathrm{V_{ex}}$.

Conversely a "diode BOTTOM" approach can be implemented by using either an avalanche diode with an insulated anode (Figure 5b) or an avalanche diode sharing the anode with the grounded substrate (Figure 5c). However, in the first case, each avalanche quenching and reset phase is slowed down by the presence of the parasitic capacitance C_{sub} between the deep n-well insulating the p-type anode and the p-substrate. More importantly, such a capacitance provides an additional charge during the avalanche which enhances the after-pulsing probability. Special care must be taken in order to make sure that the deep n-well/p-substrate diode never turns on or, on the other direction, that it doesn't experience any premature breakdown, i.e. its breakdown voltage must be larger than the avalanche diode one. When the avalanche diode has a grounded anode, no parasitic capacitance has to be accounted for. However, as the breakdown voltage of an avalanche diode is normally larger than $10V$, it is always necessary to use a DC decoupling capacitor between the output of the diode and the input of the read-out electronics (configuration presented in Figure 5c).

2.2.4 Quenching circuit for the 3D-SiCAD sensing levels

In this section, a novel time-integration based quenching circuit implemented in the AMS High-Voltage CMOS 0.35 μm technology for a $50 \times 50\ \mu m^2$ active area avalanche diode is presented [6]. The circuit has been designed by adopting a passive quenching and an active recharge approach with the aim of minimizing the overall avalanche charge and thus reducing as much as possible the after-pulsing degradation of the detector noise counts. Moreover the circuit features a

tunable hold-off time within a wide range, i.e. $\sim 50\,ns - 5\,\mu s$, primarily for a full characterization of the SPAD noise performance, but also to mitigate the dark count enhancement due to after-pulsing phenomena. Figure 6 (left) shows the proposed quenching circuit. Thanks to a passive quenching approach, the total avalanche charge is indeed limited by the total parasitic capacitance on the diode moving node (node A in Figure 6). The quenching electronics would add only a few femtofarads on node A, which is practically negligible with respect to a $50 \times 50\ \mu m^2$ avalanche diode parasitic capacitance (on the order of a few hundreds of fF). The active recharge approach enables an external user-defined hold-off time ranging from a few tens of nanoseconds up to a few microseconds. Figure 6 (right) shows some key nodes waveforms of the proposed circuit after an avalanche event. Before every avalanche ignition, the voltage across the diode is $V_{bd} + V_{ex}$ and node A is at ground. As soon as an avalanche is triggered, e.g. by an incoming photon, node A promptly rises to $V_A = V_{ex}$ (i.e. the voltage across the diode drops to V_{bd}) since transistor Q1 (in the OFF state) acts as a high resistance quenching element. Q2 turns consequently ON and pulls node OUT from V_{dd} to ground (Q3 is OFF). In this way Q4 is turned OFF and the DC current provided by Q5 (externally controlled by a hold-off voltage V_h) starts to be integrated over time through the capacitor C_h. Once node B crosses the threshold of the inverter chain U1-U2, the recharge phase is started through Q7 and U3 by turning Q1 ON.

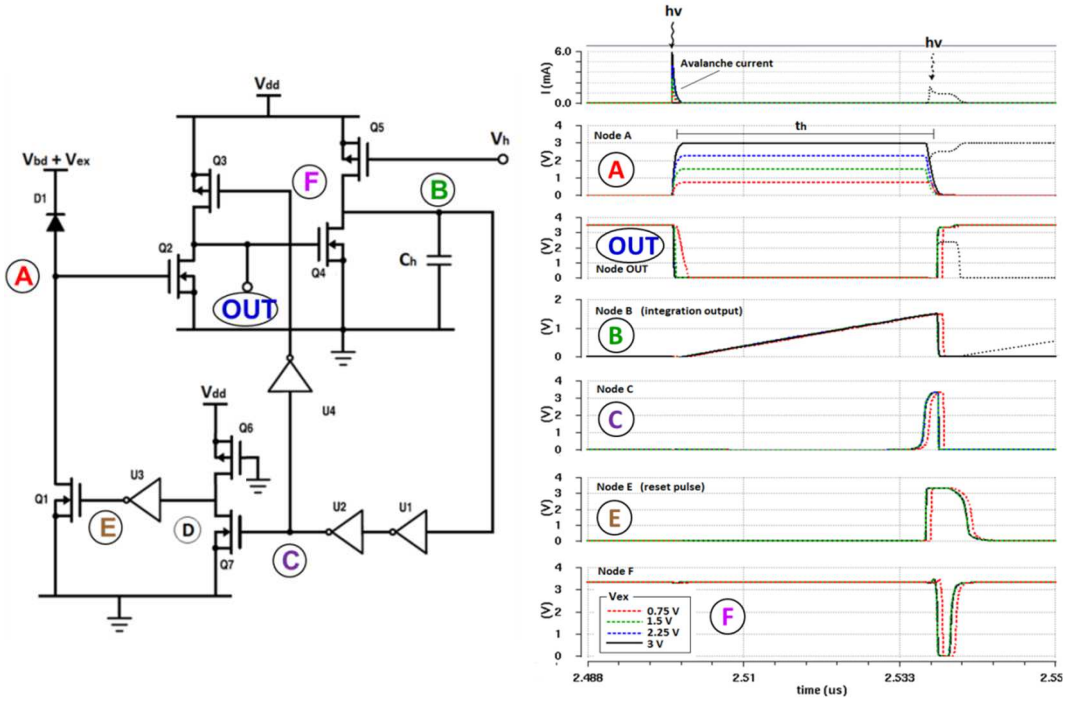

Figure 6: *Left: schematic diagram of the time-integration based quenching/recharge circuit proposed in this work. Right: simulation results showing key node waveforms after an avalanche event for an hold-off time of* $t_h = 35ns$.

At the same time, the integration process is stopped through U4 and Q3 by turning Q4 ON and thus discharging C_h. The commutation of node B to ground propagates to node E which eventually commutes to ground as well. Q1 turns eventually OFF and the recharge phase is stopped. The diode is now ready to detect another event.
Few additional comments are worth mentioning. It is indeed important that Q1 stays ON for a sufficiently long time in order to fully reset node A to ground. This has been accomplished thanks to a sufficiently long reset pulse and a slightly delayed time-integration interruption (Figure 6 right, Node E and F respectively). Moreover, even if an avalanche event occurs during the reset phase, the proposed circuit successfully assesses the event occurrence. If, for instance, an avalanche is triggered right after the recharge phase starts (black short dotted lines in Figure 6 right), node A would probably stack close to $V_A = V_{ex}$ meaning that Q2 keeps staying ON. However, as Q3 is still ON, node OUT is forced to cross the threshold voltage of Q4. In this way, the integration is interrupted anyway and, a few nanoseconds later, the recharge phase too. Therefore Q3 turns OFF and since Q2 is still ON, node OUT is pulled to ground, starting a new time-integration process. The avalanche event is not missed. Worthless to say, a proper sizing of Q2 and Q3 is crucial for a successful design.
Several observations on the circuit sizing are worth mentioning. First of all, both Q1 and Q2 have to be sufficiently conductive in order to enable a reset phase and a commutation time for the node OUT in a few nanoseconds, even for the lowest allowed diode excess bias, i.e. when Q2 is biased only a few tens of mV over its threshold voltage. Of course the sizing has to take into account the additional parasitic capacitance introduced by Q1 and Q2 on node A which should not be larger than a few femtofarads in order to reduce as much as possible the after-pulsing probability. On the other hand Q3 has to be sized accordingly to Q2 in a way that the former is always more conductive than the latter when both are fully ON. This guarantees that if an avalanche event occurs during the reset phase, Q3 is strong enough to allow node OUT crossing the logic threshold of a possible read-out electronics which can consequently detect the avalanche event.
It is worth noticing that also the inverter chain U1-U2 following the time-integration stage plays an important role. It “regenerates” indeed the slow ramp signal from the time-integration circuit to a good rectangular pulse in order to correctly drive both the reset phase and the time-integration stop.
The time-integration process is carried out by Q5 through C_h and it deserves more detailed comments. The integration time, i.e. the hold-off time, can be estimated as:

$$t_h \approx \frac{V_{th-chain} C_h}{I_{Q5}} \tag{3}$$

where $V_{th-chain}$ is the inverter chain threshold required for the low-to-high commutation of node OUT (turning Q4 ON and thus interrupting the integration process) and I_{Q5} is the current supplied by Q5. The $(W/L)_p$ ratio of Q5 can be expressed as a function of the surface A_h of C_h:

$$\left(\frac{W}{L}\right)_p = \frac{V_{th-chain} C'_h A_h}{2 I'_p L_{min} t_h} \left(\frac{V_{ov,Q5}^{max}}{V_{ov,Q5}}\right)^2 \quad (4)$$

where C'_h is the capacitance per unit surface, I'_p is the pMOS maximum saturation current density with respect to the transistor width, $V_{ov,Q5}^{max}$ and $V_{ov,Q5}$ are the maximum and the adopted respectively overdrive voltages for Q5. Assuming that $(W/L)_p = 1$, $V_{th-chain} = V_{dd}/2 = 1.65V$, $C'_h = 1fF/\mu m^2$, $I'_p = 200\mu A/\mu m$, $L_{min} = 0.35\mu m$, $V_{ov,Q5} = 500mV$, $V_{ov,Q5}^{max} = 2.8V$ and by considering the longest hold-off time, e.g. $t_h = 1\mu s$, the capacitor surface will be $A_h = 52 \times 52\ \mu m^2 \approx A_{int}$ (where A_{int} is the total area of the integration circuit). Optimization is definitely needed. From this simple estimation, it is apparent that a $(W/L)_p < 1$ has to be used. The overall area can be therefore expressed as follows:

$$A_{int} = A_h + \left(\frac{L}{W}\right)_p W_{min}^2 \quad (5)$$

The optimum sizing can be found by setting the first derivative of (5) to zero, after substituting $(W/L)_p$ in (5) with (4).

$$\begin{cases} A_h^{opt} = W_{min} \left(\frac{V_{ov,Q5}}{V_{ov,Q5}^{max}}\right) \sqrt{\frac{2 I'_p L_{min} t_h}{V_{th-chain} C'_h}} \\ \left(\frac{W}{L}\right)_p^{opt} = W_{min} \left(\frac{V_{ov,Q5}^{max}}{V_{ov,Q5}}\right) \sqrt{\frac{V_{th-chain} C'_h}{2 I'_p L_{min} t_h}} \end{cases} \quad (6)$$

Using the numbers of the previous example the total area of the integration circuit is now $A_{int}^{opt} = 4.9 \times 4.9\mu m^2$, i.e. only the 1% of the surface in the not optimal case.

As discussed before, it is really important that the recharge phase is long enough to fully reset node A to ground. For this purpose, the interruption of the integration process can be conveniently delayed by properly sizing U4 (which can be seen, more generally, as a simple delay element). In order to further improve the recharge effectiveness, the transition high-to-low of node E can be delayed by implementing a "weak" sizing of Q6. Indeed if a simple inverter were adopted instead of Q6-Q7, the reset phase would probably be prematurely interrupted. As soon as the time-integration process is stopped, the electronics would indeed immediately turn Q1 OFF with the risk that the reset phase has not yet completed. For this reason, a pseudo n-MOS inverter based on Q6-Q7 has been placed before U3. By a proper sizing of Q6-Q7 and inverter U3, the reset phase can be sufficiently extended to avoid an incomplete recharge. The sizing can be faced by considering that as soon as the time-integration is interrupted, node C is promptly pulled to ground, turning Q7 OFF. In order to extend the reset phase it

is necessary to slow-down the low-to-high transition of node D. The delay introduced by Q6 charging the input capacitance of U3 $C_{IN,U3}$ is given by:

$$\Delta t = \frac{V_{dd}}{2I_{Q6}} C_{IN,U3} \tag{7}$$

where it has been assumed that the logic commutation threshold of U3 is $V_{th}^{U3} = V_{dd}/2$. Therefore the optimal sizing can be found by following the procedure adopted for the time-integration block. Δt can be estimated as the time Q1 needs to pull node A at ground. In practice U3 is implemented with an inverter cell from the standard digital library that will be sized according to the optimal area found for the input capacitance. Q6 will be sized accordingly to U3 and (6).
A correct operation of the proposed circuit is assessed through circuit simulations as well as Monte Carlo analysis in the Cadence Environment for a High-Voltage $0.35\,\mu m$ CMOS technology. The avalanche diode has been modeled thanks to a behavioral circuit based on the work of Zappa et al. [7] and adapted to the quenching/reset circuit proposed in this work. In this way, the avalanche ignition due to photon absorption, the fast avalanche current build-up, the self-sustaining charge-multiplication process, and the self-quenching of the avalanche pulse have been accurately taken into account in simulations. Figure 6 (right) shows typical waveforms for the simulated circuit for several applied excess bias and for an hold-off voltage $V_h = 1.5\,V$. It is possible to observe that the reset pulse (node E) is large enough to allow a complete recharge of node A to ground and that the time-integration is interrupted after a small desired delay with respect to the start of the recharge phase. Furthermore, in case of an avalanche occurring during the reset phase (black short dotted lines in Figure 2), Q3 is able to pull node OUT above $V_{dd}/2$ ($V_{dd} = 3.3\,V$ in this technology), allowing a correct detection of the avalanche event by a possible read-out electronics.
A Monte Carlo analysis of the proposed circuit has been performed as well. The results of this analysis show that for a given hold-off voltage, the integration time may vary according to a statistical variation of the process parameters even if the circuit continues working correctly. The variation is as strong as the hold-off time is larger because of the dependence of the hold-off time on the overdrive voltage, i.e. $V_{ov} = V_{sg} - V_{Th\_p}$ applied to Q5 (V_{sg}, V_{Th_p}: source-gate voltage and threshold voltage respectively of pMOS Q5). For larger values of V_{ov}, the current flowing through Q5 is less sensitive to statistical variations of the transistor threshold voltage. For this reason, as the hold-off times of the proposed circuit are tuned to be shorter and shorter, they are expected to approach to the nominal case ones. Statistical variations of the parameters due to mismatch are conversely not really affecting the performance of the device thanks to the required large transistor sizing. A t_h versus V_h pre-characterization step is thus necessary in order to successfully implement the proposed circuit for characterization of a Geiger-mode avalanche diode, since process corners might lead to hold-off time discrepancies with respect to nominal simulation results. These aspects will be discussed in more detail in Chapter 3.

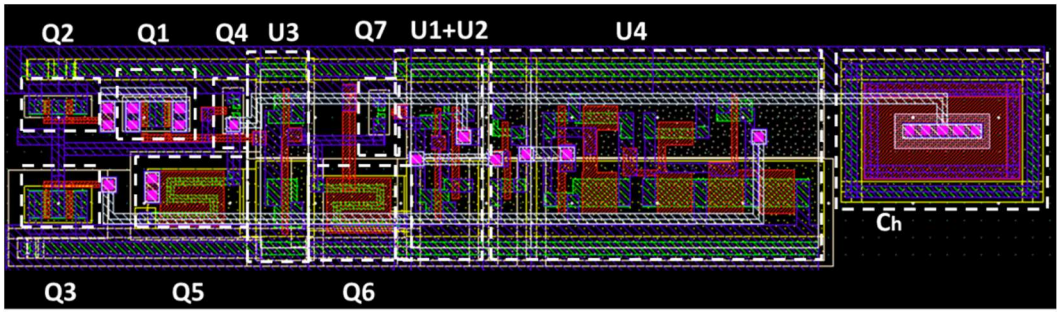

***Figure** 7: Layout of the time-integration based passive quenching – active re-charge circuit proposed in the present work.*

Static power consumption of the circuit is mainly due to transistor Q5. The maximal static power dissipation occurs when the shortest hold-off time is desired and it is around $5\mu W$. This value comes from the optimal sizing discussed previously where the main objective was to search for a "minimal surface" design and is a consequence of implementing a wide range of the externally-tunable hold off time. Power dissipation can be naturally reduced considerably by adopting a not optimal design at the expense of a larger area. In reality, in the present work power consumption has not been considered as a major concern as the proposed circuit is meant to provide a characterization tool for Geiger-mode avalanche diodes. Thanks to the adopted optimal sizing the circuit requires a total surface of $60 \times 13\ \mu m^2$, which represents 10% of the total pixel size in case a $50 \times 50\ \mu m^2$ active area avalanche diode is adopted (the diode total surface is indeed $80 \times 80\mu m^2$ due to the substrate metal-ring). The fill-factor of the SPAD pixel is expected to be around 35%. Figure 7 shows the layout of the proposed circuit.

2.2.5 3D-level pixel electronics

As discussed in Chapter 1, a 3D-SiCAD pixel requires dedicated electronics responsible of providing a proper interfacing between the two sensing levels and, more importantly, assessing the occurrence of coincidence hits. A coincidence event can be defined, in principle, as the simultaneous occurrence of an avalanche event in both sensing levels of a 3D-SiCAD pixel. From an electrical point of view, a coincidence can be recorded only if there is an infinitesimal delay (ideally zero, and hereafter referred to as "inter-levels delay") between the leading edges of the avalanche pulses coming from the two SPAD devices. The electronics should be thus capable of discriminating inter-levels delays shorter than a certain value, referred to as "coincidence time-window", that should be naturally kept as small as possible. This concept can be smartly implemented by shortening the output pulses produced by the 3D-SiCAD sensing levels down to a well-defined coincidence time-window. In this way, the use of a simple logic gate is sufficient to process the resulting pulses and eventually record a valid count in case they overlap in time, as illustrated in Figure 8.

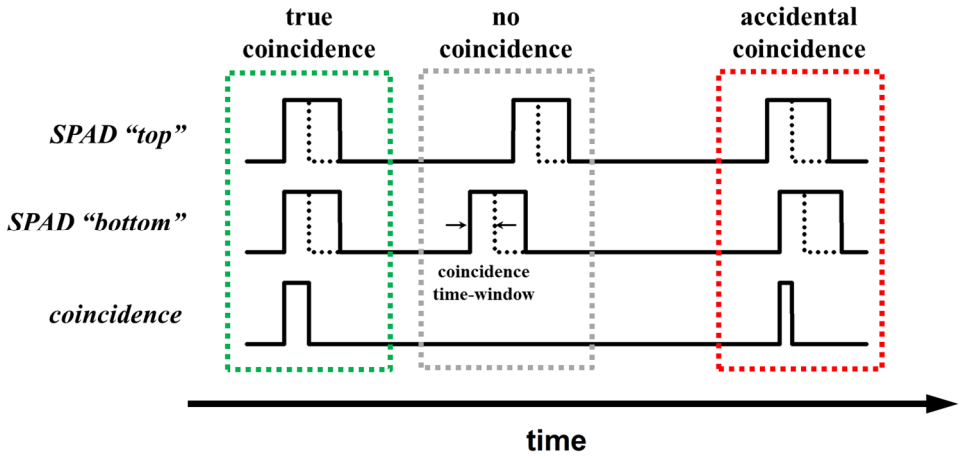

Figure 8: *Qualitative time diagram representing the main waveforms in a 3D-SiCAD pixel. Solid line waveforms represent the output signals produced by the quenching electronics in each sensing level. The dotted lines represent the ultra-short pulses that are synchronous to leading edge of an avalanche event. The width of these synchronous pulses represent the coincidence time-window. Even a partial overlap between these pulses produces a coincidence count.*

It is worth noticing how the shortening operation is very important since it provides ultra-short pulses that are synchronous to the avalanche leading edge in each sensing level. Figure 9 schematically represents the way the 3D-pixel electronics has been conceived.

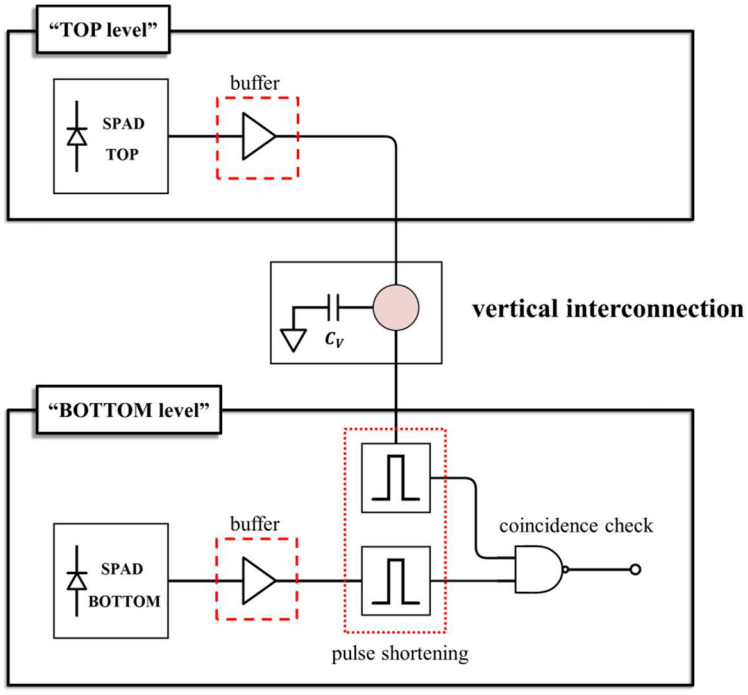

Figure 9: *Schematic block diagram of the 3D-SiCAD pixel electronics.*

The routing of the "SPAD top" output signal towards the bottom level requires a proper buffer in order to ensure minimal rising/falling time for the avalanche pulse signal which might be otherwise degraded by the parasitic capacitance introduced by the vertical interconnection. Of course, this buffer produces a systematic delay on the propagation of the signal towards the coincidence circuit. This has to be compensated by placing an identic buffer in the "bottom pixel". In this way, in case of a true coincidence (i.e. an ionizing particle crossing the pixel), the two signals reach the coincidence circuit simultaneously. This is of crucial importance because it allows using very short coincidence time-windows, which is essential in order to achieve an excellent noise rejection capability. In reality the coincidence time-window is limited by the shortest time that the electronics is capable to resolve and, for this reason, represents an important figure for the device noise performance: the smaller the time-window is, the lower the accidentals counts will be, as discussed in Chapter 1. On the other hand, it is worth noticing that even if the electronics had no timing limitation, the resolving time would be anyway bound by the time jitter (in the worst case, a few hundreds of picoseconds [5], as discussed in Chapter 1) intrinsically affecting the detector response to an incoming particle. Therefore, in order to avoid losses of true coincidence events, the time-window should be always larger than the SPAD jitter, reasonably on the order of (at least) a couple of nanoseconds. This point will be discussed in more details in Chapter 4. The pulse shortening block represented in the schematic diagram of Figure 9, is implemented by means of a mono-stable circuit that provides a user-adjustable coincidence time-window. Figure 10 shows the circuit implementation of the coincidence circuit.

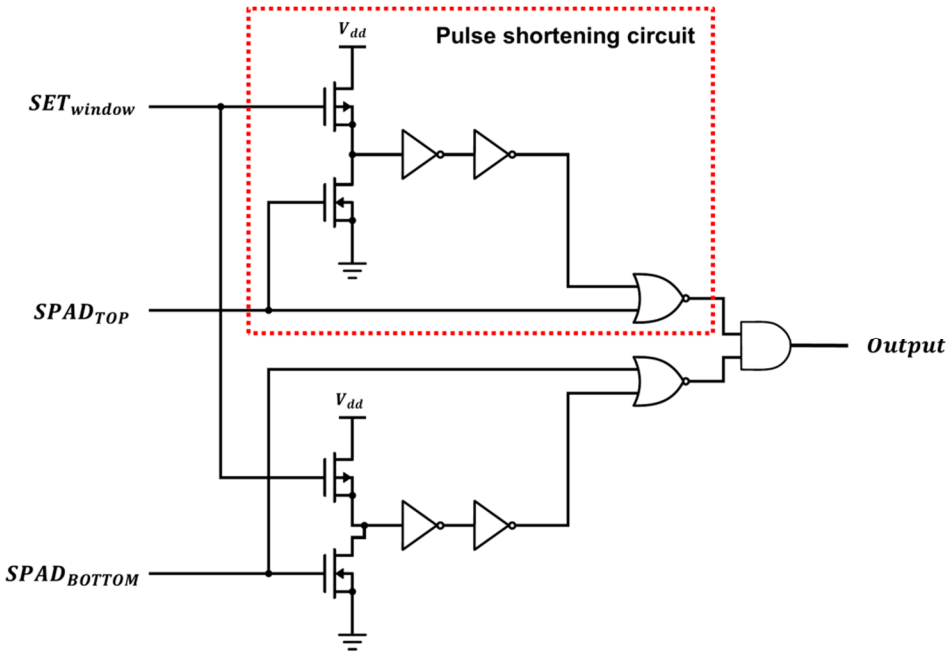

Figure 10: *"On-chip" coincidence circuit implementation. The output pulses originating from the avalanche diodes are shortened by means of a mono-stable circuit and sent towards an AND logic gate for the coincidence check.*

The mono-stable circuit operates as follows (refer to Figure 8 for a qualitative understanding of the signal waveforms). The output signal originating from one of the two avalanche diodes of the 3D-SiCAD pixel, for instance the signal $SPAD_{top}$, is sent towards the pulse-shortening block in a reversed logic, i.e. the occurrence of an avalanche event implies a "1" to "0" commutation of the SPAD output signal. This signal is then split into two electrical paths: one is sent to the input of a 2-inputs NOR logic gate; the other is sent to a 3-stages inverter chain whose output is sent to the other input of the NOR gate. It is worth noticing that the commutation time of the inverter chain can be delayed by a user-adjustable voltage SET_{window} which is responsible of controlling the time-width of coincidence time-window. The output of the NOR logic gate is thus on the "0" state before the occurrence of an avalanche event: one input is indeed at the "1" state, the other is at the "0" state. As soon as the $SPAD_{top}$ signal commutes from "1" to "0", the NOR output suddenly commutes to the "1" state, as both inputs are now on the "0" state, until the other input of the NOR gate, after a delay introduced by the inverter chain (i.e. the coincidence time-window), commutes to "1". Therefore the output of the NOR gate is sent to an AND gate, which provides a "1" output signal (coincidence event) in case both of its inputs are simultaneously on the "1" state. For the sizing of the circuit, a similar approach as the one described in section 2.2.4 has been adopted (not discussed here for the sake of brevity) in order to provide coincidence time-window ranging from around $\Delta t = 0{,}5\ ns$ up to $\Delta t = 50\ ns$.

Figure 11 reports an "off-chip" FPGA implementation of the coincidence circuit, conceived as an alternative approach with respect to the "on-chip" one discussed previously. The occurrence of asynchronous avalanche events originating from the SPAD cells of the 3D-SiCAD pixel are detected by means of the left-hand side part of the circuit shown in Figure 11 (enclosed in the red box).

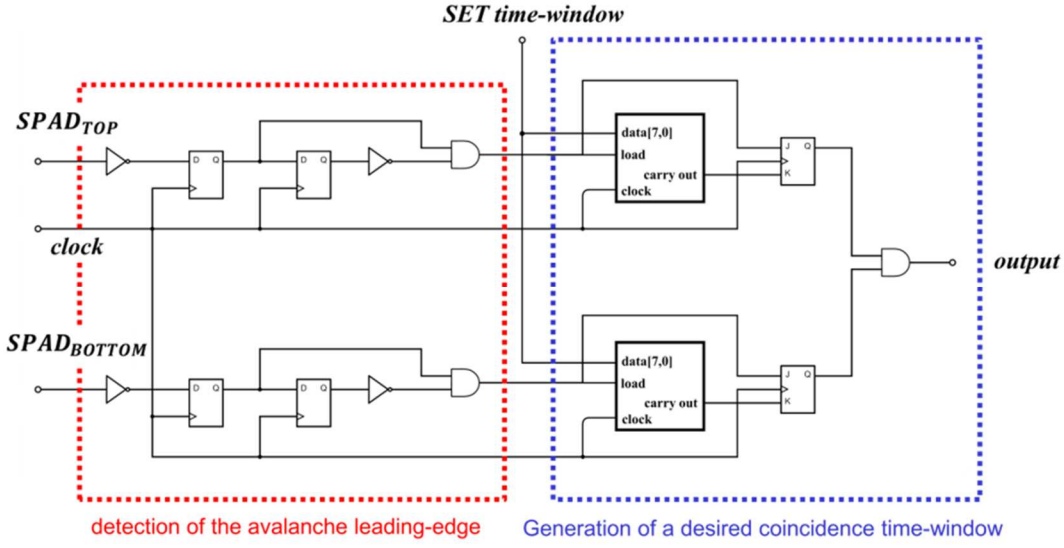

Figure 11: *"Off-chip" FPGA implementation of the coincidence circuit.*

This mono-stable block produces indeed a short pulse lasting a single clock cycle, which is synchronous to the leading edge of an avalanche signal. This pulse is then sent to the right-hand side part of the circuit in Figure 11, which generates a voltage pulse lasting a user-defined time-width, i.e. the coincidence time-window, ranging from around $\Delta t = 10\,ns$ up to $\Delta t = 100\,ns$. The resulting pulse is finally sent to one input of the 2-inputs AND logic gate for the coincidence check.

2.3 Floor plan and 3D assembling strategy

The realization of a 3D-SiCAD prototype requires the stack of two different CMOS circuits, in a way that each sensing level of a 3D-SiCAD pixel is correctly connected and well-aligned with its counterpart. In general, the choice of the 3D integration technique is mainly driven by the required density of vertical interconnections, the complexity of the ICs stack and, more importantly, the manufacturing costs. These aspects are examined in detail in this section.

2.3.1 3D assembling/integration techniques.

Vertical stacking can be accomplished by means of dedicated 3D integration technologies that are nowadays increasingly pursued by the microelectronics industry for obtaining higher performance, increased functionality, lower power consumption, and a smaller footprint from the current generation of integrated circuits. It is worth noticing that the word "3D integration" considers actually a wide variety of methods and processes that exploits the 3rd dimension to improve the electrical performance of a device. These include for instance 3D wafer-level packaging, 2.5D and 3D interposer-based integration, 3D Stacked Integrated Circuits (3D-SICs), monolithic 3D ICs, 3D heterogeneous integration and 3D systems integration. Among them, 3D-SICs technologies would probably represent the most appropriate solution for the realization of a 3D-SiCAD prototype. This approach consists of stacking silicon wafers and/or dies while providing vertical interconnections among the different levels so that the whole stack behaves as a single device. The different ICs can be stacked together by means of micro-bumps / micro-pillars structures acting as the bonding medium between the layers, or in a more advanced way (high density 3D interconnects), with direct bonding techniques such as copper-to-copper thermo-compression. However, the most attractive key feature of 3D-SICs technology is certainly the possibility to etch a direct access through the silicon substrate towards the front-end-of the line in a CMOS circuit, thanks to special vias commonly referred to as "through silicon vias" (TSVs) [8]. This allows a full interconnection capability between the different layers in a 3D stack as represented in Figure 12.

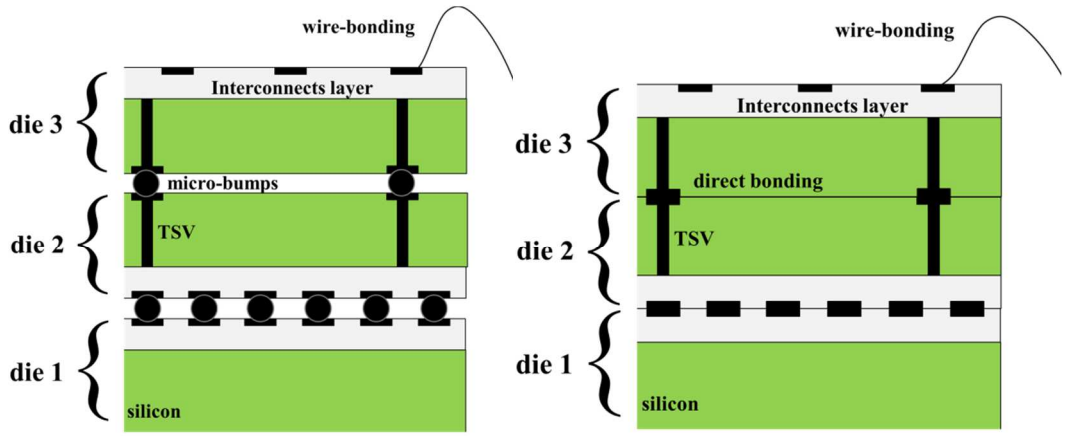

Figure 12: *Schematic representation of a typical 3D-SIC in case of a three level stack. Die 1 and Die 2 are bonded "face-to-face". Die 2 and Die 3 are bonded "back-to-back" (a mixed "back-to-face" configuration is naturally possible too). Through Silicon Vias (TSVs) allow interconnecting die 3 with the other levels. (a) Wafer bonding with micro-bumps; (b) direct wafer bonding.*

2.3.2 Adopted 3D assembling technique

It is useful to define the minimal 3D-integration features that are required for the realization of our first 3D-SiCAD prototype. It is also important to stress that the main objective of the present work is to demonstrate the feasibility and the benefits of this novel detector, with no claim of realizing a cutting-edge full detection system. The prototype is meant to be a test-chip featuring different test structures consisting of simple pixels and small pixel arrays. Therefore the density of the vertical interconnections is expected to be rather low, which relaxes the requirements on the inter-layer bonding approach. For all these reasons, the 3D stack between the two layers can be realized in a simple die-to-die approach with micro-bump bonding with no need of TSVs.
The resulting stack can be thus obtained by means of a flip-chip assembly between two different dies, as shown in Figure 13. The two dies are bonded together in a "face-to-face" configuration by means of micro-bumps. The "face-to-face" term refers to the fact that, for both the dies, the bonding surface is on the top-metal layer of the IC. The whole structure is then placed into a dedicated package, whose leads are wire-bonded only with the "bottom die". Therefore this latter is also responsible of delivering voltages and signals from the "external world" to the "top die" and vice-versa, as this latter does not have any direct wiring with the package's leads.
To a first insight, this approach would require the manufacturing of two different ICs, one for the "bottom" level and another for the "top" one, which would naturally double the prototyping costs.

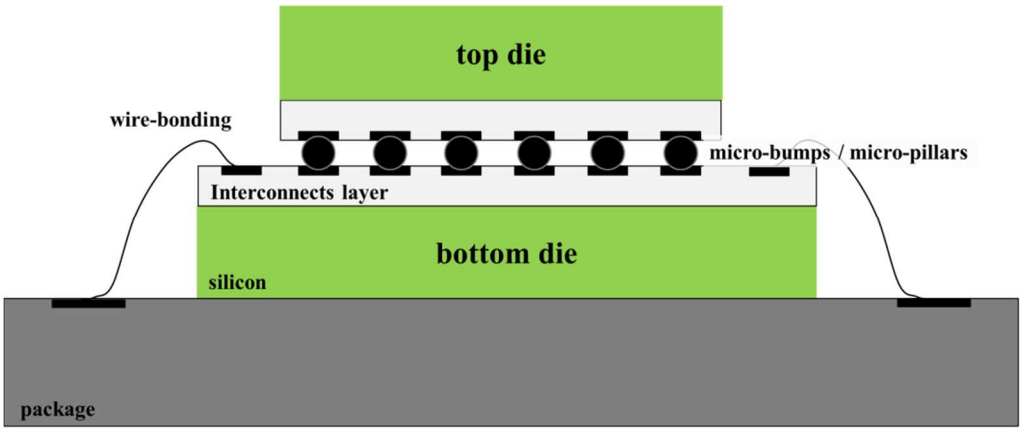

Figure 13: *3D assembly technique adopted for the 3D-SiCAD prototype.*

It would be instead more convenient to stack together two identical test-chips, with a layout conceived to provide two different regions that are superimposable when the dies are assembled according to the scheme shown in Figure 14. The "bottom" and "top" levels are laid-out, respectively, in the red and yellow triangular regions of the test chip. In this way, the desired stack between the two levels can be obtained by assembling together two identical test-chips, provided that one of the two has been first flipped and rotated by -90°. It is important to properly size "y-dimension" of the test-chip in order to avoid, after the assembly, any obstruction for the wire-bonding towards the PAD ring on the two sides. Moreover, it is worth noticing that this approach produces two copies of the same stack, i.e. "bottom + top" and "top + bottom", resulting in a 50% waste of silicon area. Nevertheless, the waste is less expensive than the fabrication of two different test-chips, as in case of a multi-project wafer run there is always a "minimal silicon area per test-chip" to be purchased for the manufacturing.

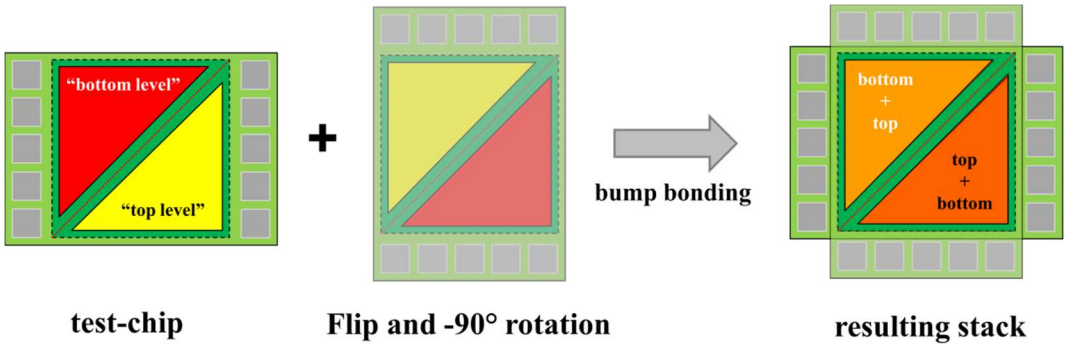

Figure 14: *Flip-chip assembly strategy: the vertical stack is obtained by assembling two identical test-chips.*

2.3.3 Floor Plan and Tape-out

Figure 15 shows the overall layout of the test-chip. The floor-plan has been designed based on the 3D integration strategy discussed previously. The cells have been indeed laid-out with respect to a certain axis of symmetry, represented by the red dashed line. In this way, each level of a 3D-SiCAD cell can be correctly aligned with its counterpart after the 3D-assembly procedure.

Observe that cells on the left-hand side are linked with the cells on the bottom side, and they all have a time-integration based passive-quench active-recharge circuitry for the SPADs. The cells on the right-hand side are conversely linked to the cells on the top side, and they have a simple passive quench circuitry for the SPADs. The resulting test-chip provides 3D-SiCAD cells arranged as a test-structure consisting of "pseudo 3D-SiCAD cell" based on two side-to-side SPAD pixels in coincidence-mode (Figure 16a), a single pixel (Figure 16b), and a 2×2 matrix (Figure 16c).

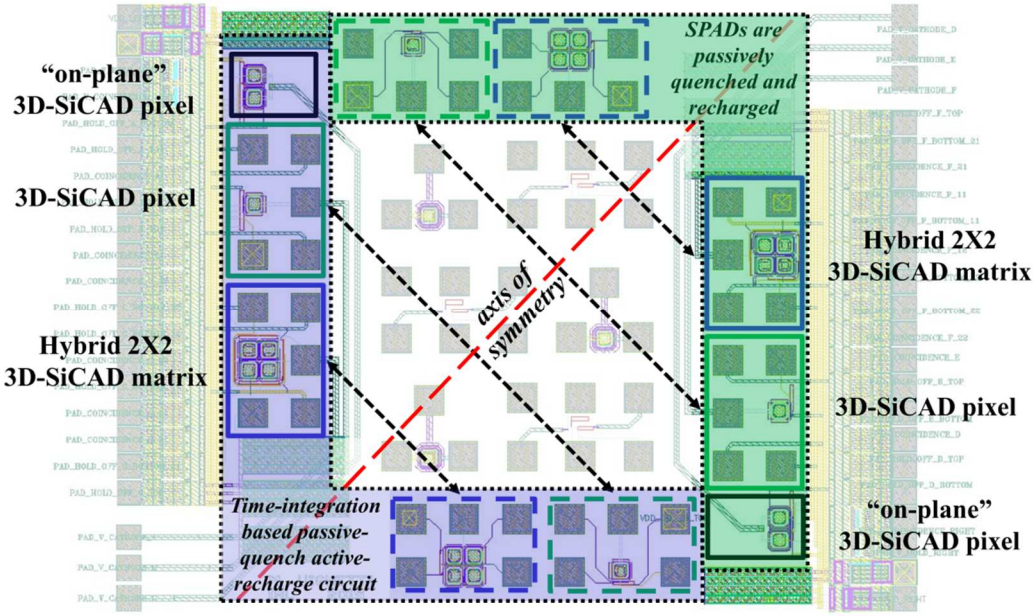

Figure 15: *Integrated circuit layout compatible with flip-chip assembly strategy. After 3D stacking, the cells on the left-hand side are linked with the cells on the bottom side (they all have a time-integration based passive-quench active-recharge circuitry) and the cells on the right-hand side are conversely linked to the cells on the top side (they have a simple passive quench circuitry).*
The other cells in the middle are described in Figure 17.

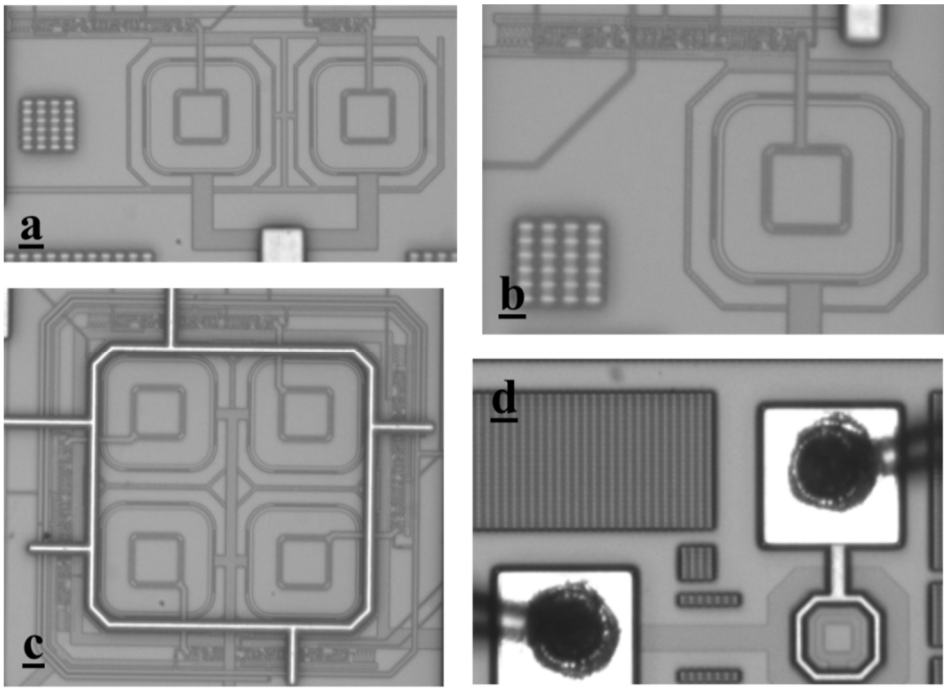

Figure 16: *Microphotographs of a) "side-to-side" SPAD pixels b) single 3D-SiCAD pixel, c) small* 2×2 *3D-SiCAD matrix, d) simple avalanche diode.*

Different stand-alone cells consisting of avalanche diodes (Figure 16d) and quenching electronics are placed in the middle of the chip in order to allow direct measurements by means of a probe station (Figure 17).

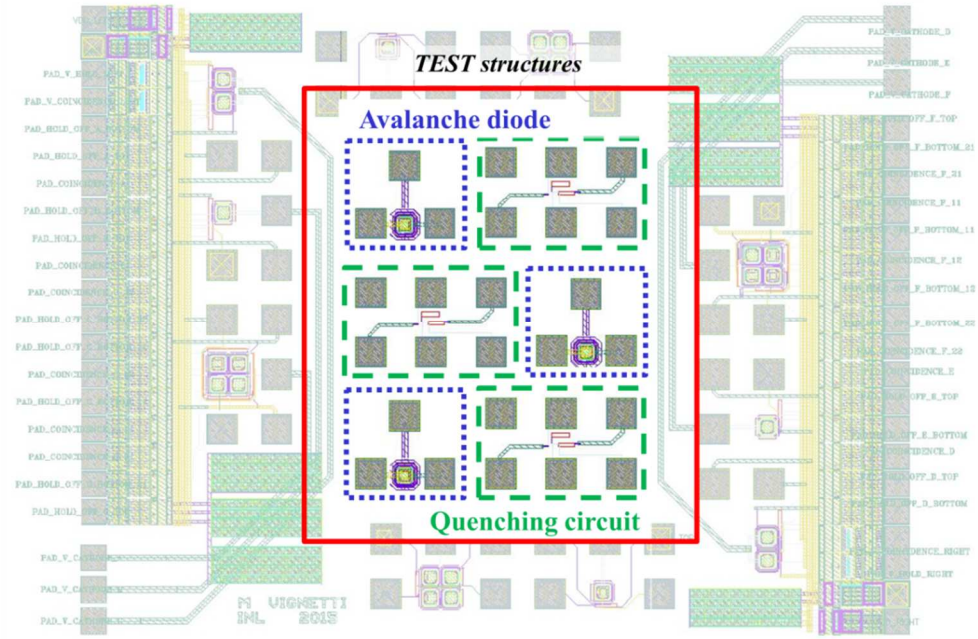

Figure 17: *Different stand-alone cells (avalanche diodes, quenching circuit) are placed in the middle allowing direct measurements by means of probe station.*

The resulting 3D prototype has been realized in collaboration with CEA-LETI and CIME Nanotech (Grenoble, France) and it is finally shown in Figure 17.

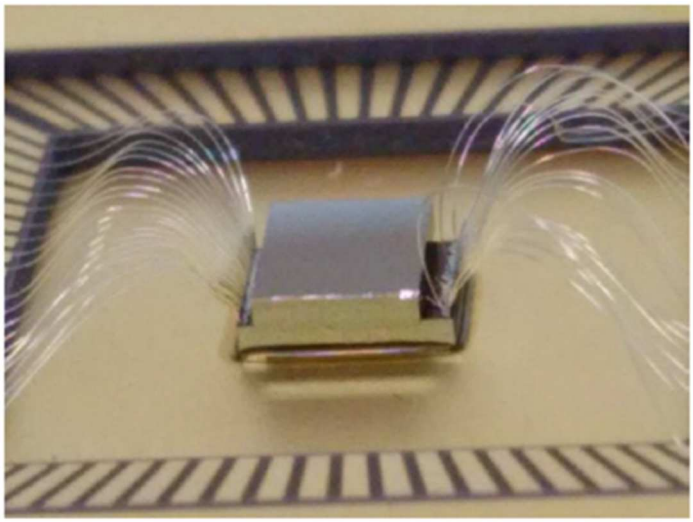

Figure 17: Microphotograph of the 3D-prototype showing the two stacked dies and the wire bonding on the left and right sides.

Conclusions

This chapter discussed in detail the main development steps that have been faced for the design of a first demonstrator of a 3D Silicon Coincidence Avalanche Detector (3D-SiCAD). The SPAD pixels have been fabricated in the AustriaMicroSystem 0,35 μm High Voltage CMOS process while adopting a "diffused guard-ring" architecture for the design of a square-like 50 μm side-length avalanche diode. The design of the associated pixel electronics has been subsequently tackled by discussing the different quenching strategies commonly adopted in order to ensure correct Geiger-mode operation. The design of an original time-integration based passive quenching / active recharge circuit has been then introduced and studied. The 3D-level pixel electronics has been conceived to ensure a proper interfacing between the two sensing levels in a 3D-SiCAD pixel and, more importantly, assessing the occurrence of coincidence hits by providing a user-adjustable coincidence time-window. The realization of a 3D prototype has been tackled by choosing a simple die-to-die flip-chip approach by means of gold micro-bumps featuring a 70 μm diameter. For this purpose, the test-chip has been laid-out in a way that a correct stack between the sensing levels of a 3D-SiCAD pixel could be ensured by assembling together two identical test-chips. The resulting 3D-assembled prototype provided 3D-SiCAD pixels arranged as single cells and 2×2 matrix cells.

The following chapter will address the characterization of the SPAD cells constituting the building block for the 3D-SiCAD pixels.

References

[1] M. M. Vignetti, F. Calmon, R. Cellier, P. Pittet, L. Quiquerez, and A. Savoy-Navarro, "Design guidelines for the integration of Geiger-mode avalanche diodes in standard CMOS technologies," *Microelectronics J.*, vol. 46, no. 10, pp. 900–910, 2015.

[2] A. Rochas, A. R. Pauchard, P. A. Besse, D. Pantic, Z. Prijic, and R. S. Popovic, "Low-noise silicon avalanche photodiodes fabricated in conventional CMOS technologies," *IEEE Trans. Electron Devices*, vol. 49, no. 3, pp. 387–394, Mar. 2002.

[3] A. Lacaita, M. Ghioni, and S. Cova, "Double epitaxy improves single-photon avalanche diode performance," *Electron. Lett.*, vol. 25, no. 13, pp. 841–843, 1989.

[4] A. Gallivanoni, I. Rech, and M. Ghioni, "Progress in Quenching Circuits for Single Photon Avalanche Diodes," *IEEE Trans. Nucl. Sci.*, vol. 57, no. 6, pp. 3815–3826, 2010.

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

[6] M. M. Vignetti, F. Calmon, R. Cellier, P. Pittet, L. Quiquerez, and A. Savoy-Navarro, "A time-integration based quenching circuit for Geiger-mode avalanche diodes," in *New Circuits and Systems Conference (NEWCAS), 2015 IEEE 13th International*, 2015, pp. 1–4.

[7] F. Zappa, a. Tosi, a. Dalla Mora, and S. Tisa, "SPICE modeling of single photon avalanche diodes," *Sensors Actuators, A Phys.*, vol. 153, no. 2, pp. 197–204, 2009.

[8] "3DinCities." [Online]. Available: http://www.3dincites.com/3d-incites-knowledge-portal/what-is-3d-integration/.

Chapter 3: Electrical and Optical Characterization of SPAD pixels

This chapter presents a characterization study of the SPAD cells constituting the building block for the 3D-SiCAD pixels. The characterization includes: the I-V curves of the avalanche diode and the breakdown voltage as a function of the temperature, the electro-luminescence test to assess the quality of the avalanche diode architecture for Geiger-mode operation, the validation of the quenching electronics, the study of the SPAD noise performance in terms of dark counts and after-pulsing, and the photon detection efficiency.

3.1 Avalanche Diode characterization

Current-voltage curves measurements and luminescence imaging of the p-n junction constituting the sensitive region of the SPADs in a 3D-SiCAD pixel have been conducted in order to assess a correct device operation and thus to validate the adopted avalanche diode architecture.

3.1.1 I-V curves

Figure 1 shows the forward bias I-V curve of the SPAD active region.

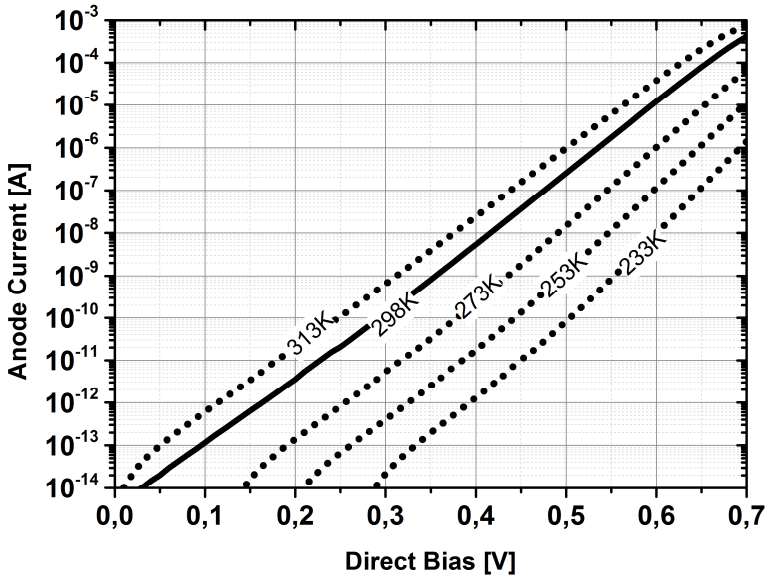

***Figure 1**: Forward bias I-V curves of the p-n junction constituting the active region of the SPAD under study within the temperature range* 233K– 313K.

The measurements have been performed on one of the "test diodes" included in the tape-out chip within the temperature range going from -40 °C up to 40 °C (i.e. 233 K – 313 K) by means of a Keithley 4200 parameter analyzer directly connected to the terminals of the sensor with the help of a cryogenic probe station. The plot shows the typical exponential relationship between the anode current and the forward bias voltage, and a correct thermal behavior according to the Shockley diode equation [1]. The carrier generation rate in the diode space charge region decreases indeed at lower temperatures, resulting in a shift of the current lines towards the right in the log-scale plot provided in Figure 1. Similarly, the slope of the lines increases at lower temperature due to a lower thermal energy kT, intervening in the exponential term of the diode equation.
Figure 2 shows the reverse bias current-voltage curves at room temperature (i.e. $T = 298\,K$) with a particular focus on the breakdown region of the p-n junction. Observe that the picture reports two different measuring conditions depending on whether the diode has been characterized in the dark (symbols) or under a controlled lighting condition (lines). If the SPAD is kept in the dark, it is rather difficult to measure the avalanche current for biases slightly above the breakdown threshold and thus to evaluate the breakdown voltage of the junction with an acceptable accuracy. Under this scenario, there is indeed a faint flickering of the diode current, toggling between the "on" (avalanche ignited) and the 'off' (device in quiescence, with a very small current flow) branches above breakdown [2]. Therefore a well-defined reverse bias plot can be effectively obtained if the measurement is performed under a controlled lighting condition.

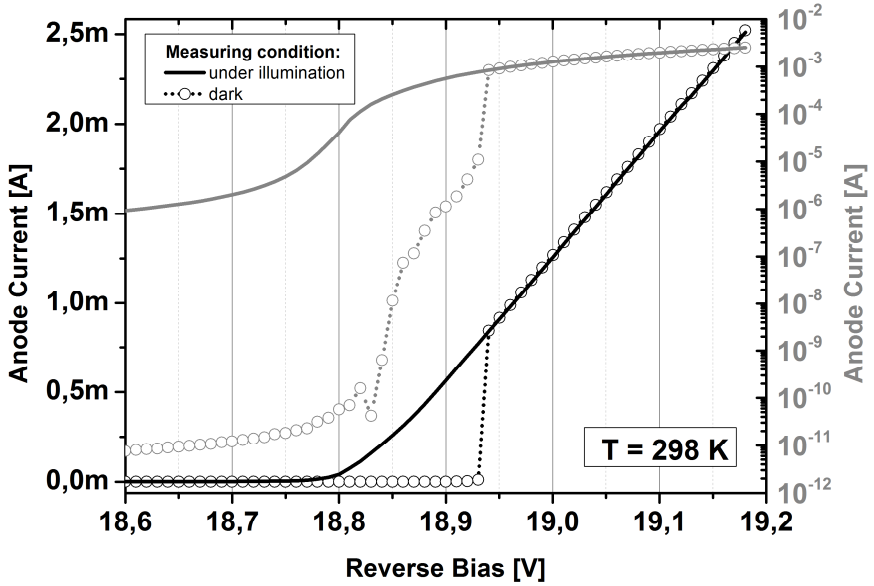

***Figure 2**: Reverse bias I-V curves at T=298 K in both linear (black) and logarithmic (grey) scale. Two different measuring conditions are adopted: device in the dark (symbols) or under a controlled lighting condition (lines).The breakdown voltage is easily extracted in the latter case (lines) thanks to a continuous and clear current transition from micro-ampere values to milli-ampere ones around a breakdown threshold of* $\mathrm{V_{bd}} = 18{,}8\,\mathrm{V}$.

This allowed the extraction of the device breakdown voltage which, in the device under test, is $V_{bd} = 18.8\,V$ at room temperature. The measurement has been then repeated within the same temperature range as for the forward bias case in order to study the thermal behavior of the avalanche diode in the breakdown region. Figure 3 shows that the breakdown voltage of the device increases with temperature. That's because the former is strictly related to the carrier ionization coefficients which in turn decrease with temperature [1]. Basically this implies that a voltage bias providing impact ionization at a certain temperature may not be sufficient to provide ionization at higher temperatures. Therefore the breakdown voltage increases with temperature and, according to Figure 4, the thermal behavior can be well extrapolated by a linear-fit. From the plot, it is possible to extract a breakdown temperature coefficient of around $dV_{bd}/dT = 20.4\,mV/°C$ meaning that by varying the temperature from $-40°$ C to $+40°\,C$, the breakdown voltage varies of just ±4.3% with respect to the room temperature condition.

3.1.2 Electron Hole Pair (HEP) generation study

The saturation current of the avalanche diode can be studied to evaluate the physical nature of the main Electron – Hole Pair (EHP) generation mechanisms occurring in the space charge region of the device. In case of pure thermal carrier generation, the saturation current can be indeed derived according to the Shockley-Read-Hall theory, and it can be expressed as follows [1]:

$$I_{0,sat} \sim T^{3/2}\, e^{-E_g/2kT}$$

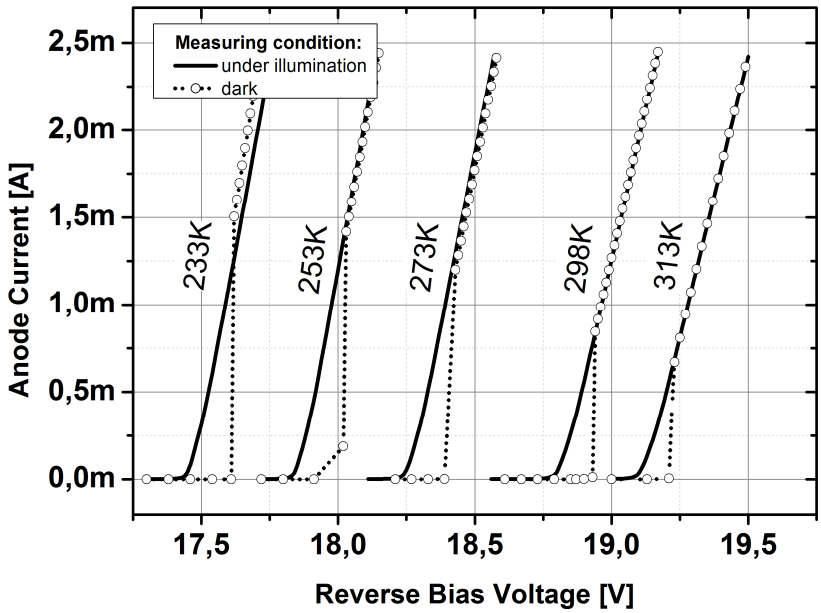

***Figure 3**: Reverse bias I-V curves within the temperature range (*233K – 313K*).*

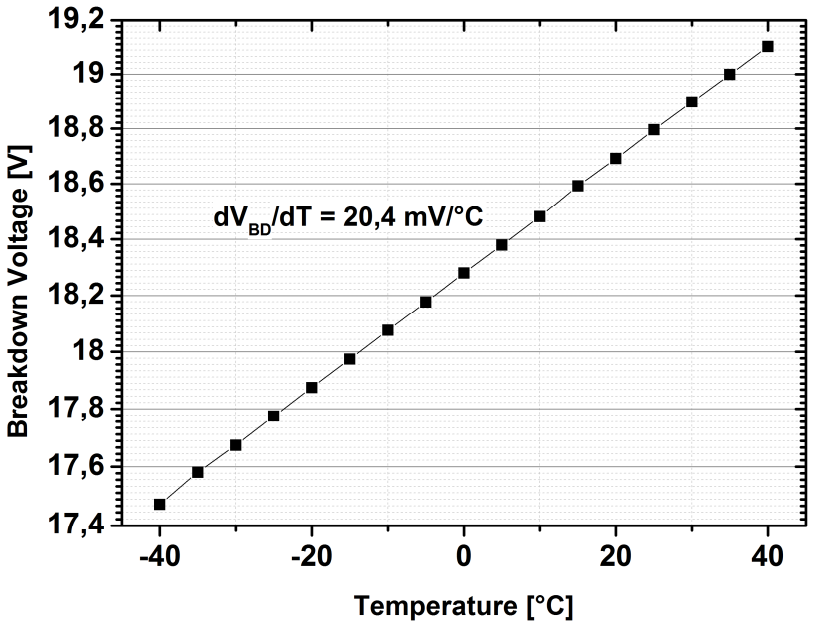

***Figure 4**: Breakdown voltage as a function of the ambient temperature of the p-n junction constituting the active region of the SPADs in a 3D-SiCAD pixel. The breakdown temperature coefficient is* $\mathrm{dV_{bd}/dT} = 20.4\ \mathrm{mV/°C}$.

where T is the temperature, E_g is the silicon band-gap and k is the Boltzmann constant. Therefore the EHP generation mechanisms can be studied with the help of an Arrhenius plot, extracting the activation energy as the slope of the resulting lines. If the saturation current is mainly produced by thermal generation of carriers, than the activation energy should be close to half of the silicon energy band-gap, i.e. $E_g/2 = 0.56\ eV$. If, conversely, field-assisted generation, i.e. trap-assisted tunneling (TAT) or band-to-band tunneling (BBT), concur to the generation of carriers in the device, the resulting saturation current is expected to show weaker dependence on the temperature. This translates into activation energies smaller than $E_g/2$. Figure 5 shows the Arrhenius plots for the measured avalanche diode under three different reverse biases. For temperature values higher than $10\ °C$ the extracted activation energies are $E_A^{13V} = 0.383\ eV$, $E_A^{15V} = 0.336\ eV$ and $E_A^{17V} = 0.337\ eV$ which look rather close to the mid-gap energy. Therefore the EHP generation in the diode space charge region might be due mainly to thermal generation with probably some field-assisted mechanisms such as trap-assisted tunneling. In case of a reverse bias of $17\ V$, field assisted generation seems to start becoming the dominant mechanism for temperature values lower than $10\ °C$, as the extracted activation energy is $E_A^{17\,V} = 0.212\ eV$. Worthless to say, a better analysis of the generation mechanisms occurring in the avalanche diode space charge region would rather consider an Arrhenius plot of the device dark count rate (DCR) instead of the diode saturation current. This latter is indeed including not only the generation mechanisms inside the multiplication region of the SPAD, but also the carrier generated inside the guard-ring region.

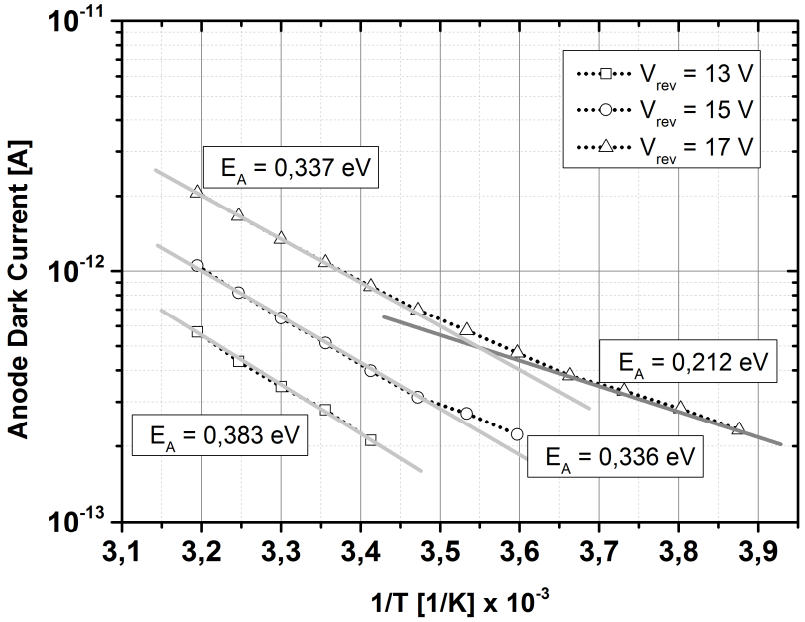

***Figure 5**: Arrhenius plot of the avalanche diode saturation current at different reverse bias voltages.*

3.1.3 Luminescence imaging

An electro-luminescence light emission test has been performed to further validate the compatibility of the adopted avalanche diode architecture for Geiger-mode operation. Hot-carrier induced luminescence is the physical phenomenon responsible for the emission of visible photons within a broad-band spectrum range when a p-n junction is biased above the breakdown threshold. The physical mechanism behind that has been subject to a large debate since its discovery in 1955 by Newman [3]. However the last studies focusing on this topic suggest that carrier energy relaxation between states of the same band (intra-band relaxation) should be the most likely physical process responsible for hot-carrier induced luminescence in silicon [3][4]. Therefore an electro-luminescence light emission picture of an avalanche diode allows evaluating the guard ring effectiveness in preventing premature breakdown at the device periphery and the homogeneity of the electric field distribution all over the device active region, revealing possible electrical weak spots or localized breakdown. When the diode is biased above the breakdown threshold, the hot carriers crossing the junction can lead to a white faint luminescence that is observable even to human eyes. Brighter regions relate to a higher local current density through the junction and thus this is ultimately related to a higher local electric field across the diode active region. The diode under test has thus been biased to operate in Geiger-mode at $V_{rev} = 20\,V$, i.e. $V_{ex} = V_{rev} - V_{bd} = 1.2\,V$ above the breakdown voltage, and observed in the dark with a microscope connected to a scientific CMOS camera with a collection time of 2 seconds. The resulting luminescence picture is even-

tually obtained by averaging ten different acquisitions. Figure 6 refers to the luminescence emission test over the avalanche diode under investigation. The acquired picture (Figure 6b) shows a slightly stronger emission around the top and bottom sides of the diode. Nevertheless it is reasonable to assess that there is a rather good uniformity of the electric field distribution all over the junction and more importantly that there are no clusters of defects since no hot spots are observed. It is worth noticing that the dark ring observed in the middle of the luminescence picture it is not due to a real active area inhomogeneity but it is simply produced by a "shadowing effect" due to the anode metal ring of the diode, which is probably reflecting back the photons emitted right beneath it. No photoluminescence is occurring in proximity of the cathode metal ring, thus implying that premature edge breakdown prevention is effective.

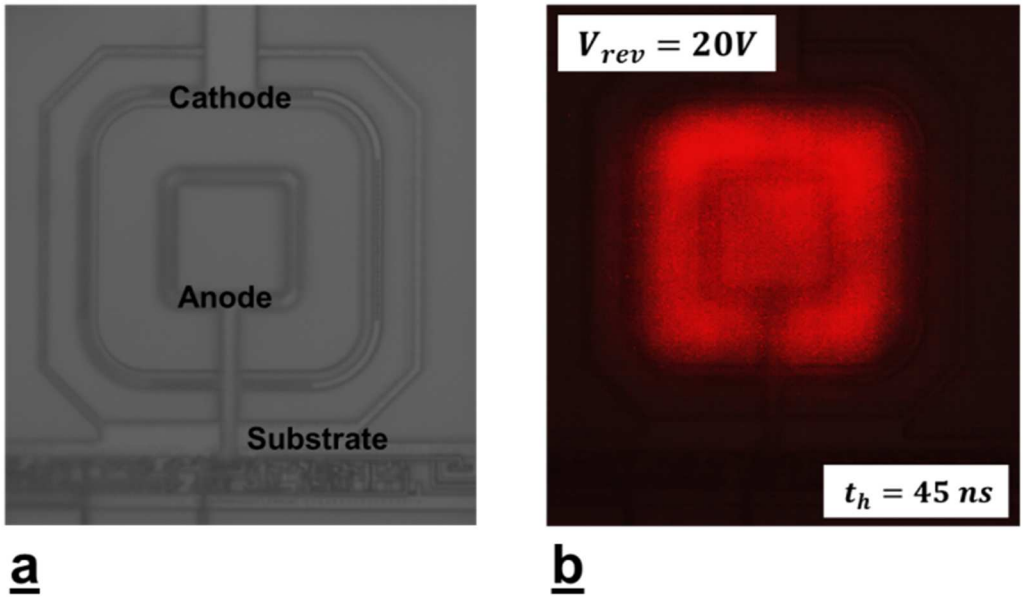

***Figure 6**: Luminescence emission test over the avalanche diode under investigation.*

3.2 Characterization of SPAD pixels in Geiger-mode operation

The dynamic behavior of the different cells on the manufactured test-chips has been characterized with the help of a dedicated testing board allowing to read-out the signals by means of an oscilloscope for a direct waveform analysis, but also to interface the device with an FPGA, for a convenient data acquisition via computer. The measurements have been performed in dark condition at a controlled room temperature of $T_{room} = 25\,°C$. Figure 7 shows a typical output waveform in dark condition of a SPAD pixel (among those available in the cells of the test-chip) consisting of a sequence of digital pulses that are randomly distributed in time. As discussed in Chapter 1, the SPAD is indeed affected by spurious avalanche pulses (i.e. dark counts) that are produced by EHP generation mechanisms in the diode space charge region leading to undesired avalanche ignitions. It is worth noticing that the output pulses are provided in a reversed logic, i.e. an avalanche pulse is represented by a "0" logic level. Figure 7 shows that the device is correctly working in the Geiger-mode since each avalanche ignition is promptly quenched by the electronics and the reverse bias is sharply restored above the breakdown threshold after a certain hold-off time. Each pulse has indeed a well-defined time-width that can be easily adjusted by the user thanks to the hold-off circuit integrated in the SPAD pixel. As discussed in Chapter 1, this feature is very important for a proper analysis of the noise performance degradation due to after-pulsing phenomena, which are strictly related to the time elapsed between the last avalanche ignition and the recharge phase (i.e. the hold-off time). Therefore, the pixel electronics has been first pre-characterized in order to assess a correct functioning of the hold-off circuit presented in Chapter 2 and evaluate the dispersion of the "hold-off time versus voltage" curve within a reasonably large set of SPAD pixels.

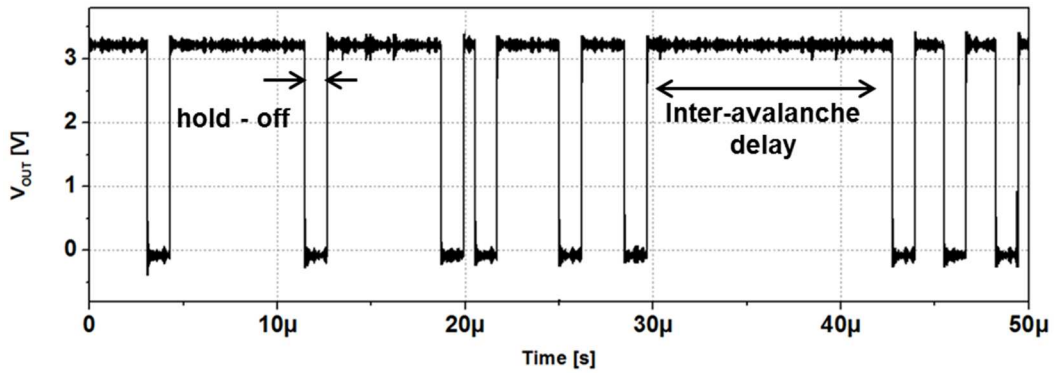

***Figure** 7: Typical SPAD output waveform (dark condition): the spacing between consecutive avalanche pulses, referred to as "inter-avalanche delay", is a statistical parameter obeying a Poisson distribution in absence of after-pulsing. Observe that at around* 20 µs *there may be an after-pulse produced right after a primary avalanche pulse.*

3.2.1 Hold-off circuit characterization

The measurements have been performed over a set of 28 SPAD pixels, from four different test-chips and the results have been compared with Monte Carlo circuit simulations performed during the design phase of the hold-off electronics. According to Figure 8, the measured hold-off time ranges monotonically from around $t_h = 45\ ns$ up around $t_h = 6{,}5\ \mu s$ within the hold-off voltage range going from $V_h = 1{,}6\ V$ to $V_h = 2{,}4\ V$. Interestingly only the simulation scenarios that did not consider the layout parasitics provided a quite reasonable agreement with the measured data, likely due to over-estimated values from the parasitic extraction tool of the simulator. Conversely, the Monte Carlo analysis accounting for process mismatch or both process mismatch and variations, provided a quite good agreement with the measurements up to a hold-off voltage of around $V_h = 2\ V$, while at higher voltages the measured values are smaller than the simulated ones. In reality, the discrepancy between the measured curve and the simulated ones might be simply justified by the statistical variation of the CMOS process parameters. This can be better understood by looking at Table 1, reporting both the measurements and simulation results in terms of mean value and standard deviation. The measured curve falls indeed within the simulated $\pm\sigma$ standard variation bounds around the mean values obtained from the mismatch- and process- dependent Monte Carlo simulation.

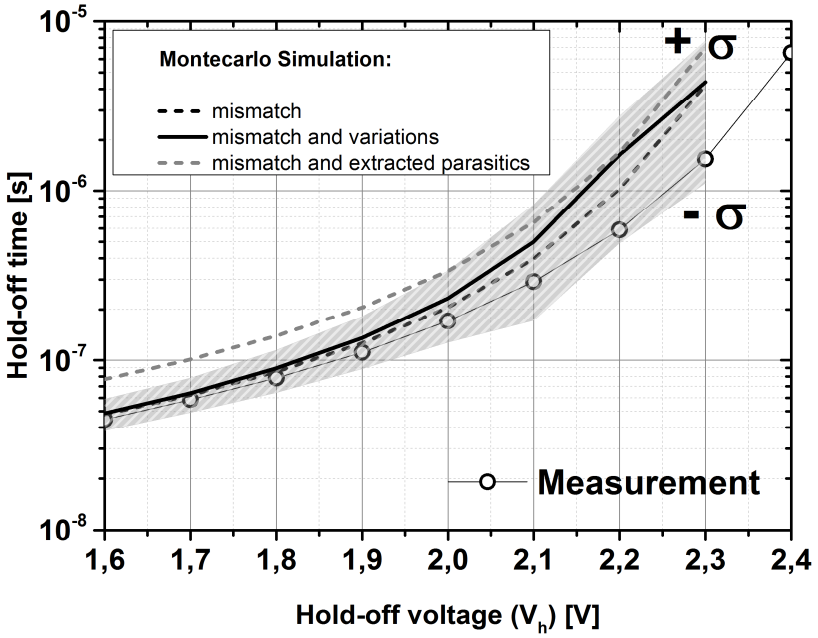

***Figure 8**: Experimental hold-off time as a function of the applied voltage (symbols) compared with Monte Carlo simulation results under three different scenarios: mismatch between devices on the same test-chip with or without considering parasitic extraction (grey and black dashes, respectively); mismatch between devices on the same test-chip and process parameter variations between different test-chips (black line).*

Table I. Hold-off time data from measurements and Monte Carlo simulations.

	Measurements		Monte Carlo Mismatch		Monte Carlo Mismatch and Process	
Hold-off voltage [V]	$\mu(t_h)$	$\sigma_r(t_h)$	$\mu(t_h)$	$\sigma_r(t_h)$	$\mu(t_h)$	$\sigma_r(t_h)$
1,6	44,6 ns	5,1 %	47,8 ns	1,9 %	48,6 ns	21,3 %
1,7	58,5 ns	4,8 %	62,3 ns	1,9 %	63,9 ns	23,2 %
1,8	78,6 ns	5,5 %	85,6 ns	2,2 %	89,7 ns	28,2 %
1,9	111 ns	6,4 %	126 ns	2,6 %	136 ns	34,4 %
2	171 ns	6,3 %	206 ns	3,2 %	232 ns	45 %
2,1	292 ns	7,7 %	397 ns	4,3 %	500 ns	65,6 %
2,2	587 ns	10,5 %	1,03 μs	6,3 %	1,63 μs	70 %
2,3	1,54 μs	15,6 %	4,25 μs	9 %	3,28 μs	75 %
2,4	6,55 μs	26,1 %	N.A.	N.A.	N.A.	N.A.

However the measurements show standard deviation values that are comparable to the ones resulting from a mismatch Monte Carlo analysis, suggesting that the measured devices might come from the same silicon wafer.

3.2.2 Dark Count Rate

The Dark Count Rate (DCR) has been extracted by evaluating the mean time elapsed between two consecutive avalanches (i.e. the mean inter-avalanche delay $< t_{av} >$, see Figure 7) over the available set of SPAD pixels. According to the discussion made in Chapter 1, this value corresponds indeed to the reciprocal value of the DCR, which, in this way, can be easily derived without the need for a "dead-time" correction. The measurement has been performed with the help of a LeCroy HDO6104-MS oscilloscope, thanks to its capability of running fast and continuous measurements of user-defined parameters all over the acquired waveforms. The DCR data has been extracted for several excess bias voltages above the nominal breakdown threshold of $V_{bd} = 18{,}8\,V$ extracted in the I-V curves measurements, while the impact of after-pulsing on the observed dark counts has been evaluated by varying the hold-off time (t_h) within a sufficiently large time-range. Figure 9 shows the "DCR versus t_h" curves of the SPAD cell "A top" in the sample "test-chip_01" for various excess biases V_{ex}. In agreement with the discussion made in Chapter 1, the measured DCR increases with the excess bias voltage, as indicated in Figure 9 by the vertical shift of the curves towards higher values for larger reverse biases.

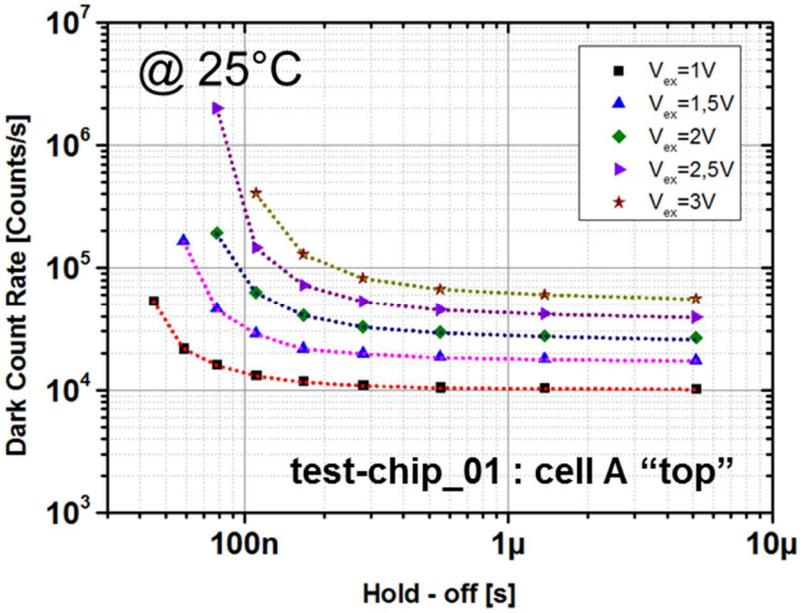

Figure 9: DCR *versus hold-off time* t_h *curves for the SPAD cell "A top" in the sample "test-chip_01".*

The higher electric field across the SPAD multiplication region leads indeed to a higher avalanche triggering probability for the carriers crossing the junction, but also to an enhancement of the field-assisted EHP generation mechanisms. The impact of the after-pulsing on the SPAD noise performance translates into an increasingly larger counting rate at shorter hold-off times: the smaller this time is, the more likely the carriers are released after the diode multiplication capability is restored, resulting in a higher probability to have an avalanche (after-) pulse and thus higher DCR. The effect is actually more severe at higher biases where a larger amount of charge is trapped, due to the increased amount of carriers flowing through the junction within an avalanche cycle. After-pulsing looks indeed negligible for a hold-off time of (roughly) $300\,ns$ in case of an excess bias of $1\,V$, while it is really detrimental on the measured dark count rate for an applied excess bias of $3\,V$. Extensive DCR measurements have been then conducted over a set of 35 SPAD pixels from 5 different test-chips for several excess bias voltages with the aim of analyzing its variability all over the available samples. The measurements have been performed by adopting a long hold-off time of around $5\,\mu s$, in order to ensure a negligible after-pulsing probability. Figure 10 shows the cumulative distribution of the intrinsic *Dark Count Rate density* for different excess bias voltages, obtained by dividing the measured DCR by the diode active surface in order to define a technology dependent parameter expressed in $Hz\,\mu m^{-2}$. In case of $V_{ex} = 1\,V$, the distribution has a median value of around $70\,Hz\,\mu m^{-2}$ (i.e. the value that corresponds to a cumulative percentage of the 50 %), which is in good agreement with respect to what has been obtained by Vilella et al. (see ref. [5], page 111, Picture 4.13 - left) for the measurements of a SPAD detector operated in free-running mode.

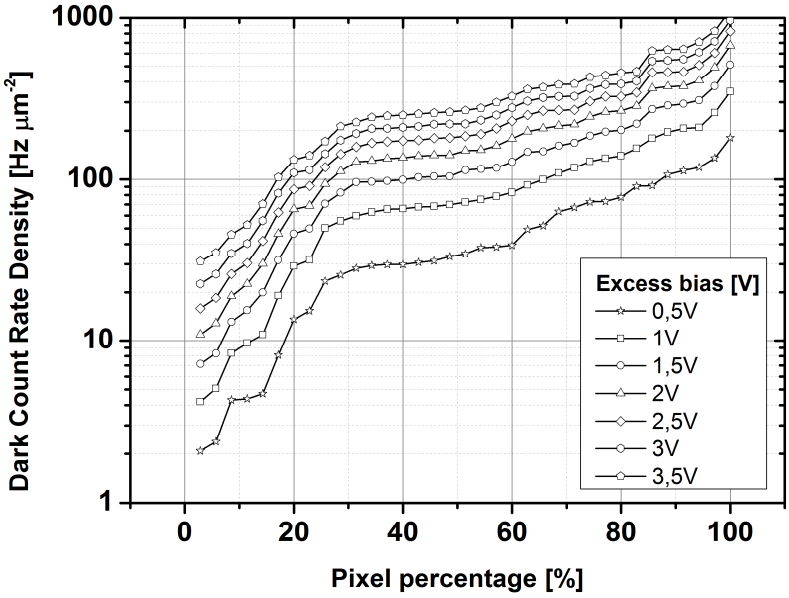

***Figure 10**: Cumulative distribution of the intrinsic Dark Count Rate density for different excess bias voltages, obtained from measurements over a set of 35 SPAD pixels (hold-off time around* 5 µs*, i.e. negligible after-pulsing).*

It is possible to notice that the measured DCR density is affected by a rather severe statistical variation spanning, in case of $V_{ex} = 1\,V$, within two order of magnitude from a few $Hz\,\mu m^{-2}$ up to a few hundreds of $Hz\,\mu m^{-2}$, even if about 40 % of the pixels shows a rather good uniformity close to the median value. The DCR is indeed strictly related to the concentration of impurities and defects in the silicon crystal, which in this case seems to be particularly subject to a strong variability.

3.2.3 Breakdown voltage uniformity

Thanks to the available DCR data over the measured set of SPAD pixels, it is useful to estimate the statistical variation affecting the breakdown voltage of the avalanche diodes. The typical approach normally adopted to estimate the breakdown voltage without performing I-V curve measurements, consists of monitoring the output dark counts as a function of the reverse bias voltage. The voltage sweep is typically performed within the range wherein the breakdown threshold is expected to be found (according, for instance, to a previously obtained I-V curve measurement on a single device). However the pixel electronics employed in the present work is not able to provide counting information in case of excess bias voltages lower than the threshold of the comparator at the input of the hold-off circuit. For this reason, the breakdown voltage has been extracted by extrapolating the available data down to the voltage corresponding to zero counts, as shown in Figure 11.

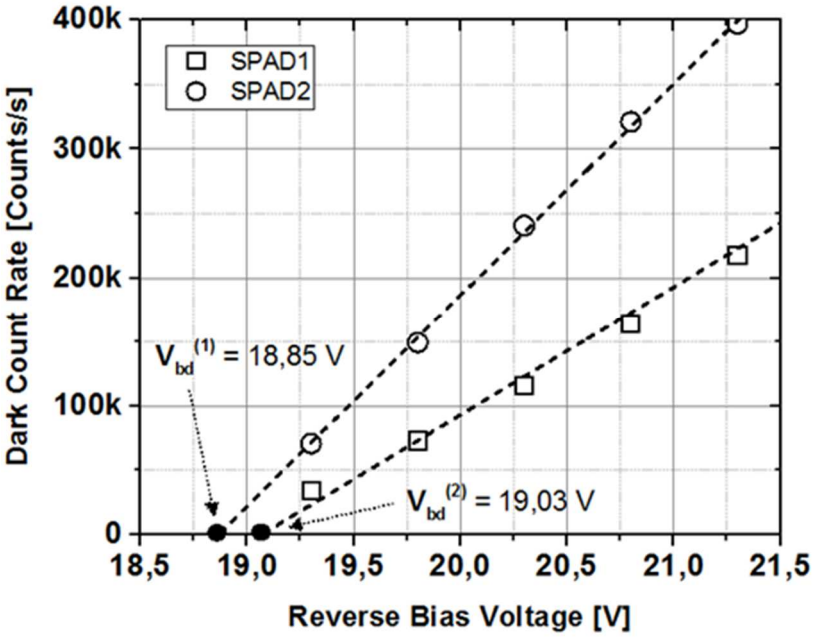

***Figure 11**: Breakdown voltage extraction procedure based on* DCR *versus* V_{rev} *measurements.*

The analysis has been performed over a set of 28 pixels (a larger pixel population would have been more statistically significant) from 4 different test-chips (7 pixels per chip) and the resulting histogram is reported in Figure 12. The breakdown voltage has a mean value $\mu(V_{bd}) = 18{,}77V$ and a standard deviation of $\sigma(V_{bd}) = 13\ mV$.

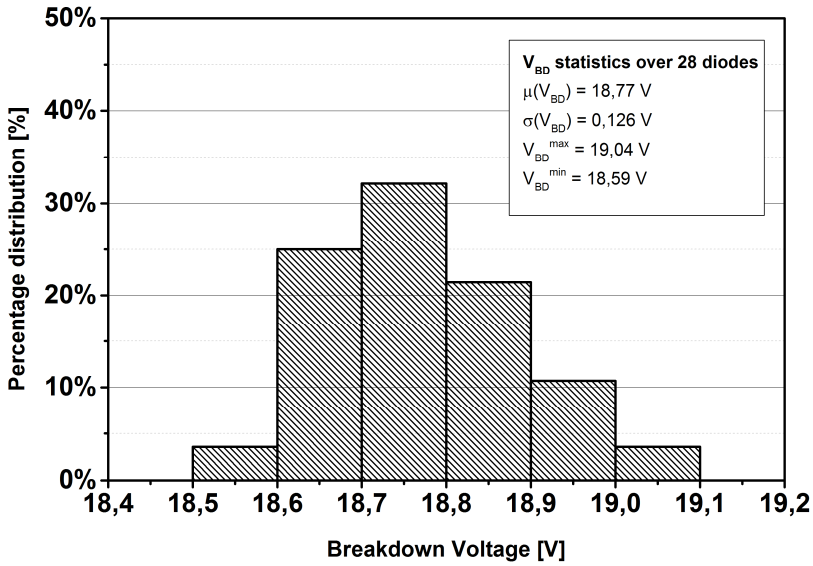

***Figure 12**: Breakdown voltage statistical distribution over a set of 28 measured SPAD pixels.*

3.2.4 After-pulsing

As for the DCR statistical analysis, an extensive after-pulsing experimental study has been performed over a set of 28 pixels from 4 different test-chips for several excess bias voltages with the aim of analyzing its variability all over the available samples. The measurements have been performed in the dark, by adopting a hold-off time of $t_h = 150\,ns$, while the after-pulsing probability has been estimated with the help of equation (15) from Chapter 1, reported here for convenience:

$$\lambda^* \approx \frac{\lambda_0}{1 - P_{ap}} \tag{1}$$

The above formula refers to the overall DCR observed in a SPAD pixel affected by an intrinsic DCR of λ_0 and after-pulsing probability P_{ap}. From an experimental stand-point, the after-pulsing probability can be estimated as follows:

$$P_{ap}(t_h = 150\,ns) \approx 1 - \frac{\lambda_0}{\lambda^*} = 1 - \frac{DCR(long\ t_h)}{DCR(t_h = 150\,ns)} \tag{2}$$

Figure 13 (left) shows the cumulative distribution of the after-pulsing probability for different excess bias voltages. In case of $V_{ex} = 1\,V$, the distribution has a median value of around 12 %, which is actually a rather severe value, considering that this probability will be increasingly higher at shorter hold-off times. Nevertheless, it is possible to notice that there is a rather modest statistical variation on the measured after-pulsing probability for the entire range of adopted excess bias voltages. Figure 13 (right) stresses this fact more effectively, showing that the absolute standard deviation of the obtained after-pulsing probability all over the ensemble is just a few percent.

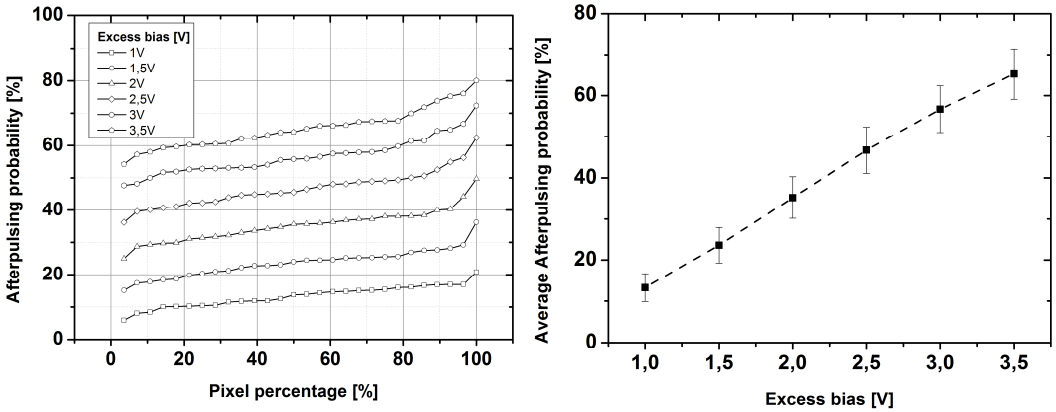

***Figure 13**: Left: cumulative distribution of the after-pulsing probability for different excess bias voltages, when adopting a hold-off time of* $t_h = 150$ ns. *Right: average after-pulsing probability as a function of the applied excess bias voltage.*

As discussed in Chapter 1, the after-pulsing probability increases with the excess bias because of the larger amount of charge flowing through the junction (and thus a larger amount of trapped carrier within an avalanche cycle) but also because of the higher probability that the trapped carriers successfully trigger an avalanche event once they are released by the trap-states.

3.2.4.1 An accurate method for after-pulsing analysis

The after-pulsing effect can be accurately studied by analyzing the time-distribution of the inter-avalanche delay. This distribution represents the relative frequency at which an avalanche event is observed within the time interval between a certain t and $t + dt$ after the last avalanche ignition, and it is indeed the experimental representation of the avalanche probability density function, described by the equation (14) in Chapter 1 and reported here for convenience:

$$p_{av}(t) = \big(\lambda_0 + f(t)\big)e^{-\int_0^t \left(\lambda_0 + f(t')\right)dt'} \tag{3}$$

The formula describes a Poisson process with a combined time-dependent rate parameter $\lambda(t) = \lambda_0 + f(t)$ [6]. The parameter λ_0 is the noise counting rate that is observed in absence of after-pulsing, referred to as *primary dark count rate* while the function $f(t)$ is strictly related to the after-pulsing phenomenon. This latter represents the probability that a carrier is released by the trap levels in the diode space charge region and successfully initiates an avalanche process within t and $t + dt$. Figure 14 shows the experimental histogram of the inter-avalanche delay for the SPAD cell "A bottom" of the sample "test-chip_04" when the diode is reverse-biased at $V_{rev} = 19{,}8\,V$ (at room temperature). The figure considers four different scenarios, depending on the adopted hold-off time, in order to highlight the impact of after-pulsing on the resulting avalanche count distribution. It can be observed that the shorter the adopted hold-off time is, the more likely an avalanche count will occur within a short inter-avalanche delay. Conversely, if a long enough hold-off time is used, the avalanche counts distribute exponentially over time, according to a Poisson statistics with rate λ_0 (black dots in Figure 14). It is worth noticing that whatever the considered scenario is, the tail of the distribution is in general an exponential function with rate λ_0, as the avalanche counts occurring after a delay t^* larger than a few μs, have a primary dark count nature. Once all the trapped carriers are indeed released, $f(t) = 0$ and $\int_0^{t > t^*} f(t')dt' = C = constant$, meaning that equation (3) can be rewritten as $p_{av}(t) = e^{-C}\lambda_0 e^{-\lambda_0 t}$. Interestingly, the attenuation factor e^{-C} could provide a fast way to estimate the after-pulsing probability of the device, as $e^{-C} = 1 - P_{ap}$ (see equation (13) in Chapter 1). In the log-scale plot of Figure 14, this factor can be immediately visualized, as it corresponds to the downwards vertical shift between the exponential tails of one of the avalanche count distribution with after-pulsing and the after-pulsing-free one (long hold-off time). However a much better quantitative evaluation of the after-pulsing effect can be obtained by recalling the way the overall avalanche probability has been defined in Chapter 1:

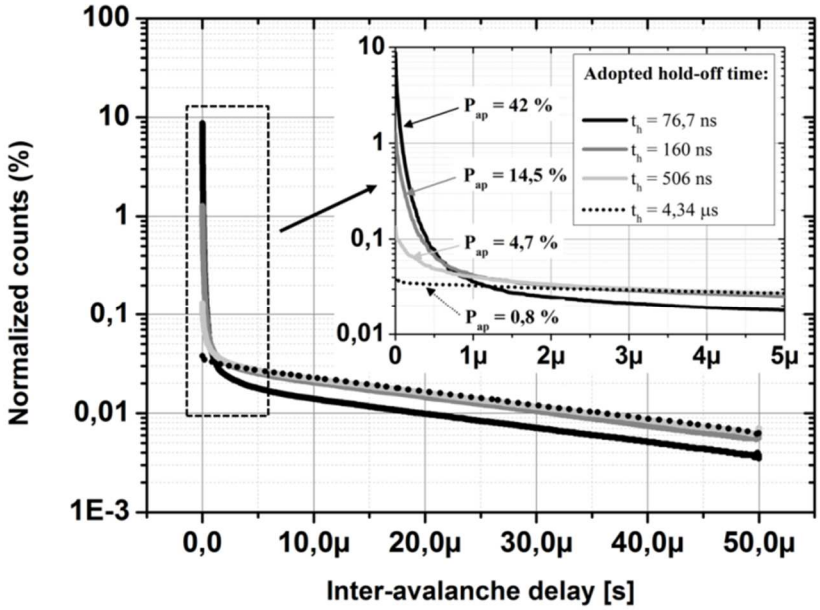

***Figure 14**: Experimental histogram at room temperature (25°C) of the inter-avalanche delay for the SPAD cell "A bottom" of the sample "test-chip_04". Four different scenarios have been considered, depending on the adopted hold-off time, in order to highlight the impact of after-pulsing on the resulting avalanche count distribution. The diode is reverse-biased at* $V_{rev} = 19{,}8$ V.

$$P_{av}(t) = P_{ap}(t) + P_d(t) - P_{ap}(t)P_d(t) \tag{4}$$

The formula expresses the fact that an observed avalanche count can be due either by a primary count $P_d(t)$ or an after-pulse $P_{ap}(t)$, but not to both. The overall avalanche probability is quickly obtained by integrating the experimental histograms of Figure 14, while the primary dark count probability can be easily evaluated by extracting the dark count rate λ_0 in absence of after-pulsing (long hold-off time). This can be done by extrapolating the slope of one of the tails of the avalanche count distribution of Figure 14:

$$P_d(t) = 1 - e^{-\lambda_0 t} \tag{5}$$

The after-pulsing probability is finally obtained as follows:

$$P_{ap}(t) = \frac{P_{av}(t) - P_d(t)}{1 - P_d(t)} \tag{6}$$

Figure 15 shows the resulting avalanche probability $P_{av}(t)$ in case of an adopted hold-off time of $t_h = 76{,}7\ ns$, obtained by integrating the black curve of Figure 14. The two components defining the overall probability are reported in the picture: the primary counts probability calculated using eq. (5) with $\lambda_0 = 28\ kHz$ (dotted line) and the after-pulsing probability extracted with eq. (6) (dashed line).

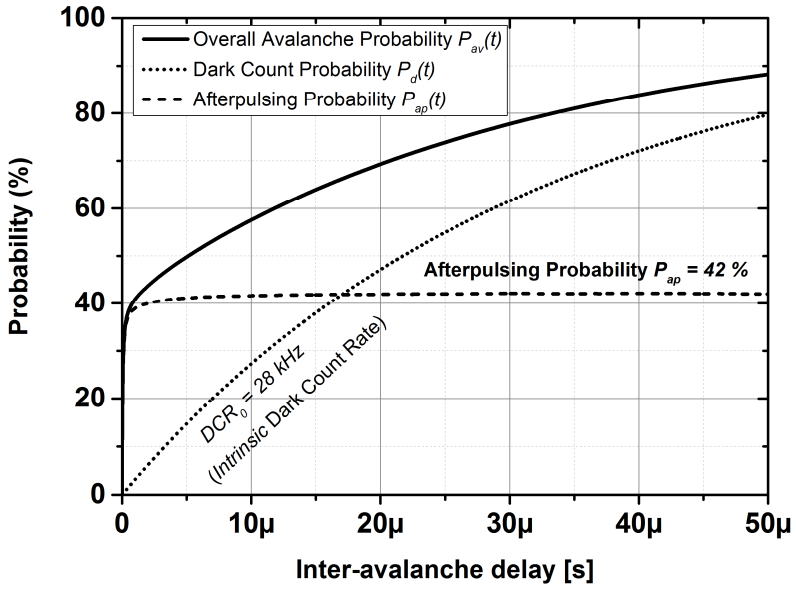

***Figure 15**: Overall Avalanche Probability at room temperature (25°C) for the SPAD cell "A bottom" of the sample "test-chip_04". The diode is reverse-biased at* $V_{rev} = 19{,}8\ V$ *and the adopted hold-off time is* $t_h = 76{,}7$ ns*. The two components defining the overall probability are reported in the figure: the after-pulsing probability (dashes) and the primary counts probability (dots).*

This analysis is rather intuitive since it allows visualizing immediately the weight of the two different physical contributions on the overall observed dark counts. According to Figure 15, the avalanche count probability raises suddenly to a value of about 40 % within the first few hundreds of nanoseconds, meaning that almost half of the observed dark counts occur within a sub-microsecond time-range. All these "fast" counts are caused indeed by the intense after-pulsing occurring in the device within the considered time-interval, as shown by the dashed curve, and quantified by an overall after-pulsing probability of $P_{ap} = 42$ % (in agreement with the results obtained by Vilella et al. [5], i.e. $P_{ap} = 52$ % in case of a $t_h = 50\ ns$ at $V_{ex} = 1\ V$). The probability to have primary counts within the sub-microseconds time-interval is indeed very low, i.e. less than a few %, but this latter is responsible for the remaining 60 % of dark counts occurring mainly at longer delays. By following this approach, the after-pulsing probability of the SPAD under study has been calculated for different hold-off times, as reported in Table II. The results show that the measured SPAD is affected by a dramatically high after-pulsing probability that can be reduced to a more acceptable value of $P_{ap} = 4{,}7$ % only for a hold-off time $t_h = 506\ ns$. This might represent an important limiting factor if the SPAD has to be implemented in high counting rate detection systems. Moreover, a comparison between the two after-pulsing evaluation methods adopted in this Chapter is proposed in Table II, showing that the "fast" one (that has been used in the previous section for a fast after-pulsing evaluation all over the measured samples, according to equation (2)) provides a fair-good quantitative estimation of the phenomenon.

Table II. Comparison between the two methods adopted for the evaluation of the after-pulsing probability.

Hold-off time	$\boldsymbol{P_{ap}}$ (accurate method)	$\boldsymbol{P_{ap}}$ (fast method)
76,7 ns	42 %	42,2 % ($DCR = 49\ kHz$)
106,8 ns	26,3 %	25,5 % ($DCR = 38\ kHz$)
160 ns	14,5 %	16,4 % ($DCR = 33{,}9\ kHz$)
506 ns	4,7 %	5,1 % ($DCR = 29{,}85\ kHz$)
4,34 μs	0,8 %	0 % ($DCR_{ref} = \lambda_0 = 28{,}3\ kHz$)

3.2.4.2 Deep levels characterization

As discussed in Chapter 1, the after-pulsing effect is strictly related to the presence of trapping centers in the deep energy level range of the depletion region, responsible of capturing a part of the free carriers flowing though the junction during an avalanche ignition and, consequently, releasing them with a statistical fluctuating delay that might be longer than the adopted hold-off time. This is described by the function $f(t)$ introduced in Chapter 1 in order to define equation (13) which is reported here for convenience:

$$P_{ap}(t) = 1 - e^{-\int_0^t f(t')dt'} \tag{7}$$

The function $f(t)$ represents indeed the probability that a trapped carrier is released by one of the traps and successfully initiate an avalanche process within t and $t + dt$. The study of this function can thus provide a better understanding of the physical nature of the after-pulsing effect [7]. This function can be re-written as follows:

$$f(t) = \frac{p_{ap}(t)}{1 - P_{ap}(t)} \tag{8}$$

where $p_{ap}(t)$ is the afterpulsing probability density function, obtained by a simple first derivative versus time of the afterpulsing probability $P_{ap}(t)$, derived in the previous analysis. Figure 16 shows the resulting $f(t)$ curve (grey) together with the after-pulsing probability density one (black), extracted from the study described in the previous paragraph and thus in case of an overall after-pulsing probability of 42 %. It can be observed that the distribution of the carrier detrapping time shows a slower decay with respect to the after-pulsing one. The collected counts for this latter are due indeed to the released carrier that "first" fires an avalanche process [8][9], thus ignoring other possible carrier releases that follow the first detected one. This discrepancy is in fact expressed in (8) by the factor $1 - P_{ap}(t)$. The trap-release becomes negligible with respect to the dark counts only for delays larger than $\sim 5\ \mu s$. Further analysis could be conducted over this topic, but this has not been investigated in the present work.

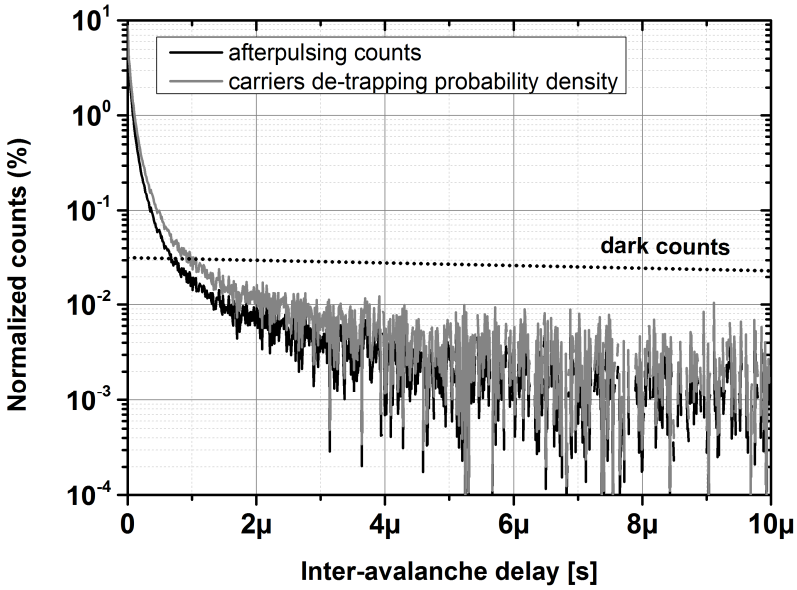

***Figure 16**: Distribution of the trapped-carrier release-time (grey curve) together with the after-pulsing probability density function (black curve).*

3.2.5 Photon Detection Efficiency

The single-photon detection capability of the SPADs under study has been tested by measuring the photon detection efficiency (PDE) within the spectral range $400\,nm - 1000\,nm$. As discussed in Chapter 1, the PDE is defined as the probability that a photon of a certain wavelength hitting a SPAD pixel, successfully fires a self-sustained avalanche multiplication process [10].

The optical set-up adopted for this measurement is based on a broadband and stable light source, a monochromator for the wavelength range of interest, optical filters, and an integrating sphere to obtain a uniform light beam over the SPAD under test and a calibrated photodiode, as shown in Figure 17. The calibrated photodiode exposed to the same photon flux, has been employed to precisely evaluate the optical power sent over the SPAD pixel. In order to perform single-photon measurement, it is necessary to ensure that there is on average only one photon being absorbed within the SPAD active region at a time, by keeping the optical power onto the considered pixel sufficiently low. Moreover, the counting rate R_m measured at the detector output includes not only the number of the detected photons $R_{ph}PDE$ impinging on the pixel with rate R_{ph}, but also the dark counts, occurring with rate DCR, and more generally, the additional count enhancement due to the afterpulsing effect by a factor $1/(1 - P_{ap})$ as discussed in Chapter 1:

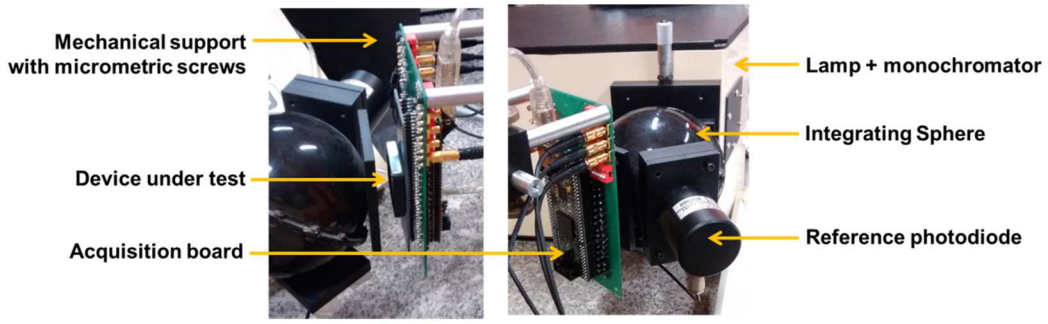

Figure 17: *Experimental setup for the measurement of the Photon Detection Efficiency (PDE). Measurements performed at ICube Lab (Strasbourg, France)*

$$R_m = \frac{R_{ph} PDE}{(1 - P_{ap})} + \frac{DCR}{(1 - P_{ap})} = \frac{R_{ph} PDE}{(1 - P_{ap})} + DCR^* \tag{9}$$

where DCR^* represents the overall DCR in presence of after-pulsing. The presence of a non-renewable dead-time (i.e. the hold-off time) may affect the device linearity with respect to the input optical power. More specifically, a counting rate $R_m(observed)$ lower than the one predicted by (9) can be observed at the device output, depending on the amount of counting losses [11] occurring during the dead-time:

$$R_m(observed) = \frac{R_m}{1 + R_m t_h} \tag{10}$$

According to (10), if the probability to have a counting loss during the dead-time $R_m t_h$ is not negligible, the observed counting rate is lower than expected. Therefore, the corrected counting rate related to true detection of an incoming photon is given by:

$$R_{ph} PDE = \left(\frac{R_m(observed)}{1 - R_m(observed) t_h} - \frac{DCR^*(observed)}{1 - DCR^*(observed) t_h} \right) (1 - P_{ap}) \tag{11}$$

A linearity measurement on the device under test has been thus conducted in order to find the optimal power range that guarantees a linear behavior for the device in case of a hold-off time of $t_h = 178\,ns$. Non-linearity can indeed arise not only from dead-time losses (i.e. the hold-off time) but also from an excessively high optical power which might produce a pile-up of absorbed photons. Figure 18 shows the result of the linearity test over a SPAD cell, obtained by adopting a reverse bias voltage of $V_{rev} = 19{,}8\,V$ and at a wavelength of $\lambda = 450\,nm$. The overall measured counting rate at the SPAD output $R_m(observed)$ has been plotted as a function of the incoming photon rate (optical power over the pixel). According to the picture, the non-linearity starts to appear for a photon rate higher than 20 $MCounts/s$.

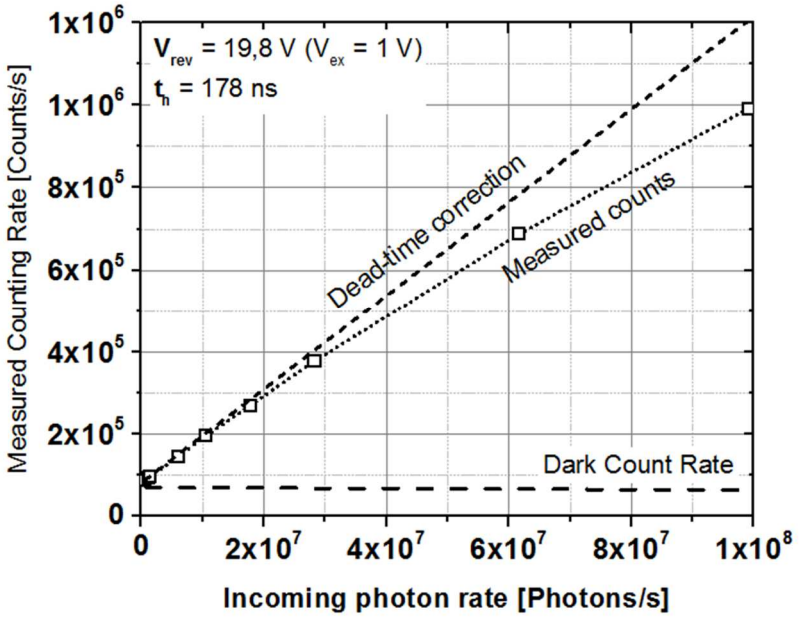

Figure 18: *Linearity test over a SPAD pixel for* $\lambda = 450$ nm *and by adopting a reverse bias voltage of* $V_{rev} = 19{,}8$ V *and a hold-off time of* $t_h = 178$ ns.

This non-linearity is due to dead-time losses only, as it can be seen by the dashed line reporting the output counting rate resulting after dead-time correction, based on (10). The corrected data show indeed excellent linearity which makes possible to consider pile up events negligible (i.e. two or more photons hitting the pixel simultaneously). Photon detection efficiency (*PDE*) has been then measured for two different SPAD pixels within the spectral range $400\,nm - 1000\,nm$, setting the hold-off time at $t_h = 178\,ns$ and considering two different excess bias voltages, i.e. $V_{ex} = 1\,V$ and $V_{ex} = 2\,V$. The resulting curves in Figure 19 show that the photon detection capability of the device is rather poor. The measured PDE reaches indeed the maximal value of 2,7 % and 4,5 % (depending on the adopted bias) around $\lambda = 600$ nm. Even if this result definitely appears quite below the expectations, it is worth noticing that it is in good agreement with what has been reported in the work of Vilella et al. [5]. The cause of these results was attributed to the polyimide passivation of the adopted CMOS process, which dramatically reduces the optical transparency of the dielectric material in the circuit back-end [5][12]. The PDE has been then evaluated as a function of the reverse bias across the avalanche diodes, by lighting the devices at the optimal wavelength $\lambda = 600$ nm. The results of this measurement are reported in Figure 20. The graph proposes an interesting comparison between the "rough" PDE (symbols, left axis) and the corrected one (solid lines, left axis). In the first case, the PDE is calculated as the ratio between the measured counting rate $R_m(observed)$ minus the observed Dark Count Rate $DCR^*(observed)$, and the overall incoming photon rate R_{ph}. In the second case, the PDE has been calculated with the help of equation (11), which required the measurement of the after-pulsing probability all over the considered reverse bias voltage range (dotted lines, right axis).

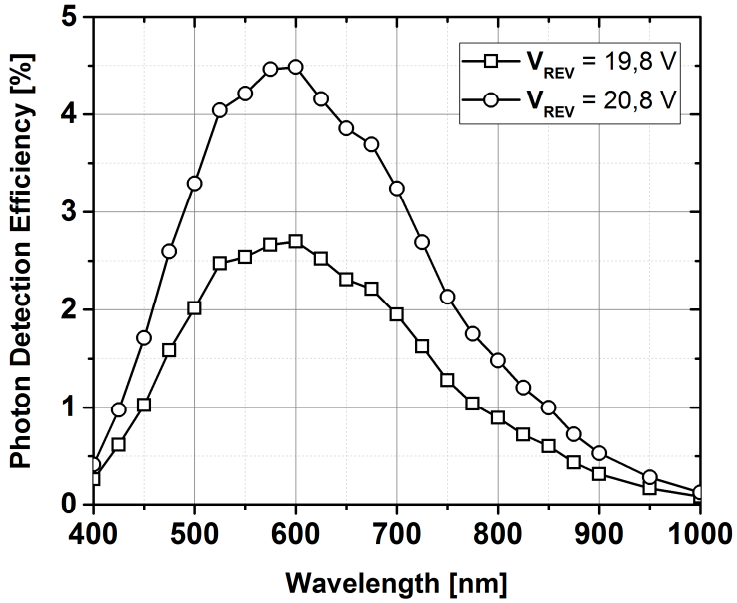

Figure 19: *Photon Detection Efficiency (PDE) of two SPAD pixels within the spectral range* 400 nm – 1000 nm*, adopting a hold-off time at* $t_h = 178$ ns *and considering two different excess bias voltages, i.e.* $V_{ex} = 1$ V *and* $V_{ex} = 2$ V.

It is interesting to observe that the curves corresponding to the corrected *PDE*, flatten for larger reverse biases, probably indicating that the avalanche triggering probability approaches the 100% value.

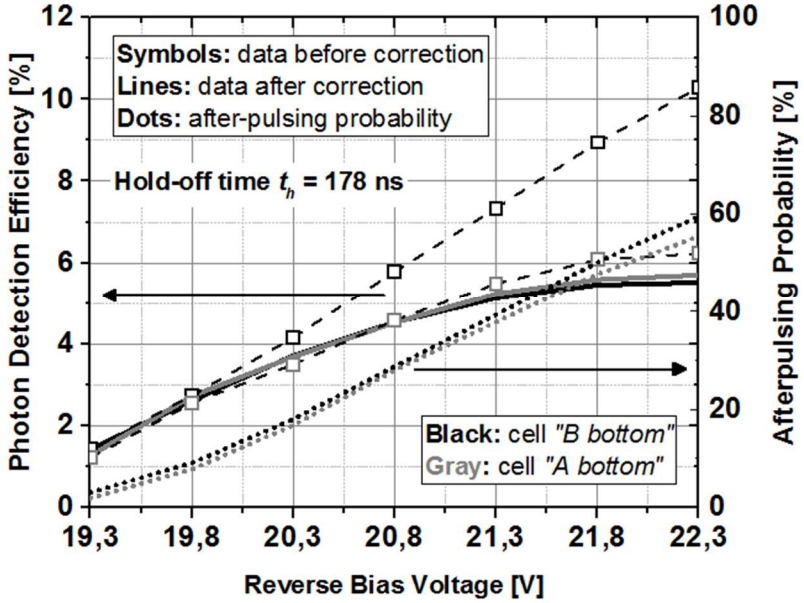

Figure 20*: Left axis: photon detection efficiency (PDE) as a function of the reverse bias across the avalanche diodes, by lighting the devices at the optimal wavelength* $\lambda = 600$ nm*. Right axis: after-pulsing probability as a function of the reverse bias.*

Conclusions

The SPAD cells constituting the sensing levels of a 3D-SiCAD pixel have been fully characterized and the results have been presented and critically analyzed in this chapter. The avalanche diode architecture adopted in the present work has been validated thanks to an electro-luminescence test showing rather good uniform light emission intensity without hot spots all over the SPAD active region and thus indicating a uniform electric field distribution all along the junction. The avalanche diode *I-V* curves and the corresponding breakdown voltage have been evaluated within the temperature range from $-40\,°C$ to $40\,°C$, allowing the study of the electron – hole pairs generation mechanisms occurring in the device active region by means of Arrhenius plot. Electrical characterization over several pixels validated the correct functioning of the quenching electronics but revealed that the dark counts are affected by a rather severe statistical variation, showing, in case of $V_{ex} = 1\,V$, a median value of around $70\,Hz\,\mu m^{-2}$. Afterpulsing measurements, in case of $t_h = 150\,ns$ and $V_{ex} = 1\,V$, showed a median value of around 12 %, which is a rather severe value, considering that this probability will be increasingly higher reverse biases and at shorter hold-off times. Photon detection efficiency (*PDE*) measurements within the spectral range $400nm - 1000nm$, in case of $V_{ex} = 2\,V$ showed a maximal value of less than 5 % at $\lambda = 600\,nm$, indicating a poor photon detection capability. The cause of this result can be attributed to the polyimide passivation of the adopted CMOS process, which dramatically reduces the optical transparency of the dielectric material in the circuit back-end [5][12].

The following chapter will address the characterization of a single 3D-SiCAD pixel from a first 3D prototype with the aim of demonstrating the expected capability in rejecting background counts and in detecting ionizing particles with an excellent efficiency.

References

[1] S. M. Sze and K. Ng, *Physics of Semiconductor Devices*, 3rd Editio. 2006.

[2] F. Villa, D. Bronzi, Y. Zou, C. Scarcella, G. Boso, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, U. Paschen, and W. Brockherde, "CMOS SPADs with up to 500 μm diameter and 55% detection efficiency at 420 nm," *J. Mod. Opt.*, vol. 61, no. 2, pp. 102–115, 2014.

[3] Sidi Aboujja, "Électroluminescence en avalanche des jonctions p-n à base de silicium et d'arséniure de gallium, et effet d'irradiation," Université de Sherbrooke, 2000.

[4] A. L. Lacaita, F. Zappa, S. Bigliardi, and M. Manfredi, "On the bremsstrahlung origin of hot-carrier-induced photons in silicon devices," *IEEE Trans. Electron Devices*, vol. 40, no. 3, pp. 577–582, Mar. 1993.

[5] E. V. Figueras, "Feasibility of Geiger-mode avalanche photodiodes in CMOS standard technologies for tracker detectors Feasibility of Geiger-mode avalanche photodiodes in CMOS standard technologies for tracker detectors," (PhD Thesis), University of Barcelona, 2013.

[6] K. E. Jensen, P. I. Hopman, E. K. Duerr, E. a. Dauler, J. P. Donnelly, S. H. Groves, L. J. Mahoney, K. a. McIntosh, K. M. Molvar, a. Napoleone, D. C. Oakley, S. Verghese, C. J. Vineis, and R. D. Younger, "Afterpulsing in Geiger-mode avalanche photodiodes for 1.06 μm wavelength," *Appl. Phys. Lett.*, vol. 88, no. 13, pp. 27–30, 2006.

[7] S. Cova, a. Lacaita, and G. Ripamonti, "Trapping phenomena in avalanche photodiodes on nanosecond scale," *IEEE Electron Device Lett.*, vol. 12, no. 12, pp. 685–687, 1991.

[8] M. Wahl, "Time-correlated single photon counting," pp. 1–14, 2014.

[9] a. C. Giudice, M. Ghioni, S. Cova, and F. Zappa, "A process and deep level evaluation tool: afterpulsing in avalanche junctions," *ESSDERC '03. 33rd Conf. Eur. Solid-State Device Res. 2003.*, pp. 347–350, 2003.

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

[11] E. Sciacca, A. C. Giudice, D. Sanfilippo, F. Zappa, S. Lombardo, R. Consentino, C. Di Franco, M. Ghioni, G. Fallica, G. Bonanno, S. Cova, and E. Rimini, "Silicon planar technology for single-photon optical detectors," *IEEE TED*, vol. 50, no. 4, pp. 918–925, 2003.

[12] 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," *Proc. SPIE 6372, Adv. Phot. Count. Tech. 63720S*, vol. 6372, p. 63720S–63720S–12, 2006.

Chapter 4: Characterization of a 3D-SiCAD prototype

In this chapter, a 3D-SiCAD pixel cell is finally characterized with the aim of validating the device capability in the rejection of the noise counts affecting the SPAD sensing levels, and, more importantly, the detection of ionizing particles.

4.1 Noise of a 3D-SiCAD pixel

The noise performance of the 3D-SiCAD pixel has been characterized by adopting the same experimental setup described in Chapter 3. In a first phase, preliminary measurements have been conducted over two adjacent "in-plane" avalanche diodes operated in coincidence-mode (Figure 1, left) with the aim of validating the coincidence electronics and demonstrating the noise rejection capability provided by the coincidence detection mode. The noise performance of a 3D-assembled prototype (Figure 1, right) has been subsequently measured and studied.

4.1.1 Preliminary coincidence-mode measurements

Figure 2 shows the measured amount of fake coincidence hits recorded between two adjacent SPAD pixels in dark condition shown in Figure 1 (left), for different reverse bias voltages, within a wide range of coincidence time-windows, which were easily adjustable thanks to a dedicated software driving an FPGA.

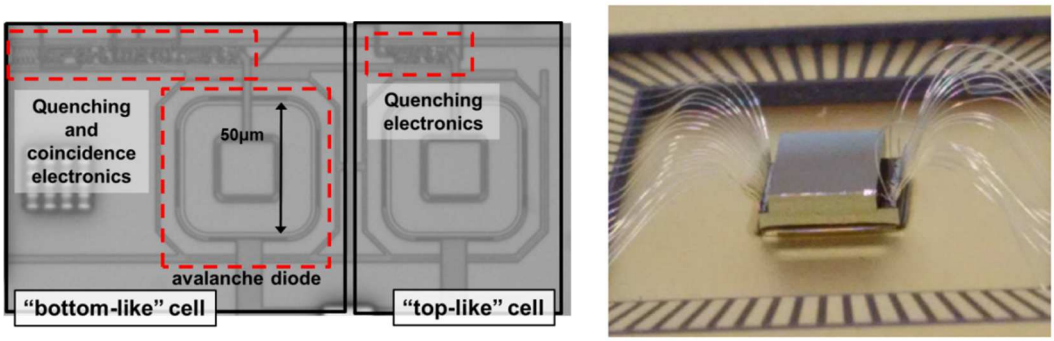

Figure 1: *Microphotograph of two adjacent SPAD cells and their associated electronics (quenching and coincidence detection). This structure has been used for a preliminary assessment of the noise rejection capability of the coincidence detection mode. (b) Microphotograph of the prototype showing the two stacked dies and the wire bondings on the left and right edges.*

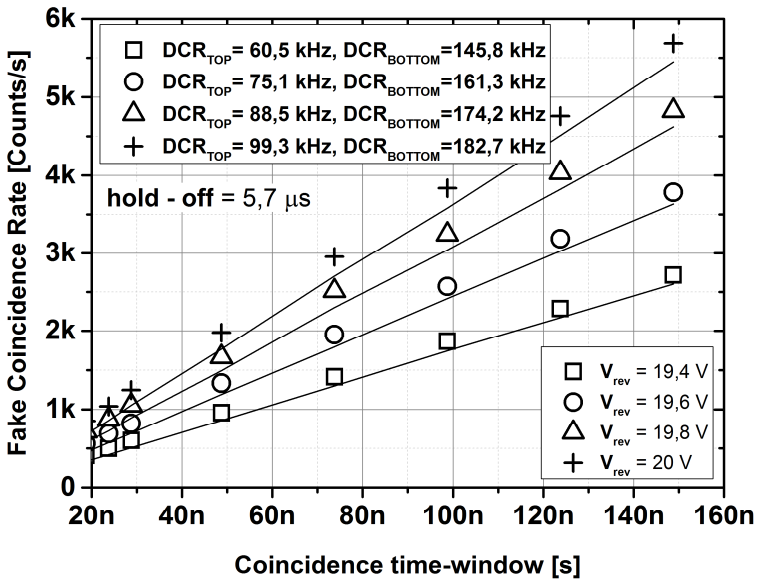

Figure 2: *Fake Coincidence Rate (FCR) recorded between two adjacent SPAD pixels in dark condition, as a function of the adopted coincidence time-window, for different reverse bias voltages. Symbols refers to measured data, while the lines represent the FCR predicted according to equation (17) from Chapter 1.*

It is worth noticing that the measured data (symbols) match very well with the counting rate (lines) predicted according to equation (17) from Chapter 1, reported here for convenience:

$$FCR = 2 \cdot DCR_{top} \cdot DCR_{bottom} \cdot \Delta t \qquad (1)$$

where DCR_{top}, DCR_{bottom} refers to the "observed" dark count rate (see Section 1.1) in the top and bottom SPAD device, respectively, while Δt is the coincidence time window. The recorded number of counts rises indeed linearly with respect to the adopted coincidence time-window. Moreover the increase of the reverse bias voltage produces a larger amount of dark counts, which in turn translates into a steeper slope and thus a higher Fake Coincidence Rate (FCR) value. A similar measurement is reported in Figure 3, but in this case the *hold-off time* provided by the quenching circuit were varied from $t_h = 165\,ns$ to $t_h = 5{,}6\,\mu s$, while keeping the reverse bias of the avalanche diodes at $V_{rev} = 19{,}8\,V$. Longer *hold-off times* result in an overall lower amount of dark counts for the SPAD*s,* as this reduces the after-pulsing probability, on the one hand, and limits the device counting capability, on the other hand, due to the larger dead-time imposed at every avalanche ignition. Therefore, the FCR curves feature a steeper slope at shorter *hold-off time,* as expected. In both cases represented in Figure 2 and Figure 3, it is anyway interesting to observe that the FCR curves approach the zero counting rate value as the coincidence time-window becomes smaller and smaller. This indicates that a great rejection of the noise counting rate can be achieved, even in case of very noisy *SPAD* pixels, as it is in the present case, provided that a sufficiently small time-window is adopted.

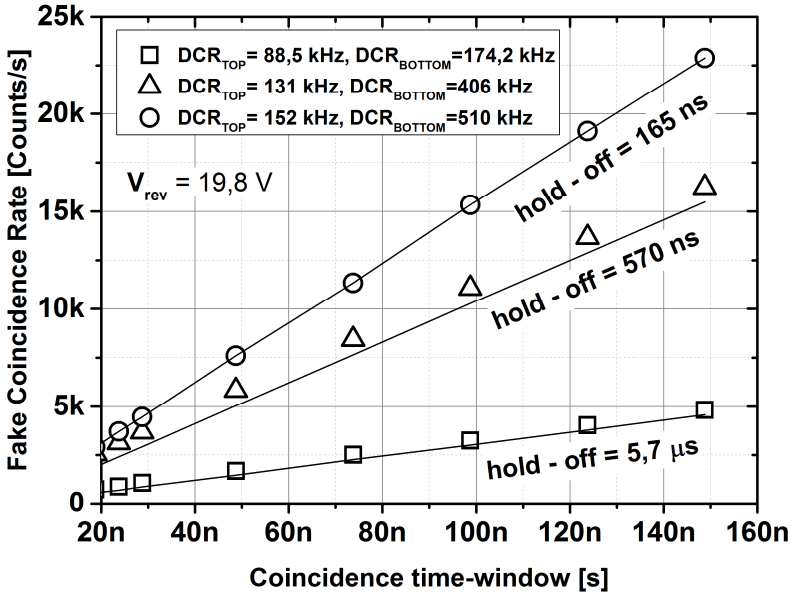

Figure 3: *Fake Coincidence Rate (FCR) recorded between two adjacent SPAD pixels in dark condition as a function of the adopted coincidence time-window, for three different hold-off times ($V_{rev} = 19.8\,V$). Symbols refers to measured data, while the lines represent the FCR predicted according to equation (17) from Chapter 1.*

The DCR measured for each single pixel is indeed in the range of $100\,kHz - 500\,kHz$ (depending on the adopted reverse bias / *hold-off time*) while, for a coincidence time-window of $20\,ns$, the recorded rate of fake coincidences falls within the range of $500Hz - 5kHz$, i.e. 2-3 orders of magnitude lower than the corresponding DCR. This clearly demonstrates the noise rejection capability achievable in the implemented coincidence detection mode, even if the results are apparently far to be close to the optimal achievable value. Less noisy SPAD*s*, i.e. a cleaner CMOS process, and a factor 10 shorter coincidence time-window, would provide indeed a great improvement on the overall FCR.

4.1.2 3D-SiCAD prototype

The noise performance in the dark of a single 3D-SiCAD pixel cell from the 3D prototype shown in Figure 1 (right) has been finally characterized. Figure 4 shows the measured FCR (symbols) as a function of the adopted coincidence time-window, for different reverse bias voltages, while adopting a *hold-off time* of $t_h = 2{,}5\,\mu s$. Table I reports the measured DCR of each sensing level of the 3D-SiCAD pixel, for the different applied reverse bias voltages.

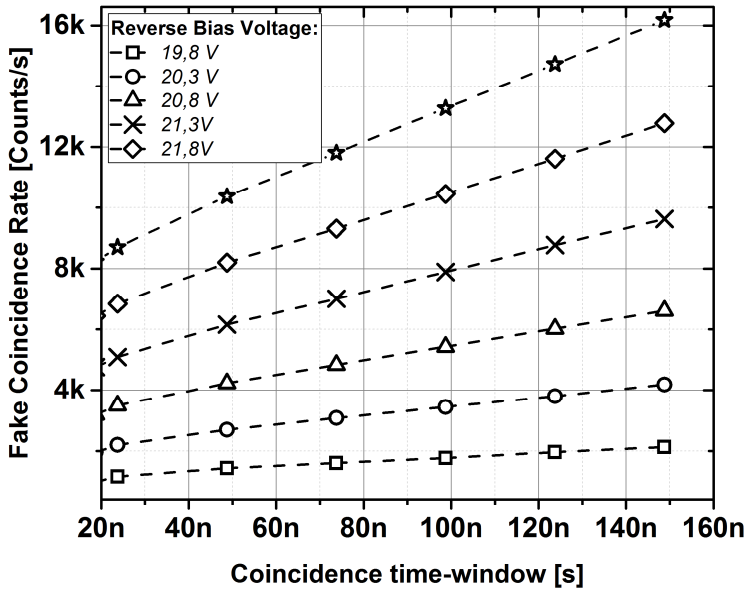

Figure 4: *Fake Coincidence Rate (FCR) recorded in a 3D-SiCAD pixel in dark condition, as a function of the adopted coincidence time-window, for different reverse bias voltages. Symbols refers to measured data (dashed lines are simply interpolating the measured data). Observe that the adopted hold-off time is* $t_h = 2{,}5\ \mu s$.

Even though in this case the SPAD cells show a lower DCR with respect to the ones considered in the preliminary measurements of section 4.1.1, the recorded FCR for a time-window of $20\,ns$, falls within the range of $1\,kHz - 10kHz$, which is a factor 10 larger than the expected one, i.e. on the order of $100\,Hz - 1kHz$. In contrast to what is observed in Figure 2, the FCR curves reported in Figure 4 seem indeed to be affected by a certain offset, which in turn seems to be related to the adopted reverse bias voltage. The measured FCR still rises linearly with respect to the adopted coincidence time-window, but the curves do not converge to a zero counting rate value towards smaller coincidence time-window, as it would have been expected. The observed offset might be due to actual coincidence events occurring for some reason in the 3D pixel.

Table I. DCR values of the SPAD cells in the 3D-SiCAD pixel

Reverse Bias	DCR Top Cell	DCR Bottom Cell
19,8 V	44 kHz	83 kHz
20.3 V	65 kHz	118 kHz
20.8 V	85 kHz	147 kHz
21.3 V	104 kHz	174 kHz
21.8V	123 kHz	200 kHz
22.3V	140 kHz	222 kHz

4.1.3 Discussion

In order to gain a better understanding on this result, the FCR has been measured by using the coincidence circuit integrated in the 3D-SiCAD pixel, as it is capable to resolve much shorter coincidence time-windows if compared to the external FPGA acquisition system. The former has been first pre-characterized by evaluating the coincidence time-window Δt as a function of the applied coincidence voltage V_c. The measurement procedure has been performed with the help of an oscilloscope, and consisted on evaluating the delay between the avalanche leading edges of the output pulses arising from the two sensing levels of the 3D-SiCAD pixel. This measurement was performed at each occurrence of a coincidence event, by setting the reference trigger for the waveform acquisition on the output signal arising from the coincidence circuit. The measured delays were then collected in a histogram, whose overall time-width Δt_{hist} allowed extracting the coincidence time-window as $\Delta t = \Delta t_{hist}/2$. The longest acquired delay between the two avalanche leading edges related to a given coincidence time-window cannot be, indeed, longer than this latter value. Since this procedure is rather time-consuming, the measurement was performed on a single cell only. The result of this measurement is reported in Figure 5, together with the data arising from circuit simulation results. The coincidence voltage has been swept from $V_c = 0$ V to $V_c = 2{,}3$ V and the measured coincidence time-window (symbols) varied from $\Delta t = 490\, ps$ up to $\Delta t = 106\, ns$. The experimental data match within 20 % with the simulated ones up to $\Delta t = 10\, ns$. At larger values the simulation results differ pretty much from the measurements. Interestingly if the curve obtained from simulation is voltage-shifted by just $\Delta V = 0{,}1\, V$ a much better match within 10 % is obtained, in the entire voltage range.

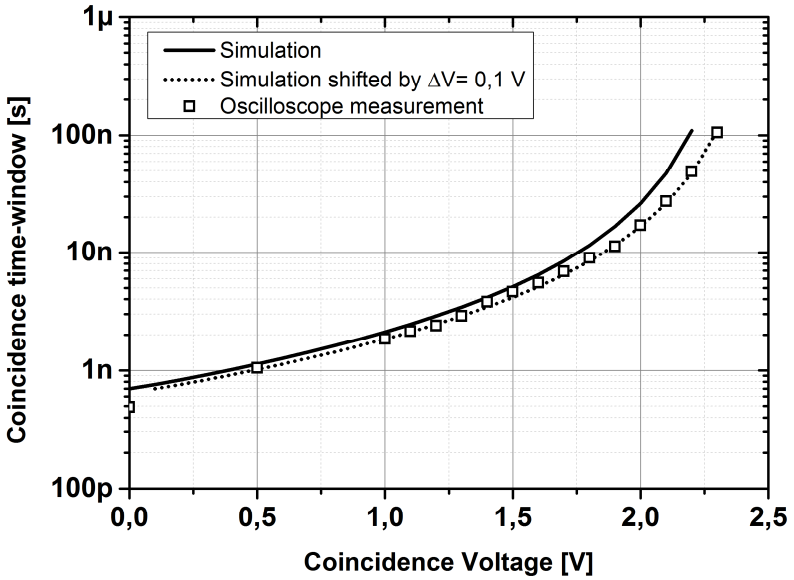

Figure 5: *Coincidence time-window as a function of the applied coincidence voltage.*

This discrepancy can be attributed to a possible threshold fluctuation of the p-MOS transistors in the circuit shown in Figure 10 from Chapter 2, with respect to the nominal value, the mismatch being larger at higher voltage biases (i.e. lower overdrive voltage applied on the p-MOS device, and thus a more threshold dependent result). Nevertheless, the simulation results can still be reasonably good within the time-window range of interest, i.e. $\Delta t < 10\ ns$ where the threshold variation can affect only in a minor way the resulting time-window due to the large overdrive voltage with respect to the threshold voltage of the p-MOS transistor. For this reason, the simulation data have been considered as a valuable reference for other measurements.
Figure 6 reports the resulting *FCR* measured with the internal coincidence circuit, in case of three different reverse biases (N.B.: the counting rate is plotted in a log-scale). The observed *FCR* continues decreasing rather slowly at shorter coincidence time-windows until a certain threshold of around $\Delta t = 1{,}5\ ns$ where the observed rate suddenly drops towards smaller values. Optical cross talk occurring between the SPAD*s* in a 3D-SiCAD pixel can be a possible cause of the observed offset of correlated counts on the *FCR* curves of Figure 6. As depicted in Figure 7, an avalanche event occurring in one of the two sensing levels produces a certain amount of photons by luminescence [1].

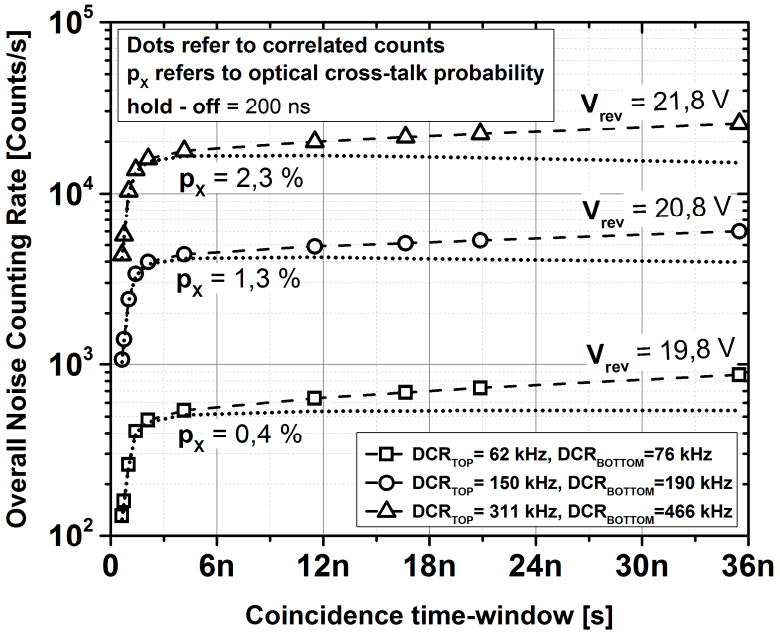

Figure 6: *Fake Coincidence Rate (FCR) recorded in a 3D-SiCAD pixel in dark condition, as a function of the adopted coincidence time-window, for different reverse bias voltages. FCR has been measured by using the coincidence circuit integrated in the 3D-SiCAD pixel, as it is capable to resolve much shorter coincidence time-window if compared to the FPGA acquisition system. Symbols refer to measured data (dashed lines are simply interpolating the measured data).*

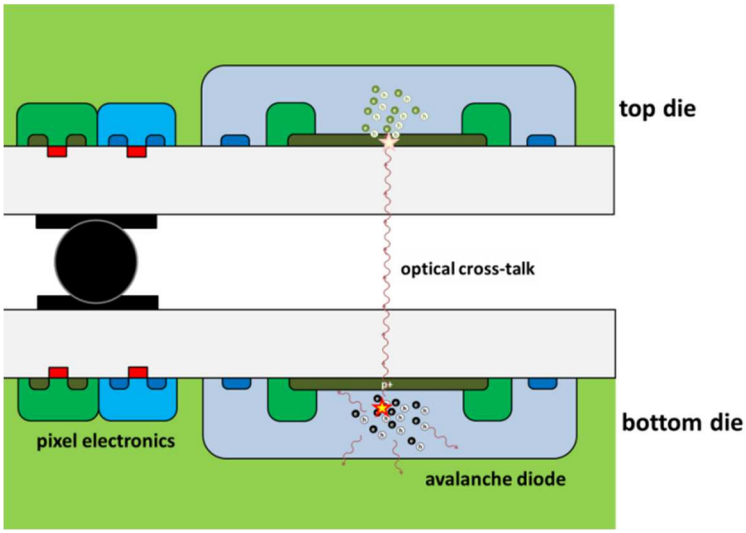

Figure 7: *Qualitative representation of an optical cross-talk event occurring in a 3D-SiCAD pixel. A primary avalanche occurring in the bottom die produces a certain amount of photons. One of them is suddenly absorbed in the top die, firing a correlated avalanche pulse.*

It may therefore happen that one (or more) of these photons is re-absorbed within the active area of the quiescent sensing level, giving rise to an avalanche pulse that is practically simultaneous to the initial one. Since the delay between two avalanches is mainly due to the time-jitter related to the avalanche build-up and the electronics, and thus falls within the sub-nanosecond range, a cross-talk event is in fact interpreted as a valid coincidence count by the electronics. This effect was actually considered to be negligible during the design phase, as the adopted CMOS technology features a polyimide passivation layer [2], which was expected to dramatically reduce the optical transparency of the dielectric material between the two sensing levels. In reality, in the discussion that follows, it is shown that even very small value amount of optical cross-talk can be the limiting factor of the noise rejection capability in a 3D-SiCAD pixel. As for the SPAD array case, the main effect of optical cross-talk is an enhancement on the dark count rate observed in the SPAD cells in the two sensing levels. If $P_{X,T\to B}$ represents the probability that an intrinsic dark count in the top SPAD, successfully induces an avalanche pulse in the bottom one, and if $P_{X,B\to T}$, represents the opposite case, then the overall DCR observed in the top and bottom levels, λ_T and λ_B, respectively, can be expressed as follows:

$$\begin{cases} \lambda_T = \lambda_{T,0} + P_{X,B\to T}\lambda_{B,0} \\ \lambda_B = \lambda_{B,0} + P_{X,T\to B}\lambda_{T,0} \end{cases} \quad (2)$$

where $\lambda_{T,0}$ and $\lambda_{B,0}$ refers to DCR of the top and bottom SPAD cells, respectively, in absence of the cross-talk enhancement. By assuming that $P_{X,B\to T} = P_{X,T\to B} = P_X$, the rate of false coincidences due to optical cross-talk can thus be estimated in a simple way, as follows [3]:

$$\lambda_X = P_X\left(\lambda_{T,0} + \lambda_{B,0}\right) = \frac{P_X}{1 + P_X}(\lambda_T + \lambda_B) \tag{3}$$

It is important to observe that the FCR formula provided by equation (17) from Chapter 1 has to be corrected from the correlated counts due to the optical cross-talk, as follows:

$$FCR(\Delta t) = 2\Delta t \lambda_{T,0}\lambda_{B,0} = 2\Delta t \frac{(\lambda_T - P_X\lambda_B)(\lambda_B - P_X\lambda_T)}{(1 - P_X^2)^2} \tag{4}$$

The overall observed Noise Counting Rate (NCR) is thus given by the combination of the real accidental coincidences and the correlated counts, as follows:

$$NCR(\Delta t) = FCR(\Delta t) + \lambda_X \tag{5}$$

This latter formula clearly shows that the optical cross-talk produces an offset on the observed counting rate curves. By extracting this offset from the measured data, it is eventually possible to calculate the cross-talk probability, as reported for each curve of Figure 6, but also, for convenience in Table II. The probability for optical cross-talk is rather small, i.e. less than 1 % at $V_{rev} = 19{,}8\,V$ and just a few % at larger biases. However, by comparing the measured FCR (FCR_m) with the ideal (and the expected) one (FCR_{id}), it is apparent that cross-talk coincidences represent the dominant contribution to the overall observed NCR. This observation can be better understood from a theoretical stand point, by considering a 3D-SiCAD pixel featuring the same DCR in the two sensing levels, i.e. $\lambda_T = \lambda_B = \lambda$. It is thus possible to define a Noise Rejection Factor (NRF) as the ratio between the overall NCR given by (5) and the DCR of a single SPAD in absence of optical cross-talk, i.e. $\lambda_{T,0} = \lambda_{B,0} = \lambda_0$, as follows:

$$NRF = \frac{NCR}{\lambda_0} = 2(P_X + \Delta t \lambda_0) \tag{6}$$

The NRF is finally plotted in Figure 8 according to equation (6), as a function of the DCR featured by the SPAD cells constituting the 3D-SiCAD pixel, by considering two different cross-talk probabilities of $P_X = 0{,}01$ % and $P_X = 1$ %, and a coincidence time-window of $\Delta t = 10\,ns$. Based on the way this figure of merit has been defined, a low NRF indicates that the 3D-SiCAD pixel features a good noise rejection capability.

Table II. Optical Cross-Talk Probability

Reverse Bias	FCR_m @$\Delta t = 4ns$	FCR_{id} @$\Delta t = 4ns$	P_X
19,8 V	542 Hz	39 Hz	0,4 %
20,8 V	4,4 kHz	237 Hz	1,3 %
21,8 V	17,7 kHz	1,2 kHz	2,3 %

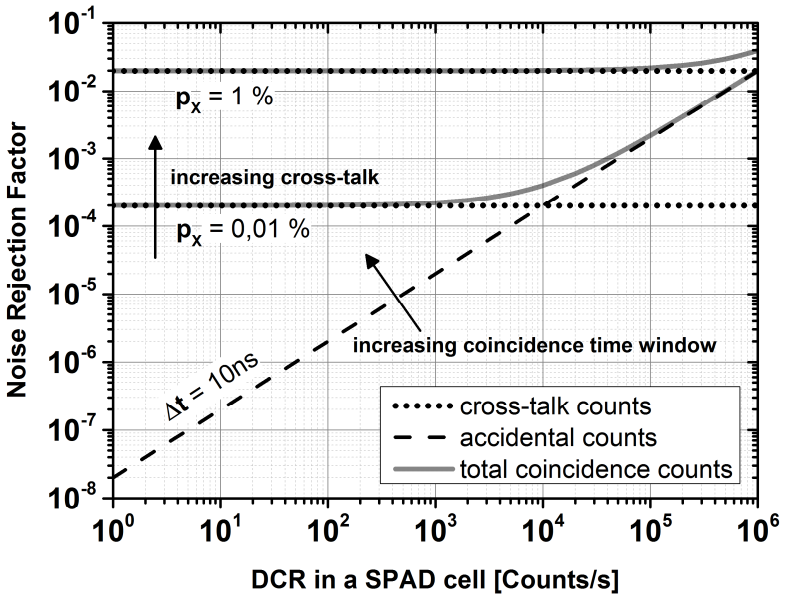

Figure 8: *Noise Rejection Factor as a function of the DCR featured by the SPAD cells constituting a 3D-SiCAD pixel. Two different cross-talk probabilities of* $P_X = 0{,}01\,\%$ *and* $P_X = 1\,\%$*, and a coincidence time-window of* $\Delta t = 10\,ns$ *have been considered in this analytical study.*

According to Figure 8, it is apparent that the noise rejection limit is set by the optical cross-talk probability, if the SPADs of the 3D-SiCAD pixel feature a DCR that is lower than a certain "corner" value given by $DCR_{corner} = 2P_X/\Delta t$. In order to achieve the lowest possible NRF, and hence to maximize the noise rejection capability of the device, the optical cross-talk phenomenon must be suppressed. This can be effectively done by interposing a fully absorbing/reflecting layer between the two sensing levels, for instance by simply covering the pixels with metal shields.

4.2 Study and Characterization of the Particle Detection Capability of a 3D-SiCAD pixel

The objective of this last section is to validate the particle detection capability of the 3D-SiCAD pixel cell studied in the first part of this chapter. In order to demonstrate this important feature, inverse square-law measurements have been performed over the device by adopting a commercial Strontium-90 radioactive source featuring a nominal activity $\lambda_{90_{Sr}} = 37 \pm 11{,}1\ MBq$. The source is shown in Figure 9 and it is characterized by an active bead featuring a spherical shape with a $1\ mm$ diameter. The whole component has been encapsulated in a Plexiglas container for a safe handling. Strontium-90 (^{90}Sr) is a radioactive isotope of strontium with a half-life of 28,8 years and undergoing a β^- decay with a maximal energy of $546\ keV$ distributed to an electron, an anti-neutrino, and the Yttrium-90 isotope (^{90}Y). This latter, in turn, undergoes another β^- decay but with a shorter half-life of 64 hours and maximal decay energy of $2{,}28\ MeV$ distributed to an electron, an anti-neutrino, and Zirconium-90 (^{90}Zr), which is stable. It is worth noticing that ^{90}Sr can be considered as a pure beta particle source, since the probability of having a gamma photon emission from the decay of ^{90}Y is typically negligible. Therefore the spectrum of a Strontium-90 radioactive source is characterized by two main branches related to the ^{90}Sr and ^{90}Y beta decays. These co-exist indeed in a so called "secular equilibrium", i.e. the two isotopes have the same activity, since the quantity of ^{90}Y remains constant because its production rate (e.g., due to decay of ^{90}Sr) is equal to its decay rate.

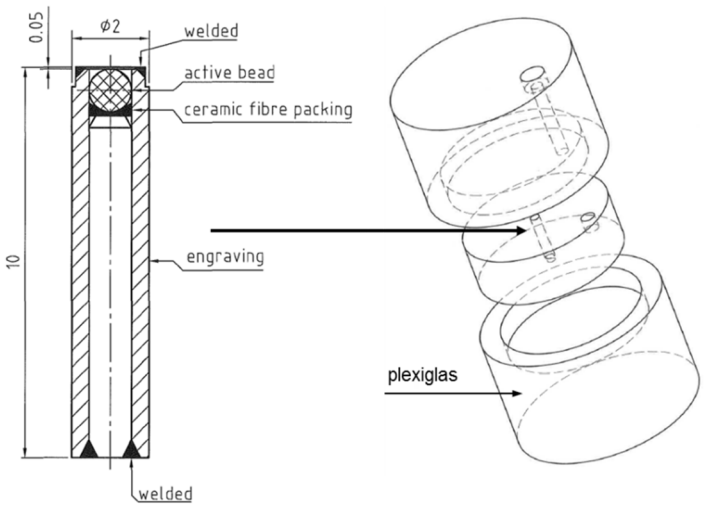

Figure 9: *Left: cross-section of the used commercial ^{90}Sr source. Right: Plexiglas container for the source (courtesy of Institut de Physique Nucleaire de Lyon). Sizes are expressed in mm.*

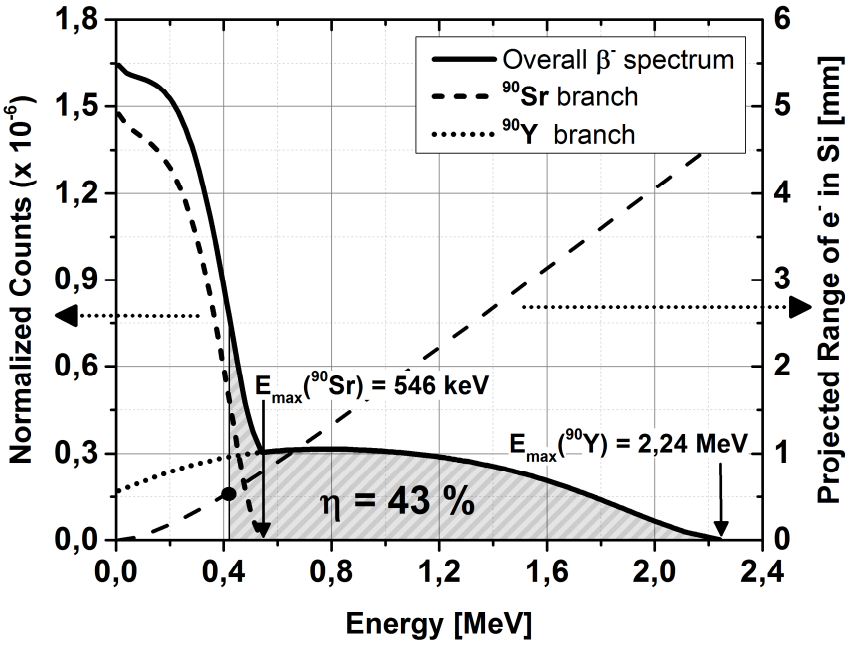

Figure 10: *β^- Spectrum of a ^{90}Sr radiation source (left y-axis) and projected range of electrons in Silicon (right y-axis) calculated according to equation (7) which is based on the Katz and Penfold empirical formula [5].*

For this reason the beta decay spectrum of the source can be derived as the balanced weighted sum of the ^{90}Sr and ^{90}Y spectra obtained from the data available in the "Laboratoire National Henri Becquerel" website [4]. The resulting spectrum $S_{90_{Sr}}(E)$ is represented by the solid line in Figure 10.

4.2.1 Methods

The adopted radioactive source can be considered as an omnidirectional isotropic emitter, i.e. beta particles are randomly emitted in all directions. Therefore inverse square-law measurements allow controlling the rate of particle hitting the 3D-SiCAD pixel by simply varying the distance between the detector and a charged particle radioactive source. Under the point-like source approximation, the resulting variation of the measured number of counts CR_m can be described as follows:

$$CR_m = FCR + \lambda_{90_{Sr}} \alpha(t_{sub}) \frac{A_{pixel}}{4\pi(d_0 + d)^2} \eta \qquad (7)$$

where A_{pixel}, FCR and η are the 3D-SiCAD pixel active area, fake coincidence rate and particle detection efficiency, respectively; $\lambda_{90_{Sr}}$ is the activity of the radioactive source; d_0 is the minimal distance between the source and the active

area of the 3D-SiCAD pixel (due to systematic limitation of the experimental setup), d is the displacement of the detector with respect to such a reference. The parameter $\alpha(t_{sub})$ quantifies the fraction of the emitted beta electrons that can successfully penetrate both sensing levels of the 3D-SiCAD pixel. An important amount of low energy beta particles is indeed absorbed in the silicon substrate of the top level test-chip whose thickness is $t_{sub} \approx 550\ \mu m$. The fraction of particles that can be effectively detected by the device can be estimated to the first order by evaluating the minimal energy at which the electron range in silicon R_{Si} is larger than the top die substrate thickness, according to the Katz and Penfold empirical formula [5] in case of electron energies lower than $2{,}5\ MeV$:

$$R_{Si} = \frac{1}{\rho_{Si}} 0{,}412\ E^{1{,}265-0{,}0954\ln(E)} \tag{8}$$

where E is the energy of an electron emitted from the source and ρ_{Si} is the silicon density. The minimal energy at which a beta particle can effectively cross both sensing level of a 3D-SiCAD pixel is thus given by the condition $R_{Si} > t_{sub} = 550\ \mu m$. This situation is represented by the black dot in Figure 10 indicating that an electron range of $R_{Si} = t_{sub}$ occurs at an energy of around $E_0(t_{sub}) = 0{,}42\ MeV$. The amount of electrons that can be successfully detected is finally obtained by integrating the spectrum from such a minimal energy, as represented by the shaded area in Figure 10, resulting in a fraction of detectable particles of around $\alpha(t_{sub}) = 43\ \%$:

$$\alpha(t_{sub}) = \int_{E_0(t_{sub})}^{\infty} S_{90_{Sr}}(E)dE \tag{9}$$

Table III reports for convenience the known parameters of equation (7). The measurement has been implemented according to the setup schematically represented in Figure 10. The distance and the correct alignment between the radioactive source and the 3D-SiCAD pixel are adjusted by means of a proper alignment system controlled via computer.

Table III. Parameters

Parameter	Value
$\alpha(t_{sub})$	43 %
A_{pixel}	$2500\ \mu m^2$
d_0	$5\ mm$
$\lambda_{90_{Sr}}$ (nominal)	$37 \pm 11{,}1\ MBq$

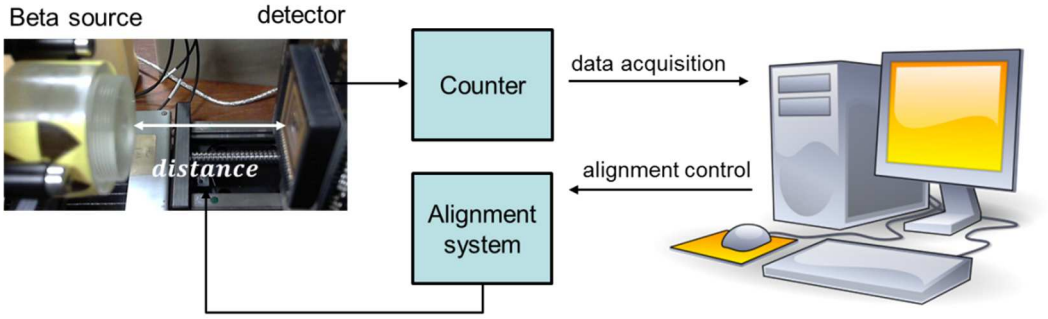

Figure 10: *Schematic representation of the experimental setup installed at the "Institut de Physique Nucleaire de Lyon".*

The number of coincidence counts arising from the 3D-SiCAD pixel is recorded by an oscilloscope which is capable of counting the number of hits of interest within a desired measurement window. The coincidence counts produced by real events (i.e. electrons emitted during β^- decay) can be recovered from the background noise (i.e. fake coincidences) by subtracting the FCR from the overall counting rate measured at a given distance from the source. The former is measured by moving away the detector from the source. The resulting β^- counting rate CR_β is thus given by:

$$CR_\beta = CR_m - FCR \tag{10}$$

In analogy with SPAD detectors [3], the fluctuation of the overall coincidence counts in the measurement interval translates into an uncertainty affecting the β^- counting rate, which has to be carefully controlled by a proper choice of the acquisition time-window. These fluctuations are indeed due to the statistical nature of both the incident rate of beta particles over the pixel S_β and the accidental coincidence counts FCR. Their impact on the beta counting rate measurement CR_β can be described in terms of standard deviation with respect to the mean values given by (10), as follows:

$$\sigma^2(CR_\beta) = \sigma^2(S_\beta) + \sigma^2(FCR) \tag{11}$$

In general, both statistical fluctuations can be considered as Poisson processes, and therefore the uncertainty affecting the measurement can be reduced to be sufficiently small by allowing the counting instrument to acquire data for large enough acquisition time-windows. Moreover the estimation of such an uncertainty is directly provided by the counter, which allowed a quick definition of an acquisition window of $T_m = 10\ s$ as a good trade-off between acceptable uncertainty and acquisition time.

4.2.2 Optimization of the 3D-SiCAD working parameters

Before proceeding with the characterization of the device by means of inverse square law measurements, it is important to analyze its behavior with the aim of optimizing the measurement conditions with respect to some key parameters, such as the hold-off voltage of the quenching circuit of SPAD cells in the two sensing levels, the reverse bias of the avalanche diodes, and the coincidence time-window. In order to properly separate coincidence counts produced by real events (i.e. beta particles) from the background noise (i.e. fake coincidences) the measurement has been performed at the minimal distance d_0 between the source and the detector. In this way, it has been possible to benefit of the highest achievable rate of beta particle hits over the pixel, with respect to the adopted measurement setup, maximizing the signal-to-noise ratio.

4.2.2.1 β^- counting rate versus hold-off voltage

Figure 12 shows the measured β^- counting rate as a function of the hold-off voltage V_h (i.e. the parameter responsible for adjusting the dead-time after every avalanche ignition in each sensing level of the 3D-SiCAD pixel). The avalanche diode is reverse biased at $V_{rev} = 19{,}5\,V$ and the adopted coincidence time-window for the measurement is $\Delta t = 10\,ns$.

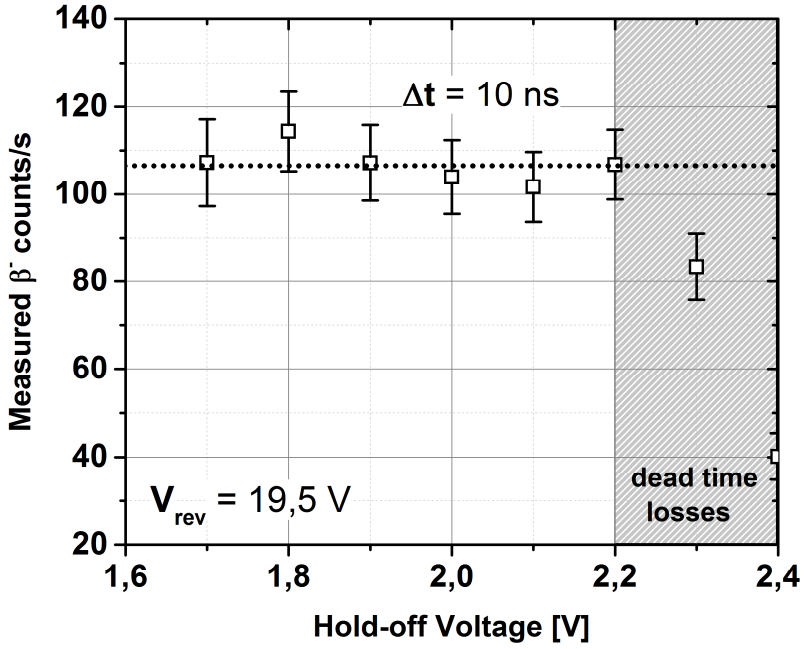

Figure 12: *Measured β^- counting rate as a function of the hold-off voltage V_h. The avalanche diode is biased at $V_{rev} = 19{,}5\,V$ and the adopted coincidence time-window is $\Delta t = 10\,ns$. The measurement is obtained by collecting coincidence counts for 10 seconds.*

The measured beta particles counting rate is approximately constant up to the hold-off voltage $V_h = 2{,}2\,V$, where it starts falling towards lower values. Larger hold-off times increase indeed the probability to find both sensing levels of a 3D-SiCAD pixel in a non-receptive state when the charged particle hits the device. A hold-off voltage of $V_h = 2\,V$, i.e. hold-off time of around $t_h = 200\,ns$, can thus guarantee negligible dead-time losses (estimated as small as 2 % based on the DCR and hold-off time data, not reported here for brevity) while keeping the after-pulsing probability at values lower than 10 %, which is beneficial in the minimization of the noise counting rate, i.e. the FCR.

4.2.2.2 β^- counting rate versus coincidence time-window

The beta counting rate has been subsequently studied by sweeping the coincidence time-window from $\Delta t \approx 0{,}6\,ns$ up to $\Delta t \approx 4\,ns$. It is indeed of crucial importance to understand how short this latter parameter can be, since the amount of accidental counts affecting the measurement is strictly related to the coincidence time-window, as discussed at the beginning of this chapter. Figure 13 reports the result of this measurement, in case of a hold-off voltage of $V_h = 2\,V$ and by considering two different bias scenarios for the avalanche diode, i.e. $V_{rev} = 19{,}5\,V$ and $V_{rev} = 19{,}8\,V$. In both cases, there is a cut-off time-window below which the detected particle rate suddenly drops from a rather constant value (within the measurement error) towards a zero value. This occurs for $\Delta t < 1{,}25\,ns$ in case of $V_{rev} = 19{,}5\,V$ and for $\Delta t < 0{,}8\,ns$ in case of $V_{rev} = 19{,}8\,V$.

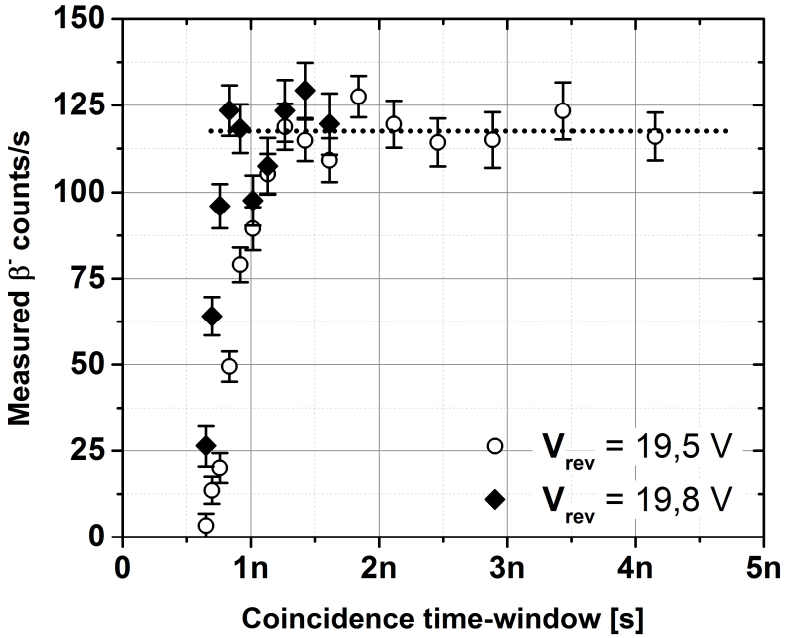

Figure 13: *Measured β^- counting rate as a function of the coincidence time-window Δt. Two avalanche diode biase scenarios have been considered and the adopted hold-off time voltage is $V_h = 2\,V$. The measurement is obtained by collecting coincidence counts for 10 seconds.*

The reason of that can be explained by accounting for the unavoidable time-jitter introduced by the electronics, which is indeed not capable to properly resolve coincidence events occurring within a time-window lower than the observed cut-off. The cut-off time-window is actually higher in the scenario considering the lower reverse bias for the avalanche diode. The comparator threshold of the SPAD quenching circuit is indeed around $V_{th} = 0{,}5\,V$, which means that the maximal voltage drop across its input V_{in} would be only $V_{ex} = V_{rev} - V_{bd} = 0{,}7\,V$, i.e. just $0{,}2\,V$ above its threshold, in this latter scenario. Such a small voltage-above-threshold value could actually affect the timing performance of the quenching electronics, and thus the time resolution of the overall coincidence detection system. In order to guarantee a precise commutation time, the leading edge of the avalanche pulse (i.e. the output voltage observed at the output of the avalanche diode) should be indeed fast enough to make thermal noise fluctuation at the comparator input negligible. This one is typically modeled by a simple RC transient with time constant τ and is strictly related to the excess bias voltage as described by (12):

$$\left.\frac{dV_{in}}{dt}\right|_{@\,V_{th}} = \frac{V_{ex} - V_{th}}{\tau} \tag{12}$$

Table IV shows the resulting voltage rising rate dV_{in}/dt evaluated when the diode output voltage crosses the comparator threshold and assuming an RC constant of $\tau = 200\,ps$. The diode space charge capacitance has been indeed estimated to be on the order of $C_D = 1\,pF$. The diode series resistance has been extrapolated from the I-V curve presented in Chapter 4 and its value is $R_D = 200\,ohm$. As expected lower reverse biases translate into a slower voltage rising rate which results in a response time of the comparator rather sensitive to thermal noise fluctuation at its input. It is worth noticing that an excess bias of $V_{ex} = 1\,V$, i.e. $V_{rev} = 19{,}8\,V$, provides a 2,5 faster pulse rising rate with respect to the case of an excess bias of $V_{ex} = 0{,}7\,V$. The same argument applies actually at the comparator output as its commutation time depends on the voltage applied at its input. The voltage comparator implemented in the quenching circuit described in Chapter 2, consists indeed of a simple n-MOS transistor. Therefore lower reverse bias voltages translate into smaller gate-to-source voltages for the n-MOS transistor comparator which in turn is less conductive and for this reason gives rise to a slower commutation. It is reasonable to consider this latter effect as being most likely the dominant cause responsible of the time-jitter worsening of the quenching electronics.

Table IV: Voltage rising rate estimation at the comparator input

V_{ex}	$dV_{in}/dt \;@\; V_{in} = V_{th}$
$0{,}7\,V$	$100\,mV/\,0{,}1ns$
$1\,V$	$250\,mV/\,0{,}1ns$
$2\,V$	$750\,mV/\,0{,}1ns$
$3\,V$	$1{,}25\,mV/\,0{,}1ns$

This aspect has not been investigated further, since in the present case it is worthless to look for an ultra-short coincidence time window, as the noise rejection capability of the 3D-SiCAD pixel is actually limited by the optical cross-talk occurring between the two sensing levels of the device. For this reason, in the case under study, a coincidence time-window of $\Delta t = 2\ ns$ has been considered sufficiently large to prevent counting losses and, at the same time, sufficiently small to allow a good accidental count rejection capability.

4.2.2.3 β^- counting rate versus reverse bias voltage

The particle detection efficiency η of the 3D-SiCAD pixel has been finally studied by measuring the detected particle counting rate as a function of the applied reverse bias voltage, within the range going from $V_{rev} = V_{bd} = 18{,}8\ V$ up to $V_{rev} = 20{,}8\ V$. Figure 14 reports the result of this measurement, in case of a hold-off voltage of $V_h = 2\ V$ and a coincidence time-window of $\Delta t = 2\ ns$. First of all, it is important to observe that beta particles cannot be detected at reverse bias values lower than $V_{rev} = 19{,}3\ V$, since the avalanche voltage pulses do not cross the comparator threshold of $V_{th} = 0{,}5\ V$. The measured particle counting rate for $V_{rev} \geq 19{,}3\ V$ is conversely rather constant within the measurement uncertainty. This result is very important because it shows that the efficiency of the 3D-SiCAD pixel in detecting beta particles is practically 100%, even at rather low excess bias voltage, i.e. just $V_{ex} = 0{,}5\ V$.

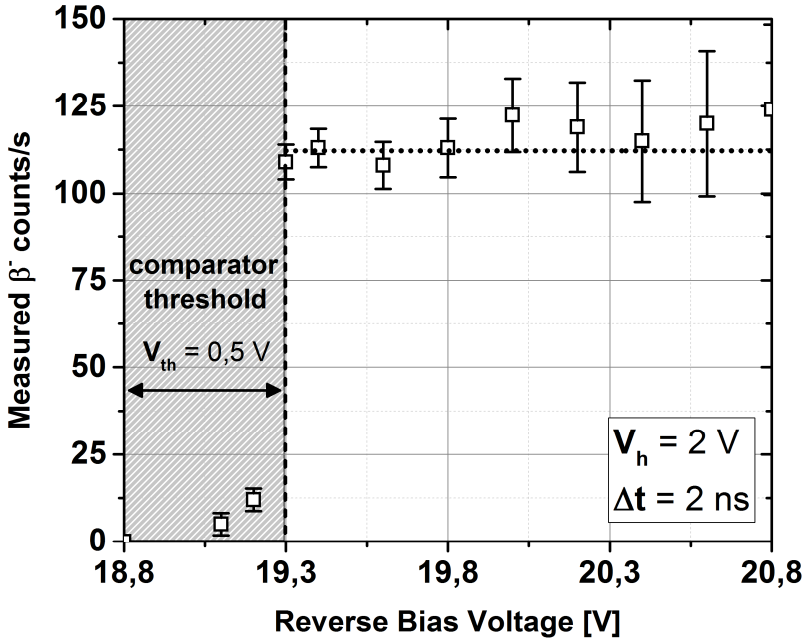

Figure 14: *Measured β^- counting rate as a function of the reverse bias voltage. The adopted hold-off time voltage is $V_h = 2\ V$ and the coincidence time-window is $\Delta t = 2\ ns$. The measurement is obtained by collecting coincidence counts for 10 seconds.*

The amount of charge generated by ionization due to the passage of a beta particle through the 3D-SiCAD pixel and effectively collected by the device, is thus sufficiently large to ensure the occurrence of an avalanche multiplication process, even if the probability for a single electron-hole pair (EHP) to fire an avalanche is much lower than 100 %. This result can be understood by recalling the discussion made in Chapter 1. With respect to equation (23) in Chapter 1, the detection probability of a Minimum Ionizing Particle (MIP) in a 3D-SiCAD pixel can be indeed expressed as follows:

$$P_{MIP_{3D-SiCAD}} = P_{MIP,SPAD_{top}} P_{MIP,SPAD_{bottom}} \eta_\alpha FF$$

where $P_{MIP,SPAD_{top}}$, $P_{MIP,SPAD_{bottom}}$ are the MIP detection efficiencies for the two SPAD sensing levels given by equation (9) from Chapter 1, FF is the fill factor for each SPAD pixel and η_α is the angular efficiency. Since the focus is on a single pixel, there is not much sense in considering the fill factor in the equation above, i.e. it can be set as $FF = 1$. Moreover the angular efficiency can be considered to be 100% since the beta particles hit the device perpendicularly with respect to the pixel surface in the current experiment. The above formula can be thus rewritten as follows:

$$P_{MIP_{3D-SiCAD}} = \left[1 - \left(1 - \frac{1}{W_{SPAD}} \int_0^{W_{SPAD}} n_c(x) P_{tr}(x) dx\right)^{R_i W_{SPAD}}\right]^2 \qquad (13)$$

where R_i is the average ionization rate in silicon produced by a MIP, W_{SPAD} is the extension of the SPAD active region, $\eta_c(x)$ is the collection efficiency of EHP generated at a position x within the SPAD active region towards the multiplication region and $P_{tr}(x)$ is the avalanche triggering probability for a single EHP generated at a position x. Based on the technology process parameters, the extension of the SPAD active region has been estimated to be around $W_{SPAD} \approx 6\,\mu m$. Therefore, by considering a MIP ionization rate of around $R_i \approx 80\,EHP/\mu m$, the average number of EHP produced by ionization should be about $N_{EHP} = 480$. In order to use equation (13) as a useful tool for the understanding of the results of Figure 13, the equation has been approximated in a simpler way by considering an average avalanche triggering probability for a single EHP being accelerated in the SPAD space charge region P_{tr} (instead of $P_{tr}(x)$), and considering an average collection efficiency η_c (instead of $\eta_c(x)$) for a carrier generated in the neutral regions to be collected in the multiplication region:

$$\eta_{MIP_{3D-SiCAD}} \approx [1 - (1 - \eta_c P_{tr})^{N_{EHP}}]^2 \qquad (14)$$

Figure 15 shows the particle detection efficiency as a function of the average EHP avalanche triggering probability, according to equation (14), by considering a carrier collection efficiency of $\eta_c = 50$ %.

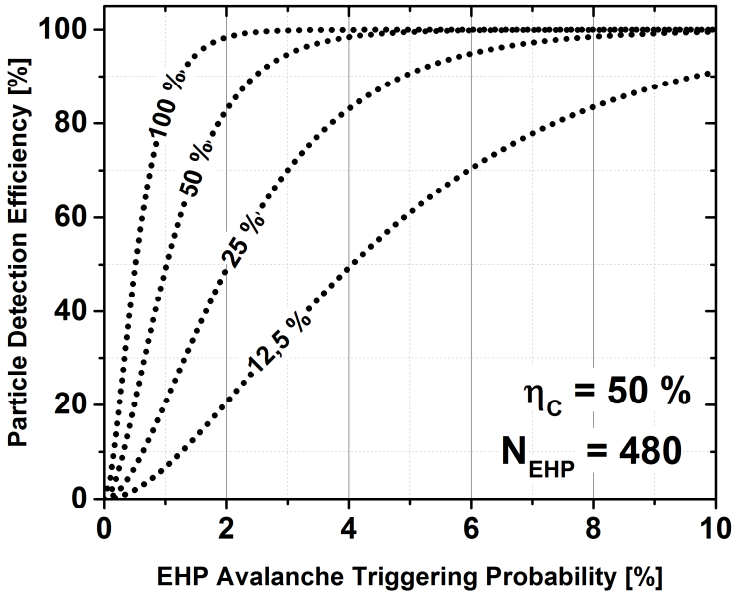

Figure 15: *Detection efficiency of a MIP as a function of the average EHP avalanche triggering probability (P_{tr}), calculated according to equation (14). Three different fluctuation scenarios on the ionization yields with respect to the value predicted by the Bethe-Bloch equation, i.e.* $100\%\, N_{EHP}$*, have been considered for this study: i.e.* $50\%\, N_{EHP}$*,* $25\%\, N_{EHP}$ *and* $12{,}5\%\, N_{EHP}$

Three different fluctuation scenarios on the ionization yields, i.e. $50\%\, N_{EHP}$, $25\%\, N_{EHP}$ and $12{,}5\%\, N_{EHP}$ with respect to the one predicted by the Bethe-Bloch equation, i.e. $100\%\, N_{EHP}$, have been considered. It can be easily observed from Figure 15 that the detection efficiency is larger than 90 % in all scenarios provided that the average P_{tr} is larger than 10 %. Therefore, the detection efficiency can attain excellent levels even accounting for a rather large fluctuation on the amount of carrier collected in the SPAD multiplication region due to the statistical nature of the nuclear ionization process.

This discussion, together with the experimental results reported in Figure 14, highlights the importance of keeping the reverse bias voltage of the avalanche diodes in a 3D-SiCAD pixel sufficiently high to guarantee a 100% particle detection probability, but also sufficiently low to minimize as much as possible the amount of dark counts and after-pulses occurring in the sensing levels of the device which in turn translates into a minimization of the amount of coincidence accidental counts. For this reason, in the inverse square-law measurements presented in the following section, a reverse bias voltage of $V_{rev} = 19{,}5\, V$ has been adopted.

4.2.3 Inverse Square-law measurements

Figure 16 shows the inverse square-law measurements performed over a 3D-SiCAD pixel. The curves have been obtained by acquiring the amount of coincidence counts by varying the distance between the radioactive source and the pixel from $d = 0$ (i.e. real distance is $d_0 \approx 5\,mm$) up to $d = 45{,}5\,mm$ (i.e. real distance is $d + d_0 \approx 50{,}5\,mm$) for three different coincidence time-windows. According to the discussion of the previous section, the avalanche diodes have been biased at $V_{rev} = 19{,}5\,V$ and the adopted hold-off voltage for the quenching circuit is $V_h = 2\,V$. As expected, it is observed that the amount of collected counts per seconds (symbols in Figure 16) decreases with the distance and that the curves are vertically shifted depending on the chosen coincidence time-window which is indeed related to the amount of accidental counts. The experimental data have been fitted with the inverse square-law model described by equation (7). The details of the fitting results are reported in Table V, while the corresponding fitted curves are shown as dotted lines in Figure 16. By observing the graph, it is possible to conclude that equation (7) fits quite well the experimental data. Moreover from the results of the fit, it has been possible to derive the FCR of the pixel and the activity of the radioactive source, as reported in Table V. The extracted activity is in excellent agreement with the radioactive source specifications. The measured data fall indeed within 13% of the nominal value of 37 Mbq which is in turn affected by an uncertainty of $\pm$ 30%.

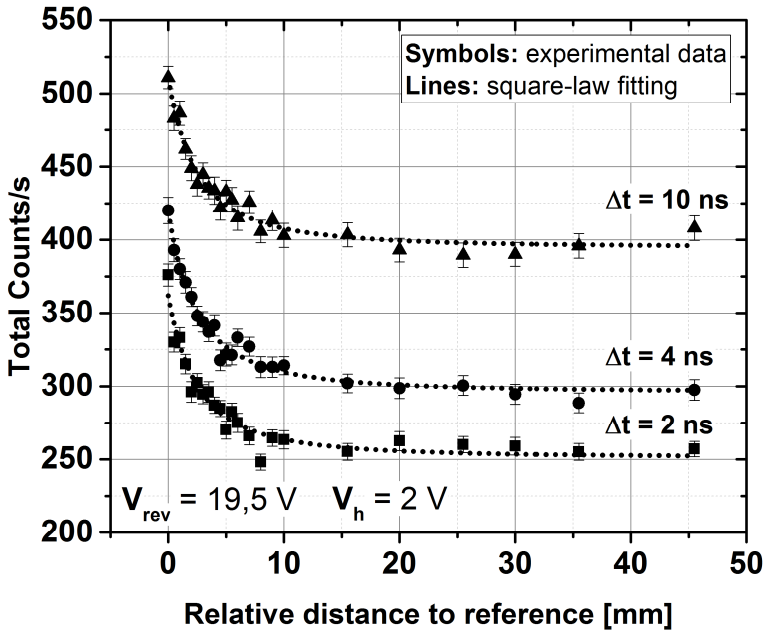

Figure 16: *Inverse square-law curves: measured β^- counting rate as a function of the distance between the radioactive source and the detector for three different coincidence time-window values. The adopted hold-off time voltage is $V_h = 2\,V$ and the reverse bias voltage is $V_{rev} = 19{,}5\,V$. The measurement is obtained by collecting coincidence counts for 10 seconds.*

Table V. Fitting results of the inverse square-law measurement

Δt	d_0	$\alpha(t_{sub})$	FCR @ $V_{rev} = 19{,}5\ V$	Extracted activity
2 ns	5 mm	43 %	$251 \pm 2\ Hz$	$32{,}3\ \pm 1{,}8\ MBq$
4 ns	5 mm	43 %	$296 \pm 2\ Hz$	$35{,}9\ \pm 1{,}4\ MBq$
10 ns	5 mm	43 %	$395 \pm 2\ Hz$	$33{,}8\ \pm 1{,}5\ MBq$

4.2.4 Discussion

The inverse square-law measurement results, together with the preliminary measurements conducted in order to optimize the operating parameters for the device, show that the realized 3D-SiCAD pixel can effectively detect ionizing particles. This is corroborated by the fitting study, resulting in a radioactive source activity which is in excellent agreement with expectations. Similar curves would have not been obtained in only 10 seconds measurement by adopting single SPAD devices (at least the ones designed in this work) as the levels of dark counts would make impossible to distinguish the beta particle counting rate signal from the background counts. There are however different uncertainties affecting the study which would deserve a deeper look. The parameter $\alpha(t_{sub})$ has been indeed estimated based on simple considerations and a more accurate study on this matter would be worth of interest. Moreover the systematic minimal distance d_0 between the source and the detector has been estimated based on some geometrical parameters of the experimental setup. Nevertheless the value of $d_0 = 5\ mm$ is considered to be a rather good estimation for that, and a maximal $\pm\ 10\ \%$ uncertainty with respect to the estimated value is reasonably expected. The precision of the measurement would certainly improve if larger acquisition time-windows were used instead of only 10 seconds. As discussed in Section 4.1, the background noise, i.e. the FCR, is unfortunately larger than expected due to the optical cross-talk occurring between the two sensing levels of the 3D-SiCAD pixel, which is responsible to dramatically enhance the overall accidental counts. This effect has to be suppressed in a future design by interposing a fully absorbing/reflecting layer between the two sensing levels, for instance by simply covering the pixels with metal shields.

Conclusion

In this chapter, a single pixel cell of a first 3D-SiCAD prototype has been fully characterized with the aim of validating the expected device capability in greatly rejecting background counts, and in detecting ionizing particles with an excellent efficiency. Preliminary measurements over two adjacent "in-plane" avalanche diodes operated in coincidence-mode demonstrated a very high noise rejection capability, i.e. up to 2-3 orders of magnitude lower than the intrinsic dark count rate of each SPAD cell of a single 3D-SiCAD pixel. Surprisingly, lower noise rejection capabilities were conversely observed on a real 3D pixel. This underperformance was found to be caused by optical cross-talk occurring between the two sensing levels of the 3D-SiCAD pixel. This drawback must be suppressed in order to restore the full noise rejection capability intrinsically related to this novel detector. This can be done rather easily by interposing a fully absorbing or reflecting medium between the two sensing levels. The detection capability of the realized prototype to ionizing radiation has been finally demonstrated by means of a commercial Strontium-90 radioactive source featuring a nominal activity of 37 MBq and inverse square-law measurements results, together with the preliminary measurements conducted in order to optimize the working parameters for the device. A fitting study over the measured data helped corroborating the validity of the results, as it allowed providing an estimation of the adopted radioactive source activity which is in excellent agreement with the expectations.

References

[1] A. L. Lacaita, F. Zappa, S. Bigliardi, and M. Manfredi, "On the bremsstrahlung origin of hot-carrier-induced photons in silicon devices," *IEEE Trans. Electron Devices*, vol. 40, no. 3, pp. 577–582, Mar. 1993.

[2] 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," *Proc. SPIE 6372, Adv. Phot. Count. Tech. 63720S*, vol. 6372, p. 63720S–63720S–12, 2006.

[3] F. Zappa, S. Tisa, a. Tosi, and S. Cova, "Principles and features of single-photon avalanche diode arrays," *Sensors Actuators, A Phys.*, vol. 140, no. 1, pp. 103–112, 2007.

[4] "Laboratoire National Henri Becquerel." [Online]. Available: http://www.nucleide.org/.

[5] L. Katz and A. S. Penfold, "Range-energy relations for electrons and the determination of beta-ray end-point energies by absorption," *Rev. Mod. Phys.*, vol. 24, no. 1, pp. 28–44, 1952.

Conclusion and perspectives

In this work, a 3D-SiCAD (3D Silicon Coincidence Avalanche Detector) demonstrator based on a commercial CMOS technology and common 3D integration techniques has been successfully developed, fabricated and characterized. The major objective was to demonstrate the feasibility of this novel device and to validate the expected performances in terms of excellent particle detection efficiency and noise rejection capability with respect to background counts, in charged particle tracking systems.

The main points tackled in this work are summarized hereafter.

In the first part of Chapter 1, a concise overview on Single Photon Avalanche Diodes (SPAD), i.e. the building block of a 3D-SiCAD pixel, has been addressed. The physics and the corresponding mathematical description of the processes behind the detection of a single photon have been discussed in detail. Consequently, a general analytical model allowing the evaluation of the detection efficiency for a single charged particle in a similar way as for single photons has been proposed. This mathematical description allows a very convenient study of the particle detection process which is expected to be sufficiently accurate in most practical situations. The physical processes responsible for the noise counts in SPAD devices have been then discussed in details together with the related mathematical modeling required for a correct data analysis of the experimental results. Other important figures of merits such as the time-resolution (or time-jitter) and cross-talk between pixels in case of matrix implementation have been further discussed. A complete overview about the state-of-the art for the SPAD devices has been finally proposed, with special focus on CMOS implementations. In the second part of Chapter 1, the working principle of the novel 3D-SiCAD device has been described in detail. More specifically the noise performance and the charged particle detection efficiency for this device have been defined in analogy with the discussion on SPAD detectors. Given the 3D structure for this novel device, a specific parameter referred to as “angular efficiency” has been introduced, in order to account for possible counting losses related to the angle of incidence of the incoming radiation. Given the novelty of the device a short state-of-the art, mostly concerning current developments, has been finally proposed.

In Chapter 2, the design of a first 3D-SiCAD demonstrator has been presented in detail. The Austria Micro-System $0{,}35\ \mu m$ High Voltage CMOS process has been chosen for the realization of the SPAD pixels. The avalanche diode in each pixel cell has been designed according to a “diffused guard-ring” implementation, with an active region defined by a p+ diffusion over a deep n-tub region in order to prevent premature breakdown at the device periphery. The diode geometry has been laid-out to provide a square-like active area featuring a $50\ \mu m$ side length. The quenching electronics necessary to ensure correct Geiger-mode operation has been realized by adopting a time-integration based passive quench-

ing / active recharge approach [1], providing user-adjustable hold-off times ranging from around $t_h = 50\,ns$ up to $t_h = 5\mu s$, and allowing excess bias voltages from around $V_{ex} = 0{,}5\,V$ up to $V_{ex} = 3{,}5\,V$. The 3D-level pixel electronics has been optimized to ensure a proper interfacing between the two sensing levels in a 3D-SiCAD pixel and, more importantly, assessing the occurrence of coincidence hits by providing a user-adjustable coincidence time-window ranging from around $\Delta t = 0{,}5\,ns$ up to $\Delta t = 50\,ns$. The realization of a 3D prototype has been tackled by choosing a simple die-to-die flip-chip approach by means of gold micro-bumps featuring a $70\,\mu m$ diameter. For this purpose, the test-chip has been laid-out in a way that a correct stack between the sensing levels of a 3D-SiCAD pixel could be ensured by assembling together two identical test-chips. The resulting 3D-assembled prototype provided 3D-SiCAD pixels arranged as single cells and 2×2 matrix cells.

The characterization results of the SPAD cells, i.e. the building blocks of a 3D-SiCAD pixel, have been presented and critically analyzed in Chapter 3. The measured avalanche diodes showed a breakdown voltage of around $V_{bd} = 18{,}8\,V$ ($18{,}77 \pm 0{,}13\,V$ over a set of 28 devices), with a temperature coefficient (extracted only from a single-device measurement) of $dV_{bd}/dT = 20{,}4\,mV/°C$. The Arrhenius plot study of the saturation current measured from a single avalanche diode, showed that thermal generation might be the dominant mechanism for the electron – hole pair generation in the device. The electro-luminescence test showed rather good uniform light emission intensity without hot spots all over the SPAD active region. This indicated that the avalanche diode features a rather uniform electric field distribution and that there are no clusters of defects over its active area. Pixel electronics characterization over 28 SPAD pixels from 4 different test-chips showed that the hold-off time ranges monotonically from around $t_h = 45\,ns$ up around $t_h = 6{,}5\,\mu s$ within the hold-off voltage range going from $V_h = 1{,}6\,V$ to $V_h = 2{,}4\,V$. The statistical fluctuation over the measured samples is lower than 10 % if $t_h < 550\,ns$. Extensive Dark Count Rate (DCR) measurements over a set of 35 SPAD pixels from 5 different test-chips and for several excess bias voltages revealed that the dark counts are affected by a rather severe statistical variation spanning within two orders of magnitude from a few $Hz\,\mu m^{-2}$ up to a few hundreds of $Hz\,\mu m^{-2}$, even if about 40% of the pixels shows a rather good uniformity close to the median value. In case of $V_{ex} = 1\,V$ the distribution has a median value of around $70\,Hz\,\mu m^{-2}$ which is in good agreement with respect to what has been obtained by Vilella et al. (see ref. [2], page 11, Picture 4.13 - left) for the measurements of a SPAD matrix detector operated in free-running mode. After-pulsing measurements over 28 SPAD pixels from 4 different test-chips, with a hold-off time of $t_h = 150\,ns$ and for a $V_{ex} = 1V$ showed a median value of around 12%, which is actually a rather severe value, considering that this probability will be increasingly higher at larger reverse biases and at shorter hold-off times. Photon detection efficiency (PDE) measurements over a single SPAD pixel within the spectral range $400\,nm - 1000\,nm$ indicated that the photon detection capability of the device is rather poor. The measured PDE in case of an excess bias of $V_{ex} = 2\,V$ reaches

indeed its maximal value of less than 5 % at $\lambda = 600$ nm. This performance is quite below expectations, but is in agreement with the results of Vilella et al. [2] where the cause of these results was attributed to the polyimide passivation of the adopted CMOS process, which dramatically reduces the optical transparency of the dielectric material in the circuit back-end [3].

In Chapter 4, a single pixel cell of a first 3D-SiCAD prototype has been fully characterized, demonstrating the expected device capability in rejecting background counts, and in detecting ionizing particles with an excellent efficiency. Preliminary measurements over two adjacent "in-plane" avalanche diodes operated in coincidence-mode showed a good noise rejection capability, i.e. up to 2-3 orders of magnitude lower than the intrinsic dark count rate of each SPAD cell of a single 3D-SiCAD pixel. The DCR measured for each single pixel is indeed in the range of $100\,kHz - 500\,kHz$ (depending on the adopted reverse bias / hold-off time) while, for a rather "long" coincidence time-window of $20\,ns$, the recorded rate of fake coincidences falls within the range of $500\,Hz - 5k\,Hz$, i.e. 2 orders of magnitude lower than the corresponding DCR. This clearly demonstrates the noise rejection capability achievable in the implemented coincidence detection mode, even if the results are apparently far to be close to the optimal achievable value. Less noisy SPADs, i.e. a cleaner CMOS process, and a factor 10 shorter coincidence time-window (i.e. $\Delta t = 2\,ns$), would provide indeed a significant improvement on the overall FCR. Surprisingly, lower noise rejection capabilities were conversely observed on a real 3D pixel. Even though in this case the SPAD cells show a lower DCR with respect to the ones considered in preliminary study, the recorded rate of fake coincidences for a time-window of $20\,ns$, falls within the range the of $1\,kHz - 10\,kHz$, which is a factor 10 larger than the expected one (i.e. on the order of $100\,Hz - 1kHz$). This underperformance was found to be caused by optical cross-talk occurring vertically between the two sensing levels of the 3D-SiCAD pixel. The probability for optical cross-talk has been measured to be rather small, i.e. less than 1% at $V_{rev} = 19{,}8V$ and just a few % at larger biases. However, this effect has been found to represent the dominant contribution on the overall observed amount of accidental counts. This limitation can be overcome by interposing a reflecting/absorbing layer between the two layers.

The detection capability of the realized prototype to ionizing radiation has been finally demonstrated by characterizing a commercial Strontium-90 radioactive source featuring a nominal activity of $37\,MBq$ by means of inverse square-law measurements, together with some preliminary measurements conducted in order to optimize the operating parameters for the device. A fitting study over the measured data helped corroborating the validity of the results, as this allowed providing a fair estimation of the activity of the chosen radioactive source which has been found to be in excellent agreement with the expectations.

More accurate results could be obtained by adopting a much longer acquisition time-window. Even if such an approach is quite time-consuming, this would certainly provide a better understanding of the device performances by means of a

finer analysis over the resulting experimental data. It would be indeed interesting to study in a more accurate way, the time-jitter affecting the quenching circuits and eventually the whole coincidence detection process. This can be done by precisely measuring both the coincidence time-window Δt and the amount of detected beta counts within the estimated time-jitter range, i.e. below a few nanoseconds down to the lowest achievable Δt. In order to push the noise rejection capability of a 3D-SiCAD device towards the best attainable performances, the study of the coincidence time-jitter can be of great interest. The ideal coincidence time-window should be indeed shortened down to few tens of picoseconds, i.e. the intrinsic time-jitter of a SPAD device, calling for an extremely challenging design for the pixel electronics.

Similarly, the particle detection efficiency deserves to be studied in more detail in order to investigate the optimal configuration for the 3D-SiCAD pixel providing the maximal signal-to-noise ratio. On the one hand, it is indeed important to bias the avalanche diode at the minimal required reverse bias in order to lower as much as possible the amount of accidental counts. On the other hand, these results have to be compared with the time-jitter study, as the electric field in the avalanche diode active region is expected to affect the timing performance of the device.

In reality, the validity of the results obtained from inverse square-law characterization should be corroborated with complementary measurements to be performed by means of a calibrated particle detector, in order to precisely evaluate the activity of the adopted Sr-90 radioactive source. This latter is indeed provided by the manufacturer in terms of "nominal value" that is affected by a rather poor uncertainty of $\pm$ 30 %. Moreover, the fraction of betas that can effectively cross both sensing levels of a 3D-SiCAD pixel (referred to as $\alpha(t_{sub})$ in Chapter 4) deserves to be evaluated in a more precise way by means of dedicated Monte Carlo simulations and, similarly, the systematic minimal distance d_0 should be controlled and fixed at a well-known value.
The timing required by the characterization of the detector could be speed-up by irradiating the device with a high fluency particle beam. A 65 MeV test-beam measurement has been indeed planned at Lacassagne Center in Nice (France). This would allow characterizing in a rather fast way other 3D pixels cells which would be of great interest in order to verify possible statistical fluctuation of the main figures of a 3D-SiCAD pixel.

As discussed in Chapter 4, optical cross-talk is a very important drawback that is currently limiting the noise rejection capability of the realized prototype. In absence of this detrimental effect, the Fake Coincidence Rate is indeed expected to be on the order of a few counts per second for the $50 \times 50\ \mu m^2$ active pixel area (i.e. a factor 100 lower than the one reported in Chapter 4) in case of a coincidence time-window of $\Delta t = 2\ ns$, by adopting the same measurement condition as for the inverse-square law one. This drawback could be solved rather easily by interposing a fully absorbing or reflecting medium between the two sensing levels such as metal planes or opaque polymers.

Currently the measurements over the available prototypes are performed with a rather low degree of "automatization". This is actually limiting the amount of devices that can be characterized and the uncertainty affecting the experimental results. Therefore, one of the most important objectives to be pursued within the mid-term is certainly the improvement, in terms of automatization, of the characterization setup. This can be supported by a parallel upgrade of the prototype electronics, which should features dedicated ASICs allowing both user-programmable parameters (such as the hold-off time, coincidence time-windows, the reverse bias etc.) and programmable read-out for a fast external control and acquisition via FPGA.

An upgraded version for the prototype should provide a detector demonstrator much closer to the final product than the present one. The upgraded demonstrator should thus consist of a pixelated architecture of 3D-SiCAD cells rather than simple single-pixel cells (or very small matrix cells). A specific improvement is actually required for the quenching electronics as it is currently not capable of detect avalanche pulses in case of excess bias voltages lower than the comparator threshold voltage at the output of the avalanche diode. This drawback has to be solved in order to allow smaller excess bias voltages which would provide much lower accidental counts and thus higher signal-to-noise ratio.
Furthermore, another important mid-term objective would consider the adoption of a real 3D integration technique based on micro-pillars featuring only a few microns pitch and through silicon vias (TSV). This would indeed enable an effective dense integration of 3D-SiCAD pixels in a CMOS device, as shown in Figure 1. Finally, the choice of a CMOS process that is more suitable for the design of SPAD detectors should be considered for the upgrade of the current design. A considerably lower DCR per pixel would allow a 3D-SiCAD pixel to achieve practically negligible levels of accidental noise counts. As discussed in the introduction (Figure 2), there exist indeed CMOS processes featuring DCR densities as low as a few fractions of $Hz\,\mu m^{-2}$ which translates, in case of a pixel size of $50 \times 50\ \mu m^2$, into an accidental counting rate of only one count every a few minutes, i.e. practically no accidental counts.

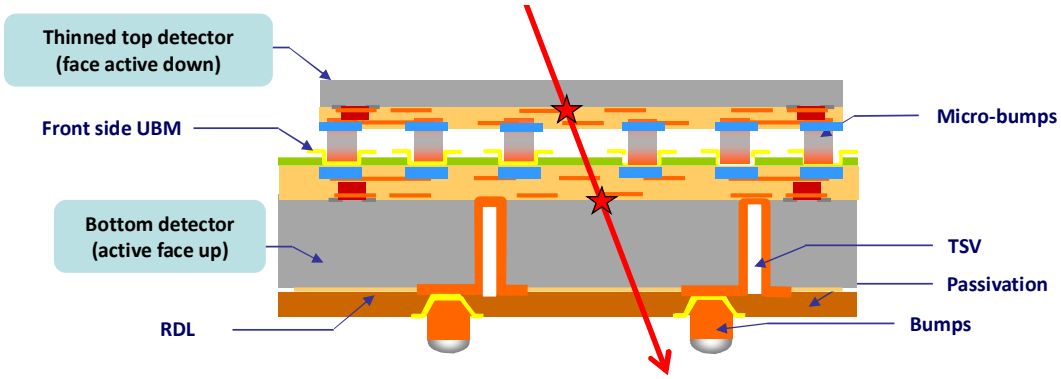

***Figure 1**: Possible 3D integration scheme (proposed by LETI)*

In the long-term development, the fill-factor should be optimized. With respect to photon counting application, where micro-lenses can be used to compensate this drawback, a non-optimal fill factor unavoidably leads to poor detection efficiency for a pixelated charged particle detector. This goal can be achieved by increasing the degree of vertical integration, by adding one or two more tiers over the stack. Two tiers would be indeed exclusively dedicated to the avalanche diodes, i.e. the sensors. The additional tiers would instead host the electronics for each pixel and for the read-out. On the one hand this approach seems to be rather straightforward. On the other hand, adding more tiers over the 3D stack would unavoidably increase the complexity of the design and thus the manufacturing costs, but more importantly there would be an increase of the material budget and an important impact on the device reliability. An alternative solution has been considered in this work, as a parallel activity to the development of a first 3D-SiCAD prototype [4][5]. The proposed solution focuses on the design of a novel architecture for the SPAD pixel cell, which makes use of the advantages and new features provided by modern FDSOI CMOS technologies (Fully Depleted Silicon-On-Insulator). This approach is expected to be also exploitable in common SPAD applications in backside illumination (BSI) mode. Compared to traditional SPAD architectures, the novel one has the great advantage to provide a monolithic 3D structure without the need of dedicated 3D integration techniques. The avalanche diode is indeed defined beneath the Buried Oxide (BOX) while the quenching electronics is sitting on top of it, in the SOI layer as depicted in Figure 2. The pixel can be designed according to the features of an advanced Fully-Depleted SOI technology, by exploiting the available implantations and diffusions that are normally meant to provide different back-biasing strategies for the transistors. The diode sensitive region is defined in the Space Charge Region (SCR) of a p-well/deep n-well junction. Premature Edge Breakdown (PEB) risk is prevented thanks to a guard-ring placed around the sensitive area. A low doped p-type region can indeed be obtained thanks to the retrograde doping of the deep n-well in the epitaxial p-type substrate. Such a region is responsible for smoothing down the electric field at the junction edge that otherwise would be too intense to allow Geiger-mode operation as shown in Figure 3.

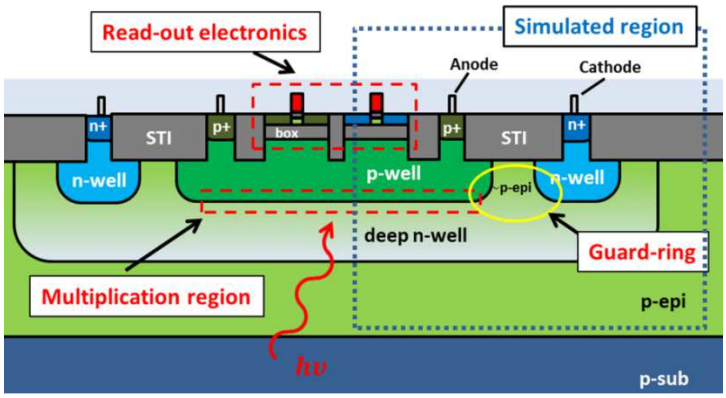

Figure 2: *Schematic representation of the proposed 3D pixel based on an advanced FDSOI technology.*

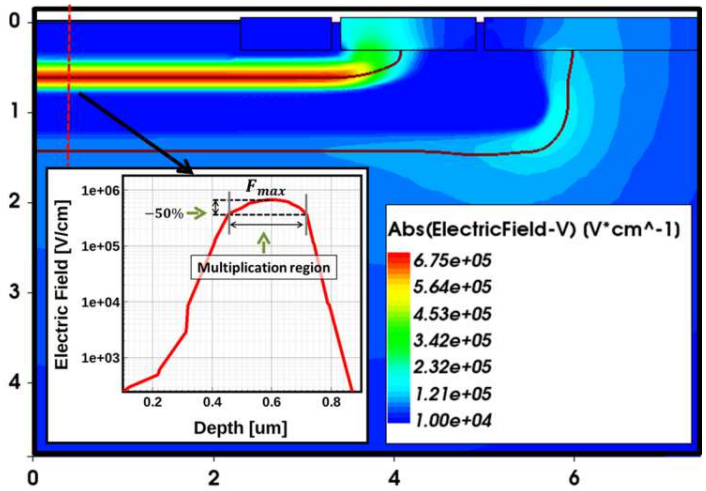

Figure 3: *TCAD simulation: Electric field color map of the pixel, when the avalanche diode is reverse biased at* $V_{rev} = 16.5\,V$ *(Spatial scales are in* μm*).*

The diode can be connected to its associated electronics thanks to back-gate contacts featured by the adopted FDSOI technology, but originally meant to enable a "tunable" threshold voltage for the transistors in the SOI. It is important to highlight that the pixel and the detector electronics can benefit of the well-known advantages brought by SOI technology with respect to bulk CMOS, such as higher speed and lower power consumption.

References

[1] M. M. Vignetti, F. Calmon, R. Cellier, P. Pittet, L. Quiquerez, and A. Savoy-Navarro, "A time-integration based quenching circuit for Geiger-mode avalanche diodes," in *New Circuits and Systems Conference (NEWCAS), 2015 IEEE 13th International*, 2015, pp. 1–4.

[2] E. V. Figueras, "Feasibility of Geiger-mode avalanche photodiodes in CMOS standard technologies for tracker detectors Feasibility of Geiger-mode avalanche photodiodes in CMOS standard technologies for tracker detectors," (PhD Thesis), University of Barcelona, 2013.

[3] 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," *Proc. SPIE 6372, Adv. Phot. Count. Tech. 63720S*, vol. 6372, p. 63720S–63720S–12, 2006.

[4] M. M. Vignetti, F. Calmon, P. Lesieur, and A. Savoy-Navarro, "Simulation study of a novel 3D SPAD pixel in an advanced FD-SOI technology," *Solid. State. Electron.*, vol. 128, pp. 163–171, 2017.

[5] M. M. Vignetti, F. Calmon, P. Lesieur, F. Dubois, T. Graziosi, and A. Savoy-Navarro, "A novel 3D pixel concept for Geiger-mode detection in SOI technology," in *2016 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, 2016, pp. 166–169.

Annex - Preliminary study of the avalanche diode architectures

This section presents a preliminary study [1] on some of the state-of-the art CMOS SPAD architectures discussed in Chapter 1, and reported in Figure 1 for convenience. In general, the architectures considered in this study, with the exception of the STI-based guard ring one, require detailed technological data for verification of the structure operation. The aim of the study was to understand the main challenges to be typically faced for the integration of a SPAD in a standard CMOS process. The study has been conducted by means of TCAD simulations and considering a 130nm standard CMOS process.

A.1 Guard-ring simulations

A.1.1 Diffused guard ring

Deep Sub-Micrometer CMOS processes typically feature an n-type deep tub in order to electrically insulate both p-channel and n-channel MOS transistors from the substrate.

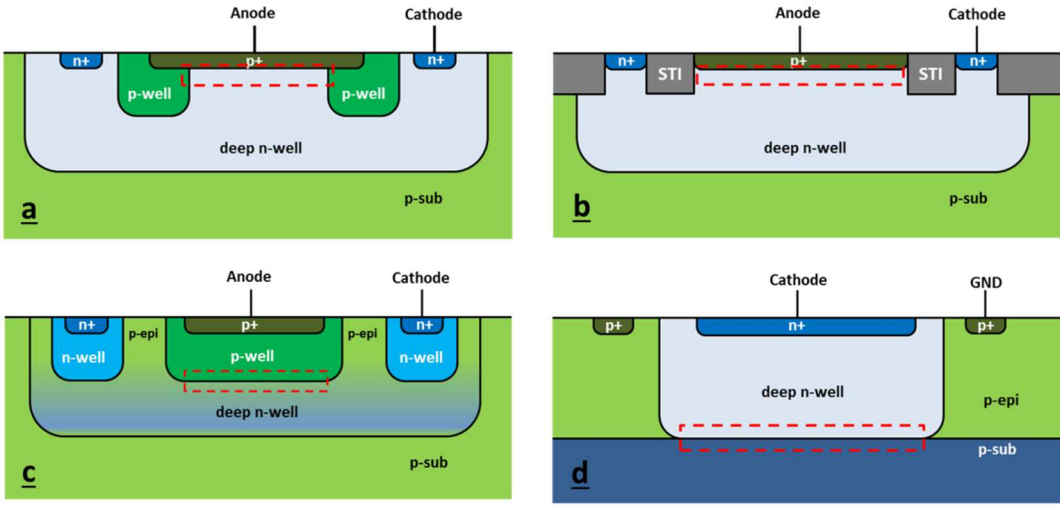

Figure 1: *Cross sections of different CMOS SPAD architectures among those found in literature. The multiplication region is highlighted with a dashed rectangle. (a) Diffused Guard Ring; (b) Shallow Trench Isolation (STI) Guard Ring; (c)retrograde-doping guard ring; (d) buried multiplying region.*

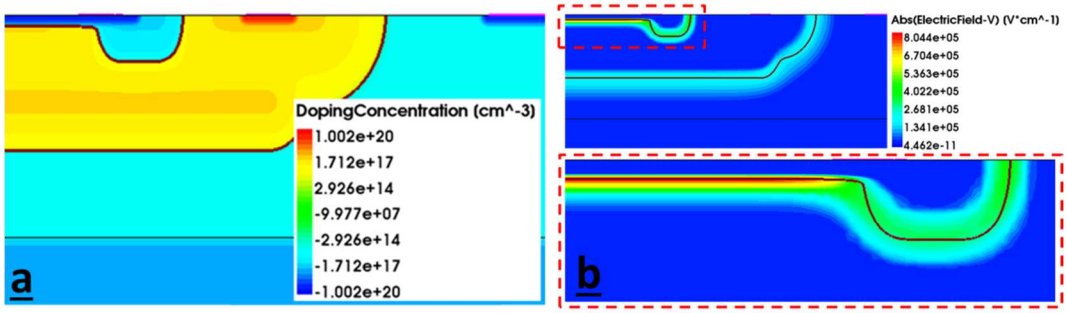

Figure 2: *Diffused guard-ring implementation in a standard 130nm CMOS process (a) Effective doping concentration color map (Positive doping concentration values refer to n-type doping whereas negative one refers to p-type) (b) Electric field color map.* $V_{rev} = 16.32V$ *(simulated* $V_{bd} = 11.66V$*).*

As shown in Figure 2a, a diffused guard-ring implementation in a 130nm CMOS process consist of a p+/n junction, obtained by means of a p+ diffusion (used for the source and drain regions of MOS transistors) over the deep n-well, while a low doped p-well (commonly used as bulk region for insulated n-type MOS transistors) surrounds the junction, defining the guard ring region. It is important to stress that the layout of such a SPAD architecture does not fully comply with the CMOS process design rules (intersection of diffusion and wells is not allowed). However foundries, in general, allow waiving these rules, but this has to be done very carefully. Figure 2b shows the electric field color map of the diffused guard-ring architecture of Figure 2a, obtained by means of TCAD simulations. As expected, premature breakdown at the device edge is successfully prevented, as the electric field over the multiplying region is higher than at the periphery.

It is interesting to evaluate the effectiveness of this solution in preventing PEB when the CMOS process features different parameters with respect to the nominal adopted values, by considering different doping profile scenarios. For this purpose, it is helpful to introduce a parameter able to quantify the effectiveness of PEB prevention with respect to both the CMOS process and avalanche diode parameters. Such a parameter can be defined as the ratio between the electric field peaks $F^{max}_{p+/nwell}$ and $F^{max}_{p-gr/nwell}$ in the p+/n-well and p-well (guard-ring)/n-well junctions respectively:

$$\alpha = \frac{F^{max}_{p+/nwell}}{F^{max}_{p-gr/nwell}} \tag{1}$$

If α is lower than one, PEB occurs at the device edges. A useful analytical expression for α can be derived under the assumption that the guard-ring/n-well junction is approximated by a linearly graded junction in order to account for the Gaussian distribution of the doping concentration. Besides, it has been shown in [2] that under this assumption, the electric field distribution is relatively independent on the junction radius of curvature. The resulting formula is [1]:

$$\alpha_{linear} \approx \frac{4}{3}\left[\frac{9q}{2\varepsilon}\frac{x_0^2}{V_{rev}}N_{n-well}\ln^{-2}\left(1+\frac{N_{p-gr}}{N_{n-well}}\right)\right]^{\frac{1}{6}} \quad (2)$$

where q is the electron elementary charge, ε is the silicon dielectric constant, N_{p-gr} and N_{n-well} are the doping concentrations in the p-type guard-ring and n-type deep tub respectively, x_0 is the guard-ring junction depth, and V_{rev} is the reverse bias voltage applied across the junction. Figures 3 shows a plot of the α parameter extracted from TCAD simulations as well as by using the linearly graded junction model (2) for a p+/n-well diffused guard-ring architecture. The curves have been obtained by varying both the doping concentration and the junction depth of the p-well guard-ring. Two different background doping concentrations, i.e. deep n-well doping, of $2\times10^{17}cm^{-3}$ or $2\times10^{16}cm^{-3}$ have been used leading to a p+/n-well breakdown voltage of 11.7V or 39V respectively (N.B.: the breakdown voltage is evaluated by the simulation tool by adopting the Van Overstraeten, De Man model for the impact ionization coefficients [3]). The reverse bias voltage has been chosen according to the p+/n-well junction breakdown in order to have the diode biased above around 30% of its breakdown voltage. The linearly graded junction model matches with the simulated data within around 20% error. For this reason a correction factor of 1.22 and 1.18 for the higher and lower background doping case respectively has been adopted in order to improve the match with the simulated data but it is worth to note that (4) predicts pretty well the simulated trend. On the one hand when $N_{background} = 2\times10^{17}cm^{-3}$, the effectiveness of PEB prevention is enhanced when decreasing the guard-ring doping concentration and increasing the guard-ring junction depth. Indeed the gradient of the net doping concentration of the p-type guard-ring / n-well junction decreases which results in a lower electric field peak in the space charge region.

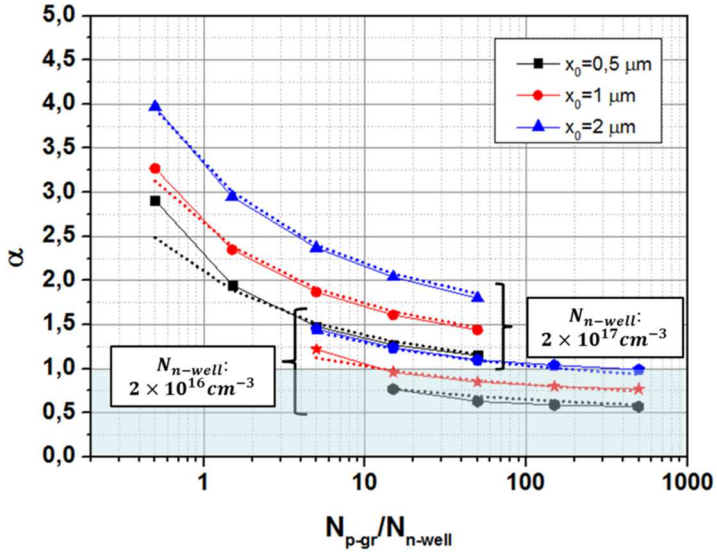

Figure 3: *α factor for a n+/p diode for $N_{background} = 2\times10^{17}cm^{-3}$ (upper curves) and $N_{background} = 2\times10^{16}cm^{-3}$ (lower curves). Dotted curves results from the implementation of the linearly graded model (2).*

On the other hand, when $N_{background} = 2 \times 10^{16} cm^{-3}$, PEB might occur if the junction is not deep enough and/or if the doping concentration is too high. In this case the background doping level is indeed rather low which widens the guard-ring/background space charge region. As a consequence of that, the guard-ring curvature effects become stronger, resulting in a higher electric field in the guard-ring region, i.e. a lower α factor. In support of this argument, it should be observed that the α factor is higher if the guard-ring is based on a deeper p-well (a larger radius of curvature according to the geometrical model adopted in the present work). A risk of PEB appears at the p+ edge for guard-ring doping levels much lower than the background doping as shown in Figure 4. Due to the low guard-ring doping concentration and small guard-ring junction depth, the space charge region punches-through the p^+/p^- junction as shown in Figure 4 for $x_0 = 0.5\mu m$. The reverse bias drops on the guard-ring/background junction but also on the p^+/p^- one. Due to the high curvature effects on the p+ diffusion edge, the electric field will be very high, and the beneficial effect of the guard-ring is lost. This issue is not observed in Figure 3 because the considered N_{gr}/N_{n-well} ranges are such that the "punch through" doesn't occur. It is indeed pointless to define an α factor outside that range, that is, where the electric field peak displaces from the guard-ring to the p+ diffusion edge. Premature breakdown at the device periphery might actually be avoided by placing the p-type well at a small distance from the edge of the p+ diffusion region, in principle without violating any design rules. An implementation of such a structure is shown in Figure 5a where the avalanche diode consists of a p+/n-well junction surrounded by a low doped p-well region.

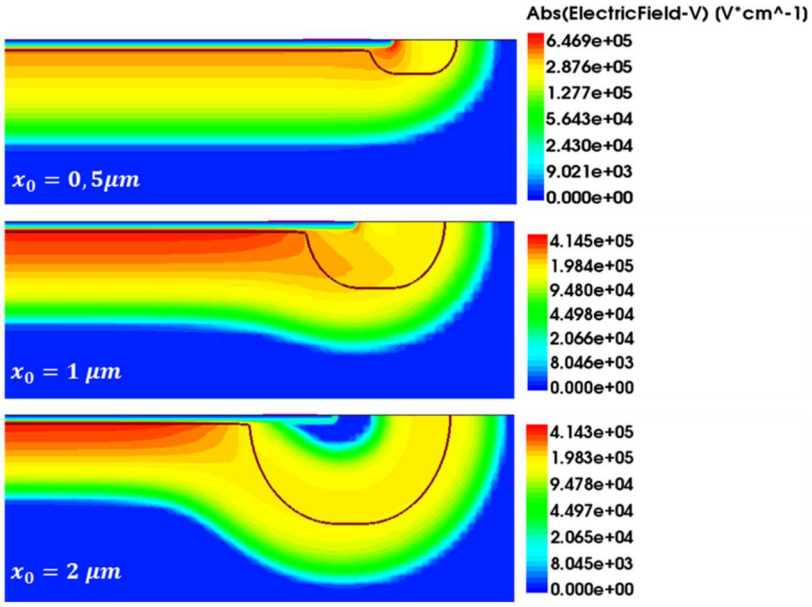

Figure 4: *Electric field color map of a diode having a very low-doped guard ring for different guard-ring depths* x_0. *Observe that* $N_{background} = 2 \times 10^{16} cm^{-3}$.

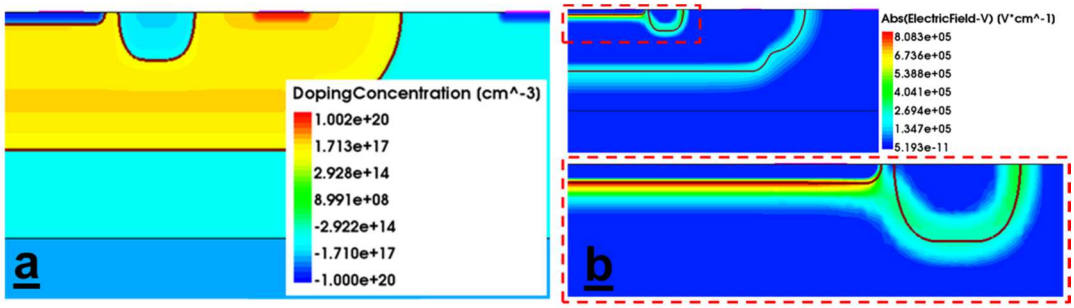

Figure 5: *Slightly spaced diffused guard-ring implementation in a standard 130nm CMOS process (a) Effective doping concentration color map (Positive doping concentration values refer to n-type doping whereas negative one refers to p-type) (b) Electric field color map of a p+/n-well diode surrounded by a p-guard ring spaced by d=100nm.* $V_{rev} = 16.3V$ *(simulated* $V_{bd} = 11.7V$*).*

In this architecture, the lateral diffusions of the doping profile in the p+ and p-well regions reduce "by compensation" the n-type doping in the gap between them. However, this topology also slightly reduces the fill-factor of the sensor with respect to the common diffused guard-ring architecture of Figure 2. Figure 5b shows the electric field color map of the avalanche diode demonstrating a successful PEB prevention when a gap $d = 100nm$ is adopted. It is important to point out that the effectiveness of this solution depends on the doping concentration of both the p-well and n-well regions and on the gap d, as summarized in Figure 6. An α factor (i.e. the ratio between the electric field in the multiplication region and the electric field at the edge of the p+ diffusion) has been plotted as a function of the distance d between the p+ diffusion and the p-well guard-ring for different guard-ring doping concentrations (Figure 6).

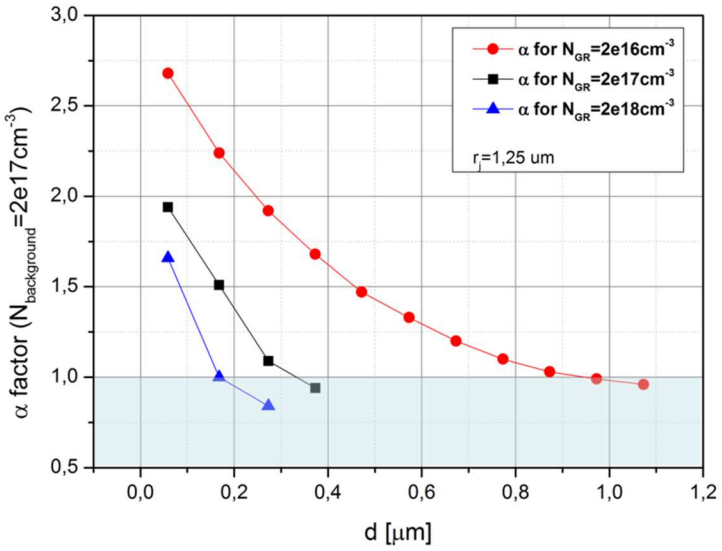

Figure 6: *α factor for a p+/n diode having* $N_{background} = 2 \times 10^{17} cm^{-3}$ *when a "slightly spaced low doped guard ring" PEB technique is implemented for different guard-ring doping concentrations.*

If the background doping is in the $10^{17}cm^{-3}$ range (which is a typical value for n-wells in deep sub-micrometer technologies) such a PEB prevention technique works pretty well for a guard-ring doping concentration of ~$10^{16}cm^{-3}$ within a wide range gap values. Conversely, it shows some lack of robustness for higher doping levels. The minimum distance allowed by the process design rules is typically on the order of the CMOS technology node, i.e. 130nm in the present case, which means that such a PEB prevention technique should work even for higher values of guard-ring doping concentration. This discussion highlights the importance to have full knowledge of the process parameters of the adopted CMOS technology in order to assess a correct operation of the desired avalanche diode architecture.

A.1.2 Shallow Trench Isolations guard ring

From the 250nm node, standard CMOS processes feature shallow trenches filled with SiO_2, commonly referred to as Shallow Trench Isolations (STI), etched all around p+ and n+ implantation areas in order to effectively prevent punch through and latch-up effects in CMOS circuits. A fully CMOS compatible guard-ring implementation can be thus realized by exploiting these isolations as shown in Figure 1b, and whose implementation in a 130nm technology is reported in Figure 7a. This solution can provide great scalability thanks to an effective physical electric field confinement due to the isolation trench surrounding the multiplying region as resulting from TCAD simulations, reported in Figure 7b. An excellent fill-factor can be obtained since the dielectric strength of SiO_2 is 30 times higher than the breakdown field of silicon which allows a 30 narrower STI guard-ring with respect to a diffused one. However, as discussed in Chapter 1, the etching process may cause important damage in the surroundings of STI regions. These are indeed sitting next to the multiplying region of the SPAD, which may lead to very high DCR values (up to a few MHz [4]).

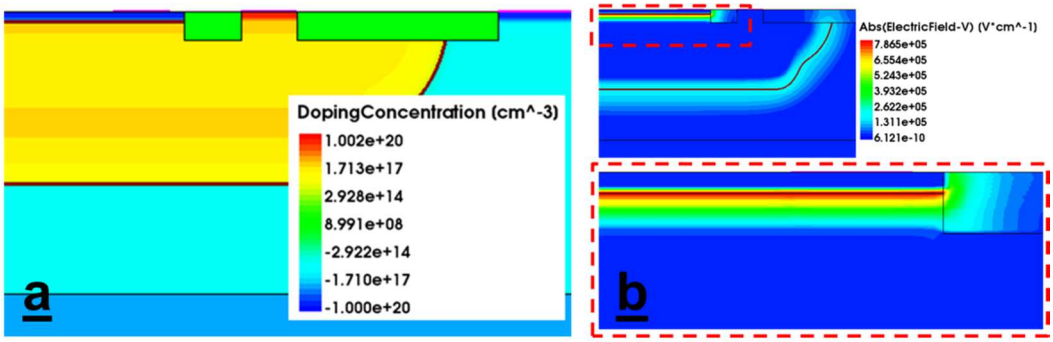

Figure 7: *STI-based guard-ring implementation in a standard 130nm CMOS process (a) Effective doping concentration color map (Positive doping concentration values refer to n-type doping whereas negative one refers to p-type) (b) Electric field color map.* $V_{rev} = 14V$ *(simulated* $V_{bd} = 11.6V$*).*

A.1.3 Retrograde-doping guard ring

This solution is depicted in Figure 1c and is based on an active region defined by a p-well / deep n-well junction. If no n-well is drawn together with deep n-well (and provided that p-well formation is explicitly prevented) the result is indeed a "buried n-well" at a certain depth, with n-type doping concentration progressively diminishing towards the upper surface of the SPAD. Therefore the p-well can be seen as a p-type doping enrichment in the retrograde n-type doping, which enhances the electric field in a well localized region, as for the "virtual" guard-ring case. The buried n-well is naturally connected to the cathode contacts by means of n+/n-well layers drawn at the outer edge. Figure 8b shows the TCAD simulated electric field color map of the retrograde-doping guard-ring architecture implemented in a 130nm CMOS process (Figure 8a). As expected, premature breakdown at the device edge is successfully prevented, as the electric field over the multiplying region is higher than at the periphery.

A.1.4 Buried multiplying region

A well localized and PEB-free high-field region can be defined by the junction between a deep n-well and a thin p-type epitaxy on a low resistivity p-substrate, as shown in Figure 1d, resulting in a "virtual" guard-ring like solution. The electric field magnitude across the lateral p-epi/deep n-well junction is expected to be certainly lower than the field maximum across the multiplication junction resulting from the large difference of doping levels between the p-substrate and the epitaxial layers. This consideration is confirmed by TCAD simulations reported in Figure 9. In reality, it is important to point out that, depending on the thickness of the thin p-type epitaxy, this solution may work or not as shown in Figure 9b,c. Therefore the implementation of this guard-ring topology requires detailed knowledge of the process parameters of the adopted CMOS technology in order to validate the architecture for a correct PEB Geiger-mode operation.

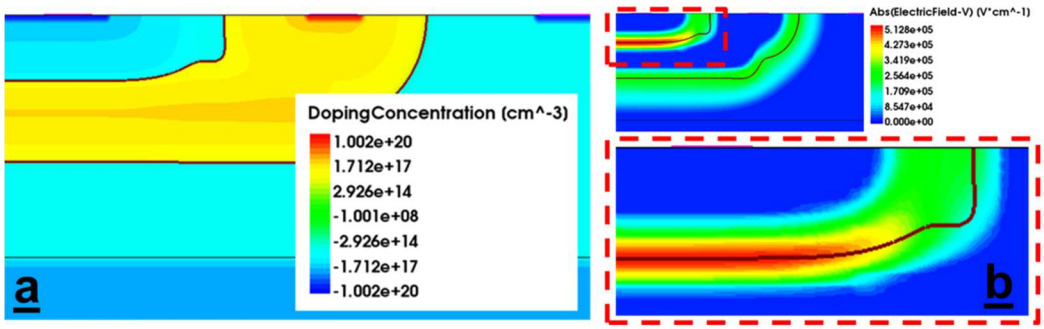

Figure 8: *Retrograde-doping guard-ring implementation in a standard 130nm CMOS process (a) Effective doping concentration color map (Positive doping concentration values refer to n-type doping whereas negative one refers to p-type) (b) Electric field color map.* $V_{rev} = 30V$ *(simulated* $V_{bd} = 20.24V$*).*

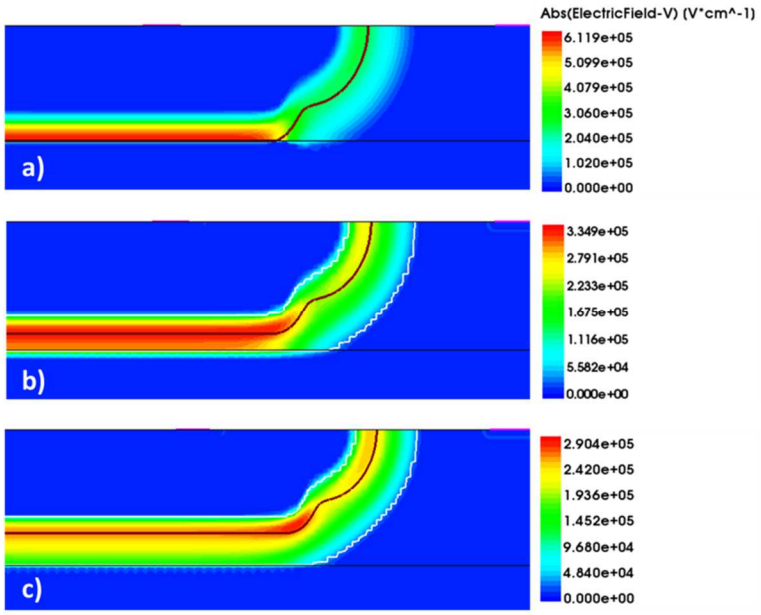

Figure 9: *Buried multiplying region implementation in a standard 130nm CMOS process. Electric field color maps for three different p-type epitaxy thickness value: (a) no p-epitaxy in between deep n-well and p-substrate; (b) thin p-epitaxy: PEB is prevented anyway; (c) thick p-epitaxy: PEB is not prevented anymore.* $V_{rev} = 30V$ *(simulated* $V_{bd} = 25.1V$*)*

A.2 Noise Considerations

The development of a CMOS SPAD requires the optimization of the noise performance of the device, quantified in terms of Dark Count Rate, as discussed in Chapter 1. For this reason it is very important to look for an avalanche diode architecture that minimizes the main carrier generation mechanisms occurring device active region, such as the Shockley-Read-Hall (SRH) generation-recombination rate (G_{SRH}), the band-to-band tunneling rate, (G_{B2B}) and the trapping/de-trapping rate due to STI interface traps, (G_{STI}). In general, the architectures relying on a p+/n-well junction, i.e. Figure 1a,b, are noisier than those based on deep n-well/p-sub or p-well/ deep n-well junctions, i.e. Figure 1c,d, especially in deep sub-micrometer technologies. The formers show indeed a considerable band-to-band tunneling generation which dramatically enhances the dark counts. Furthermore, the STI-based guard-ring architecture suffers of additional noise degradation due to deep-level carrier generation centers at the Si/SiO_2 interface. Even though this effect is negligible for large diameter avalanche diodes (the STI defects enhancement is a peripheral effects), it is in general recommended to move STI away from the multiplication region.
In this section these aspects are examined in more details.

A.2.1 Band-to-band tunneling

In deep sub-micrometer CMOS technologies the main contributor to the DCR tends to switch from Shockley-Read-Hall processes to tunneling. The high doping concentration levels cause indeed a very narrow depletion region with a very intense electric field, resulting in a significant number of tunneling-induced carriers and increased DCR. In particular, the doping levels of the n-well and p+ (p-diffusion) forming the avalanche breakdown p-n junction are excessively high. Such junctions and wells are required for reasons of efficient PMOS transistor formation in scaled CMOS technologies. However, they are contrary to the requirements for avalanche photodiode operation. The band-to-band (B2B) tunneling in three different avalanche diode structures (described in Section A.1) has been studied by means of TCAD simulations by adopting the "E2 model" [3] for the band-to-band generation rate per unit volume, provided by the TCAD simulation tool. The study is limited to the comparison of tunneling current contribution trend of the different structures. Figure 10 shows the current density contribution due to band-to-band tunneling generation as a function of a normalized cathode-to-anode voltage defined as the ratio between the actual cathode-to-anode voltage and the breakdown voltage of the simulated device. Additionally, SRH current densities are indicated for the three diode architectures. In this way, it has been possible to compare the behavior of devices having different breakdown voltages for the same excess bias-to-breakdown voltage ratio (the avalanche triggering probability does not depend on the excess bias voltage value itself, but on the V_{ex}/V_{bd} ratio [5]. In order to evaluate the band-to-band current density the multiplication region has been defined as the flat region of the p-n junction where the electric field is within 10% of its maximum value. The plots shown in Figure 10 correspond to the difference between the current at the cathode when switching from ON to OFF the B2B tunneling model while maintaining the avalanche model OFF in the simulator.

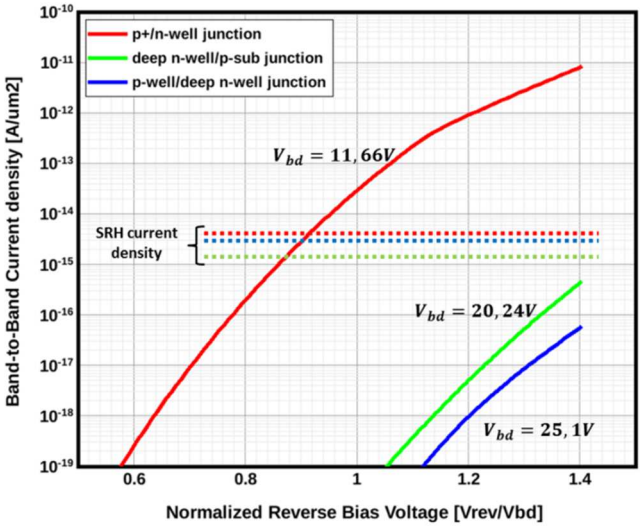

***Figure 10**: Band-to-band current density contribution for different avalanche diode architectures as a function of a normalized voltage (defined as the ratio of the applied bias over diode breakdown voltage).*

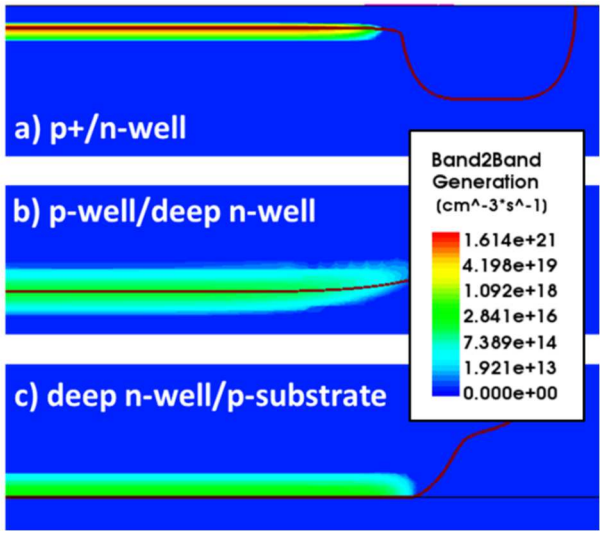

***Figure 11**: B2B generation rate color maps (arbitrary units) for the different avalanche diode architectures for an excess bias of 40%.*

It is important to notice that the plotted tunneling current densities are directly proportional to the B2B generation rate and that they come only from the multiplication region as the tunneling probability is reasonably negligible outside the high-field region. By observing Figures 10 and 11 (in the latter an excess bias ratio of 1.4 is applied), it is quite clear that the tunneling current contribution is pretty considerable in the p+/n-well avalanche diode while it is negligible in the other devices with respect to SRH current. The behavior of the p+/n-well device approaches to a Zener-like diode due to the high doping in the device active region [6]. In this specific case the tunneling process is indeed in competition with the avalanche multiplication. On the one hand, the multiplication process requires a certain electric field and a certain distance for the impact ionization to occur and generate a self-sustained process. On the other hand, the tunneling process requires only a certain electric field to reach a sufficient probability to occur. A confirmation of the DCR improvement enabled by avalanche diodes based on p-well/deep n-well junctions and deep n-well/p-substrate junctions can be found in literature [7] [8] [9] as summarized in Table 1. The breakdown voltages shown in Table 1 follow a trend (with respect to the diode topology) that is coherent with the values obtained from TCAD simulations. However the measured data differs within 20 – 30 % from the simulated one. This may be due to differences in the doping profiles between the simulated diodes and the real ones. Unfortunately, DCR data is not directly available from TCAD simulations but it is still possible to make a few considerations about the noise performance, since the dark current is somehow related to the dark count rate. For a 10% excess bias, the p+/n-well device shows a total dark current (i.e. B2B + SRH current) which is 130 and 65 times larger than the ones in the deep n-well/p-sub and p-well/deep n-well device respectively. Such ratios are on the same order of the ones between the DCR data measured in literature and shown in Table 1.

Table I: V_{bd} and DCR data for avalanche diodes based on a) a p+/n-well junction, b) a p-well/deep n-well junction and c) a deep n-well/p-substrate junction. *extrapolated from Fig. 14 in Ref [7]

130nm CMOS technology
Applied Excess Bias = 10% of $\boldsymbol{V_{bd}}$

Active Junction Type	V_{bd} [V]	DCR [Hz]	Area[μm^2]	DCR' [$Hz/\mu m^2$]
a) p+/n-well [9]	9.74	15k	78.5	191
b) p-well/dn-well [8]	14.36	80*	50	1.6
c) dn-well/p-sub [7]	20	18	50	0.36

This is of course just a first order estimation, confirming that the trend predicted by simulations is reproduced in experimental data. However, it is important to stress that the dark current does not necessarily relate one-to-one to the DCR as part of the dark current flows through the guard ring, where an avalanche breakdown cannot be triggered.

A.2.2 Shallow Trench Isolations induced Noise

In this section, the simulation study focuses on the impact of STI on the dark current and consequently on the DCR of different avalanche diode architectures.
The defects due to the imperfections at the Silicon/STI interface have been modeled as interface traps with donor and acceptor states. According to [10] the donor and acceptor states in the silicon bandgap have been described in a simplified way by two Gaussian distributions centered respectively at 0.25eV and 0.8eV above the valence band, each distribution having a standard deviation of 0.1eV. In order to consider the very worst case condition both the donor and acceptor states have a total concentration $D_i = 2 \times 10^{12} cm^{-2}$ (i.e. quite high densities are considered in this study) and electron and hole cross sections of $10^{-14} cm^2$ [10].
An avalanche diode with an STI based guard-ring has a very important drawback. Part of the Si/STI interface falls indeed inside the multiplication region of the avalanche diode. This means that electrons/holes released by the trap states at the Si/STI interface are directly injected in the high field region of the avalanche diode enhancing consequently the total DCR.
This observation can be better understood by observing the results obtained from TCAD simulations and illustrated in Figure 12 where the diode anode dark current density as a function of the reverse bias has been plotted for different values of the anode radius (B2B generation and avalanche model are OFF). By sensing the dark current at the anode of the device, only the current due to SRH genera-

tion in the multiplication region and the carriers injected from the Si/STI interface into it is extracted. In this way, it is possible to obtain an indication of the total generation rate in the multiplication region and therefore the enhancement of the device DCR. In Figure 12, the dark current density when the device is assumed to be defects-free (dotted curve) is one order of magnitude lower than the one obtained when STI defects are taken into account (red curve). It is also really important to observe that if the anode diameter increases the current density curves approach the defects-free one meaning that the degradation of the DCR produced by the STI guard ring becomes more and more negligible for large area devices. That is because the STI induced DCR is proportional to device perimeter while the other DCR components are proportional to the diode active area. It is worth noting, as well, that the contribution of STI to the dark current in case of a p+/n-well junction, i.e. the present case, is lower (or at most around the same order of magnitude) than the B2B tunneling contribution as it can be clearly observed by comparing the red curve of Figure 10 and Figure 12. According to the previous discussion, a small diameter avalanche diode cannot be implemented by adopting an STI-based guard-ring as the DCR would be unacceptably high. Special care must be taken in the design of an avalanche diode when migrating from sub-micrometer to deep sub-micrometer technologies in order to mitigate as much as possible electron/hole injections from STI/Si interface defects to the diode active region. In deep sub-micrometer CMOS processes, Shallow Trench Isolations are automatically etched all around p+ and n+ implantation areas in order to effectively prevent punch through and latch-up effects in CMOS circuits. As discussed in Chapter 1, a possible way to deal with this technological drawback consists on drawing “dummy” poly-silicon gate of a standard transistor in the region surrounding the p+ diffusion, i.e. where the STI has to be moved away, as proposed by Niclass et al. [9]. In order to prevent a high-electric field in the thin oxide layer below the poly-silicon gate, this one is kept at the same potential as the p+ anode.

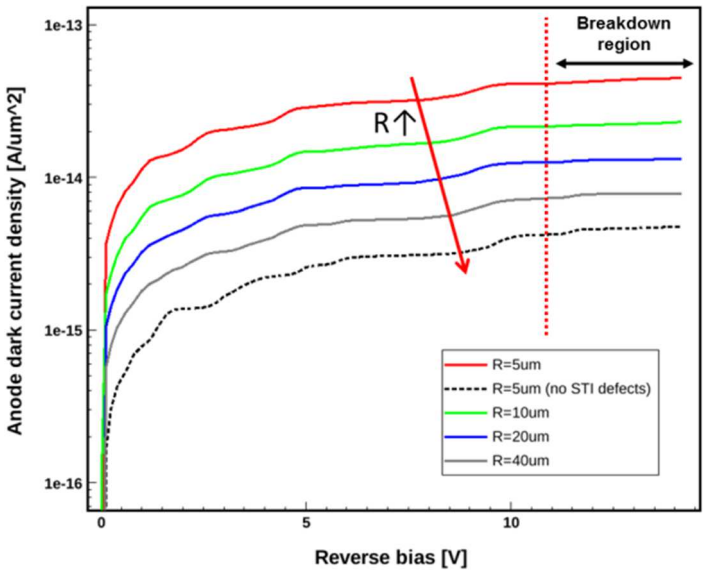

***Figure 12**: Anode dark current density as a function of the reverse bias for different values of the anode radius*

Moreover it is interesting to observe that the p-well guard-ring would act, in this case, as a passivation layer for the STI interface, which can probably mitigate the DCR enhancement. This observation was exploited in the work of Gersbach et al. [11], by surrounding the STI with several passivation implants, in a glove-like p-type structure. At the STI interface the doping level is high, which results in a very short mean free path for the minority carriers generated at the Si/SiO2 interface. As the distance from this surface increases, the doping concentration is reduced in order to lower the electric field at the p+ edge, thus preventing PEB. Table 2 shows the impact on the DCR due to STI vicinity to the multiplication region in p+/n-well based avalanche diodes. The numbers displayed in the table come from measurements found in literature which are strongly technology dependent. However it is observed that DCR decreases if the STI is moved away from the multiplication region. This is in accordance with the previous discussion. Concerning the avalanche diode architectures based on deep n-well/p-sub junction and p-well/deep n-well junction, the DCR enhancement is negligible as in these cases the multiplication region is far enough from the Si/STI interface.

Table II: V_{bd} and DCR data for p+/n-well junction based avalanche diodes having a) an STI-based guard-ring, b) a p-well based guard-ring + STI (passivated STI) and c) a p-well based guard-ring (STI "free"). (1)180nm CMOS technology, (2) 130nm CMOS technology, (3) extrapolated from Fig. 8 in Ref [11].

130nm CMOS technology
Applied Excess Bias = 7% of $\boldsymbol{V_{bd}}$

Guard-ring Type	V_{bd} [V]	DCR [Hz]	Area[μm^2]	DCR' [Hz/μm^2]
a) STI-based(1) [4]	11	1M	196	5.1 k
b) STI+passivation(2)[11]	9.4	40k (3)	58	690
c) STI "free"(2) [9]	9.7	6.5k	78.5	83

References

[1] M. M. Vignetti, F. Calmon, R. Cellier, P. Pittet, L. Quiquerez, and A. Savoy-Navarro, "Design guidelines for the integration of Geiger-mode avalanche diodes in standard CMOS technologies," *Microelectronics J.*, vol. 46, no. 10, pp. 900–910, 2015.

[2] S. M. Sze and G. Gibbons, "Effect of junction curvature on breakdown voltage in semiconductors," *Solid. State. Electron.*, vol. 9, no. 9, pp. 831–845, 1966.

[3] "Sentaurus Device User," no. March, 2013.

[4] H. Finkelstein, M. J. Hsu, and S. Esener, "An ultrafast Geiger-mode single photon avalanche diode in," vol. 6372, pp. 1–10, 2006.

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

[6] S. M. Sze and K. Ng, *Physics of Semiconductor Devices*, 3rd Editio. 2006.

[7] 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 Lett.*, vol. 33, no. 11, pp. 1589–1591, 2012.

[8] 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 Trans. Electron Devices*, vol. 58, no. 7, pp. 2028–2035, 2011.

[9] C. Niclass, M. Gersbach, R. K. Henderson, L. a. Grant, and E. Charbon, "A Single Photon Avalanche Diode Implemented in 130-nm CMOS Technology," *Sel. Top. Quantum Electron. IEEE J.*, vol. 13, no. 4, pp. 863–869, 2007.

[10] L.-Å. Ragnarsson and P. Lundgren, "Electrical characterization of Pb centers in (100)Si–SiO2 structures: The influence of surface potential on passivation during post metallization anneal," *J. Appl. Phys.*, vol. 88, no. 2, 2000.

[11] M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, and E. Charbon, "A low-noise single-photon detector implemented in a 130 nm CMOS imaging process," *Solid. State. Electron.*, vol. 53, no. 7, pp. 803–808, 2009.

List of publications

Journal articles as first author

M. Vignetti et al. "*Simulation study of a novel 3D SPAD pixel in an advanced FD-SOI technology*", *Solid State Electronics Journal (Elsevier)*, vol. 128, pp. 163–171, 2017. (http://dx.doi.org/10.1016/j.sse.2016.10.014)

Vignetti et al. "*Design guidelines for the integration of Geiger-mode avalanche diodes in standard CMOS technologies*", Microelectronics Journal, vol. 46, no. 10, pp. 900–910, 2015. (http://dx.doi.org/10.1016/j.mejo.2015.07.002)

Vignetti et al. "*Preliminary simulation study of a Coincidence Avalanche Pixel Sensor*", *Journal of Instrumentation*, vol. 10, no. 6, 2015. (http://dx.doi.org/10.1088/1748-0221/10/06/C06007)

Conference contributions as first author

Vignetti et al. "*Development of a 3D Silicon Coincidence Avalanche Detector for Charged Particle Tracking in Medical Applications*", IEEE Nuclear Science Symposium and Medical Imaging Conference", Strasbourg (France), 29 October – 6 November 2016.

Vignetti et al. "*A novel 3D pixel concept for Geiger-mode detection in SOI technology*" Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon. Vienna (Austria), January 25-27, 2016. (http://dx.doi.org/10.1109/ULIS.2016.7440079)

Vignetti et al. "*A time-integration based quenching circuit for Geiger-mode avalanche diodes*", 13th IEEE International New Circuits and Systems Conference (NEWCAS 2015), Grenoble (France), June 7-10, 2015. (http://dx.doi.org/10.1109/NEWCAS.2015.7182007)

Oral presentation at the INFIERI international workshops

M. Vignetti "New developments on Avalanche Pixels and R&D results" - INFIERI 8th Workshop, Fermi National Labs (USA), 17 - 21 October 2016. (https://indico.cern.ch/event/557734/contributions/2322855/attachments/1357210/2052316/Vignetti_-_8th_INFIERI_Workshop_-_FNAL_-_October_2016_public.pptx)

M. Vignetti "The WP2-INFIERI prototype" - INFIERI 7th Workshop, Lisbon (Portugal), 12 - 15 April 2016. (https://indico.cern.ch/event/497415/contributions/1176909/attachments/1258434/1858762/Vignetti_-_7th_INFIERI_Workshop_-_Lisbon_-_April_2016_-_Public.pptx)

M. Vignetti "Update on the Development of Novel Pixel Sensors based on 3D CMOS technology" - INFIERI 6th Workshop, Pisa (Italy), 26 - 29 October 2015. (https://indico.cern.ch/event/404880/contributions/1849120/attachments/1176978/1702106/Vignetti_-_6th_INFIERI_Workshop_-_Pisa_-_Oct_2015.pptx)

M. Vignetti "WP2 INFIERI application" - INFIERI 5th Workshop, CERN (Switzerland), 27 - 29 April 2015. (https://indico.cern.ch/event/381514/contributions/901451/attachments/760397/1043069/Vignetti_-_5th_INFIERI_Workshop_-_CERN_-_April_2015.pdf)

M. Vignetti "Development of Novel Pixel Sensors based on 3D CMOS technology for tracking devices", 4th INFIERI Workshop, Amsterdam (The Netherlands), 10 - 12 December 2014. (https://indico.cern.ch/event/352552/session/0/contribution/23/material/slides/0.pdf)

Résumé long en français

1. Introduction générale

Le développement de nouveaux détecteurs de particules chargées est nécessaire pour l'instrumentation dans les domaines de la physique des particules à haute énergie, l'astrophysique et la physique médicale avec des spécifications très exigeantes en termes de résolutions spatiale et temporelle, tenue aux radiations, épaisseur du détecteur (« material budget »), consommation, coût et modularité.

Les expériences (CMS, ATLAS, ALICE, LHCb) mises en place dans l'accélérateur de particules LHC - Large Hadron Collider à Genève) utilisent des couches de détecteurs très proches du point d'interaction pour déterminer les trajectoires des particules afin de les identifier [1]. Ces détecteurs doivent présenter une bonne résolution spatiale de $50 \times 50\,\mu m^2$ ou $25 \times 100\,\mu m^2$ avec un taux d'événements à détecter de $40\,MHz$ ($25\,ns$) pour l'upgrade du LHC (high luminosity LHC) [2][3].

De nouvelles techniques médicales ont vu le jour, telles le traitement du cancer par faisceau de particules chargées (hadrontherapy utilisant des faisceaux de protons ou d'ions carbone) et la radiographie / scanner proton (pCT). Une couche de détecteurs est nécessaire pour i) le contrôle qualité du faisceau (ex. cartographie-profil du faisceau), ii) le contrôle en temps réel du parcours des ions en l'associant à la détection de particules secondaires telles que des rayons gamma prompts [4][5], et iii) la radiographie - tomographie proton. La taille souhaitable des pixels est du même ordre qu'en physique des particules. Le taux d'événements dépend de l'application ; il est très bas pour la tomographie pCT ($< 100\,Hz$ pour une taille de pixel de $50 \times 50\,\mu m^2$) et très élevé ($> 1\,MHz$) pour réaliser la fonction hodocope (étiquetages temporel et spatial des pulses de protons).

L'objectif de ce travail de recherche est de développer un nouveau détecteur appelé *3D Silicon Coincidence Avalanche Detector* (3D-SiCAD) [6][3]. Celui-ci repose sur l'alignement vertical de diodes SPAD (Single Photon Avalanche Diode) avec un mode de fonctionnement en coïncidence permettant d'obtenir les événements corrélés temporellement (associés au passage d'une seule particule chargée qui déclenche les deux niveaux quasi-simultanément), cf. *Figure 1*. L'avantage attendu est de pouvoir discriminer les vrais événements, du bruit intrinsèque du simple pixel, grâce à une courte de fenêtre temporelle pendant laquelle les événements simultanés des deux SPADs liés à la trajectoire d'une particule unique sont détectés (*Figure 2*). Nous cherchons ainsi à développer un prototype de 3D-SICAD utilisant une technologie CMOS haute tension $0{,}35\,\mu m$ et une technique d'assemblage 3D de type flip-chip avec des billes en or.

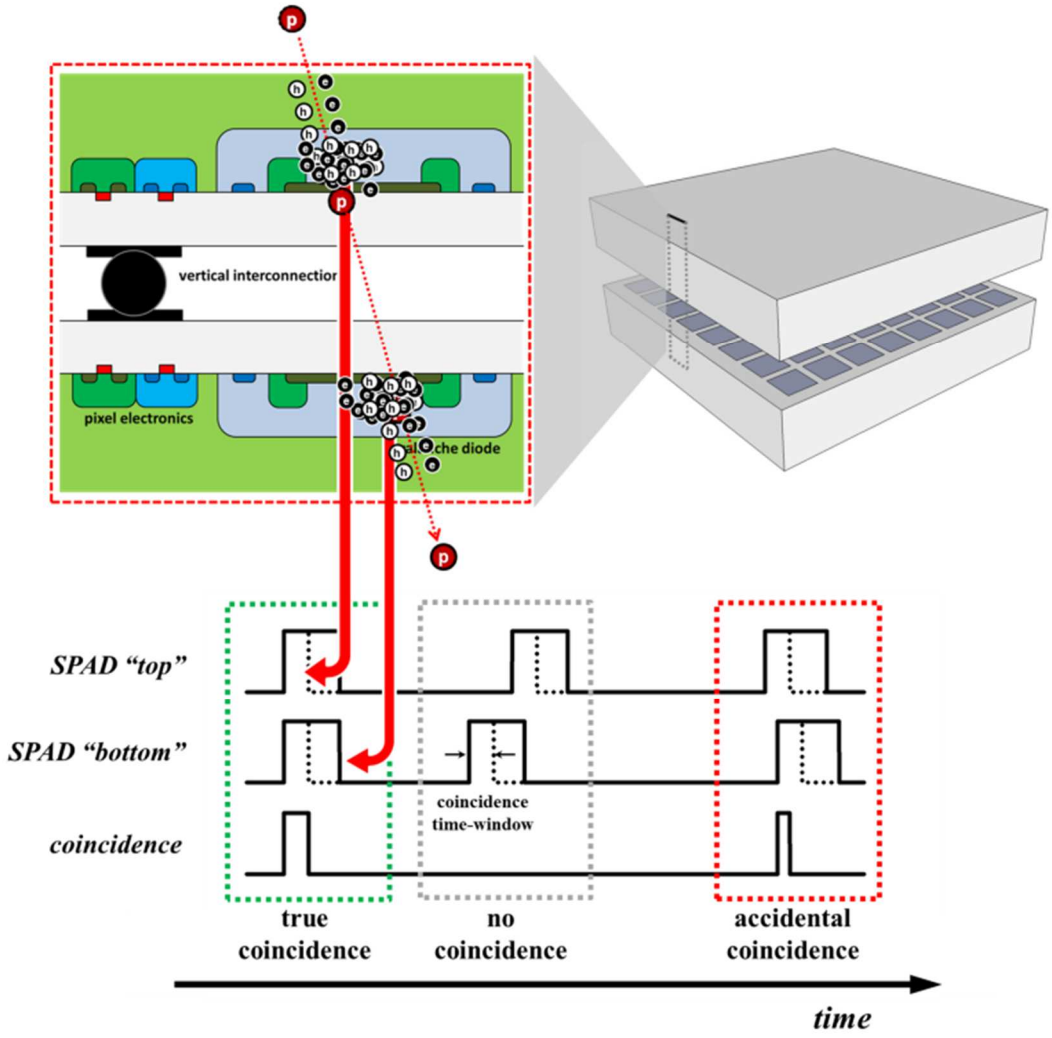

Figure 1: *Principe du détecteur 3D Silicon Coincidence Avalanche Detector (3D-SiCAD).*

Ce travail s'inscrit dans le projet Européen INFIERI – « INtelligent Fast Interconnected and Efficient Devices for Frontier Exploitation in Research and Industry », projet de type ITN Marie Curie n° [317446]. En particulier, cette thèse rentre dans le work-package n°2 dédié à l'exploitation des avancées récentes dans les technologies silicium submicroniques et les techniques d'intégration 3D pour l'instrumentation en physique des particules, astrophysique et applications médicales.

Après ce paragraphe d'introduction générale, ce long résumé en français présente les différentes étapes de conception et caractérisation du prototype en quatre paragraphes, suivis d'une conclusion générale et de perspectives (sixième paragraphe).

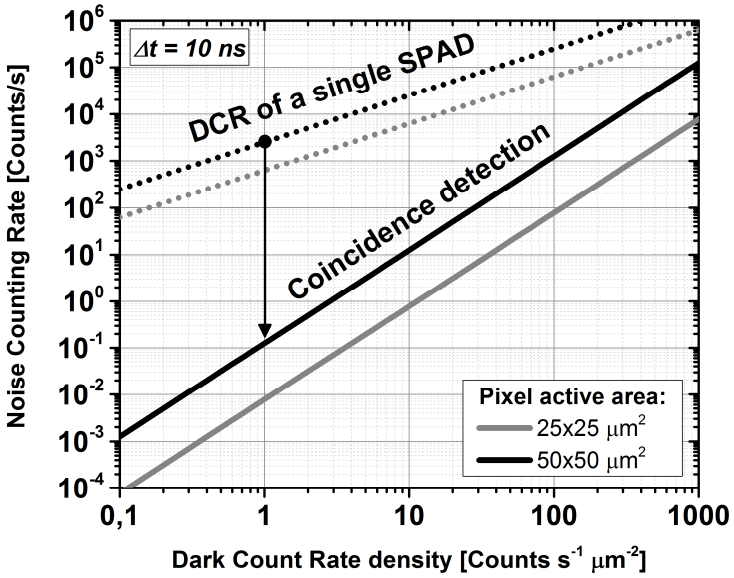

***Figure 2**: Bruit dans l'obscurité attendu pour le détecteur 3D-SiCAD pour une fenêtre temporelle de coïncidence de* 10 ns.

Le second paragraphe contient quelques rappels sur les diodes à avalanche en mode Geiger (appelés aussi SPADs pour Single Photon Avalanche Diodes), pour ensuite décrire le concept de détecteur 3D-SiCAD pour 3D Silicon Coincidence Avalanche Detector.
Le troisième paragraphe décrit la conception du démonstrateur : la diode SPAD, l'électronique associée (circuit d'étouffement – quenching et de réamorçage, circuit de coïncidence) et l'assemblage 3D.
Les résultats de caractérisation du pixel simple sont inclus dans le quatrième paragraphe ; en commençant tout d'abord par la caractérisation des diodes seules (courbes courant-tension, tension de claquage, électroluminescence), puis celle du pixel (validation du circuit de recharge, bruit dans l'obscurité / taux de comptage dans l'obscurité, déclenchements secondaires, efficacité de détection des photons).
Le cinquième paragraphe contient les résultats de caractérisation du prototype 3D-SiCAD avec les performances en bruit (en mode coïncidence sur deux pixels adjacents, puis sur le pixel 3D-SiCAD), et la démonstration de la capacité du prototype 3D-SiCAD à détecter les particules chargées avec la détermination de l'activité d'une source radioactive de strontium.

2. Détecteur à avalanche en coïncidence 3D

Dans ce paragraphe, nous introduisons, tout d'abord, quelques rappels sur les diodes à avalanche en mode Geiger (appelés aussi SPADs pour Single Photon Avalanche Diodes), pour ensuite présenter notre concept de détecteur à avalanche en coïncidence 3D (que nous avons dénommé 3D-SiCAD pour 3D Silicon Coincidence Avalanche Detector).

2.1 Quelques rappels sur les diodes à avalanche opérant en mode Geiger (SPAD)

2.1.1 Principe de fonctionnement

La diode à avalanche en mode Geiger (SPAD) est polarisée en inverse V_{inv} au-delà de sa tension de claquage V_{bd} telle que $V_{inv} = V_{bd} + V_{ex}$ (V_{ex} est appelée tension d'excès). La création d'une seule paire électron-tour (absorption d'un photon par effet photoélectrique, passage d'une particule chargée, génération thermique de porteurs etc.) suffit à déclencher l'avalanche de la diode car les porteurs générés dans la zone de multiplication vont se multiplier (en raison de l'ionisation par impact sous le très fort champ électrique) jusqu'à l'emballement du dispositif [7][8].

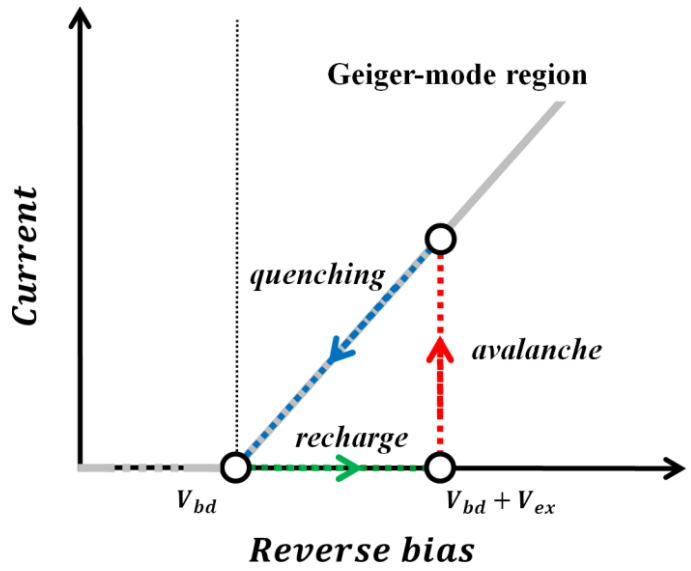

Figure 3: *Principe de fonctionnement de la diode à avalanche en mode Geiger (appelé aussi SPAD pour Single Photon Avalanche Diode).*

La diode SPAD est alors nécessairement associé à un circuit d'étouffement (on parlera de circuit de « quenching ») qui peut être simplement constitué d'une résistance en série permettant d'abaisser la tension en dessous du claquage et interrompre l'avalanche [7] (Figure 3). On désignera alors le pixel comme étant l'association de la cellule sensible (la diode SPAD), de son circuit de quenching et éventuellement d'une circuiterie permettant de retarder le réamorçage (gestion d'un temps mort t_h avant de réappliquer la tension inverse pour une nouvelle détection).

2.1.2 Les figures de mérite du SPAD

L'efficacité de détection

L'efficacité de détection (ou PDE pour Photon Detection Efficency, ou PDP pour Photon Detection Probability) est donnée par la relation théorique suivante [8]:

$$PDE(\lambda) = FF\big(1 - R(\lambda)\big)\overline{QE(\lambda)} \cdot \overline{P_{tr}}$$

avec FF : le facteur de remplissage ($FF = Surface_{active-region}/Surface_{pixel}$), $R(\lambda)$: le coefficient de réflexion qui est fonction de la longueur d'onde λ, $\overline{QE(\lambda)}$: l'efficacité quantique moyenne, $\overline{P_{tr}}$: la probabilité moyenne de déclenchement d'une avalanche.
L'efficacité quantique $QE(\lambda)$ dépend principalement du coefficient d'absorption des photons (*Figure 5*), mais aussi de la localisation de la photogénération de la paire électron-trou dans la structure et de sa collecte vers la zone de multiplication. La probabilité de déclenchement d'une avalanche $P_{tr}(x)$ peut se calculer à partir des coefficients d'ionisation pour les électrons et les trous qui peuvent être extraits d'une simulation numérique TCAD électrique du dispositif, la méthode est décrite dans [9]. La *Figure 4* illustre la probabilité moyenne de déclenchement d'une avalanche calculée pour une diode SPAD intégrée dans une technologie CMOS FDSOI en fonction de la tension inverse [10]. Cette figure contient aussi la caractéristique courant – tension en polarisation inverse. Nous notons que la probabilité moyenne de déclenchement d'une avalanche $P_{tr}(V_{inv})$ décolle lorsque la tension de claquage est dépassée, ce qui valide la méthode calcul de P_{tr}. La probabilité de déclenchement d'une avalanche $P_{tr}(x)$ augmente avec la tension inverse, elle est très faible autour de la tension de claquage est croit rapidement au-delà de la tension de claquage pour tendre vers 100 % pour un V_{ex} supérieur à $5\ V$

Bruit dans l'obscurité ou dark count noise (DCR)

Les SPAD sont affectés par des avalanches non souhaitées dans l'obscurité [7]. Plusieurs phénomènes sont à l'origine de ce taux d'événements (comptage dans l'obscurité ou DCR en Hz, on peut aussi donner le DCR en $Hz/\mu m^2$) :

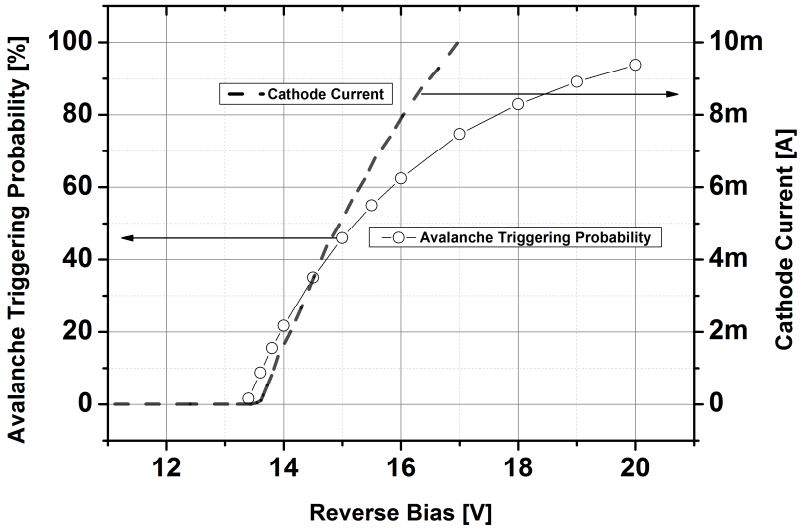

Figure 4: *Résultats de simulations, à gauche : la probabilité moyenne de déclencher une avalanche dans un SPAD conçu en technologie CMOS FDSOI 28nm, à droite : de la courbe courant-tension en inverse de cette diode de diamètre 7 µm [10].*

- Déclenchements initiaux ou primaires : ils sont reliés à la présence de paires électrons-trous dans la zone de multiplication (ou à proximité puis qui diffusent vers la zone de multiplication) avec pour origine par exemples : la génération thermique (processus Shockley-Read-Hall), le déchargement de pièges profonds localisés sous l'effet d'un fort champ électrique (trap-assisted tunneling), ou encore le passage d'électrons de la bande de valence à la bande de conduction (band-to-band tunneling) également sous l'effet d'un fort champ électrique.

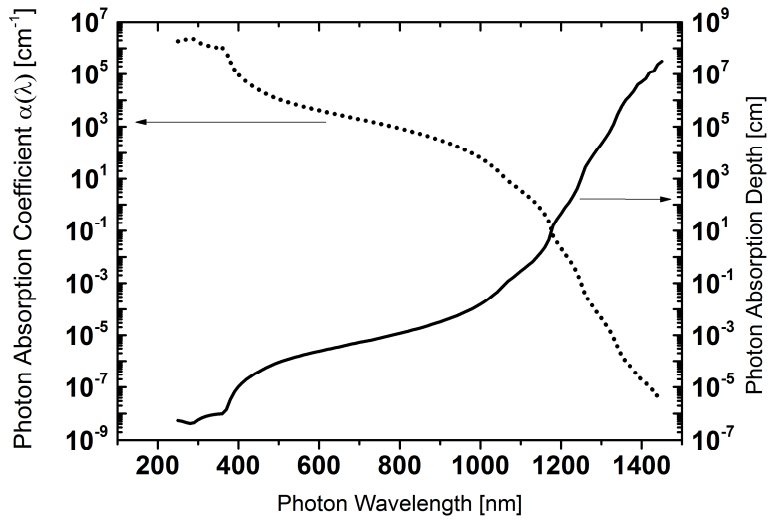

Figure 5: *Coefficient d'absorption et profondeur d'absorption des photons dans le silicium.*

- Déclenchements secondaires ou after-pulsing : ils sont corrélés à un déclenchement primaire. Lors de l'avalanche, quelques charges peuvent être capturées par des pièges puis relâchées ultérieurement déclenchant une nouvelle avalanche en cascade. Ce phénomène peut être minimisé en introduisant un temps mort t_h (avant réarmement du SPAD) suffisamment long pour que tous les pièges aient le temps de se décharger.

Le DCR observé peut s'exprimer avec la relation suivante :

$$DCR^{obs} = \frac{\lambda^*}{1+\lambda^* t_h} \quad \text{et} \quad \lambda^* = \frac{\lambda_0}{1-\mathrm{P_{ap}}(t)}$$

avec t_h : le temps mort, λ^* : le DCR intrinsèque, λ_0 : le DCR en l'absence d'after-pulsing, $\mathrm{P_{ap}}(t)$: la probabilité d'after-pulsing.

Couplage ou cross-talk

Les phénomènes de couplage ou cross-talk peuvent également déclencher des événements indésirables contribuant au DCR. Ils peuvent être de nature électrique (les porteurs diffusent d'un pixel activé vers un pixel voisin insuffisamment isolé), ou de nature optique (un SPAD en avalanche peut émettre des photons qui peuvent déclencher un SPAD voisin), voir illustration *Figure 6*.

Gigue temporelle ou jitter

Certaines applications comme le comptage de photons corrélés en temps requièrent une résolution temporelle du SPAD très élevée. Celle-ci est reliée à la gigue temporelle (jitter) qui correspond à la dispersion temporelle des temps d'observation d'avalanches, celle-ci est mesurable dans une configuration d'illumination répétitive [8] (par exemple sous éclairement d'un laser femtoseconde). En pratique, on mesure la largeur à mi-hauteur de l'histogramme des événements détectés (*Figure 7*). L'origine du jitter est multiple : dispersion des lieux de créations des photo-porteurs, dynamique de l'avalanche, architecture du SPAD (tailles des zones de multiplication, et de charge d'espace) etc.

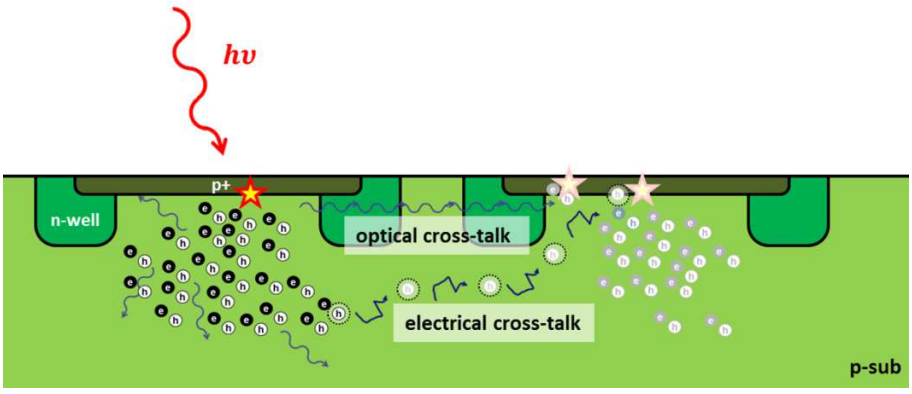

Figure 6: *Mécanismes de couplage otique et électrique dans une matrice de SPADS..*

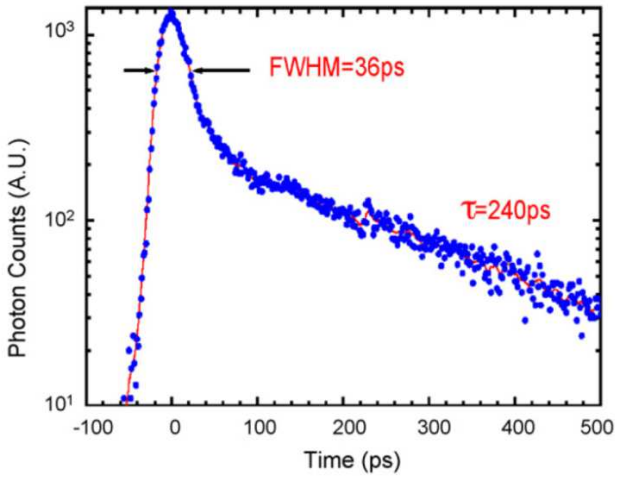

Figure 7: *Exemple de réponse temporelle d'un SPAD, extrait de* [8].

2.1.3 Quelques architectures de SPADs en technologie CMOS

Le premier SPAD en technologie CMOS a été conçu en 2003 par A. Rochas et al. [11], depuis de très nombreux travaux ont permis d'intégrer des SPADs dans des technologies avancées jusqu'au nœud 65 *nm* [12]. La *Figure 8* présente quelques exemples d'architectures de SPADs compatibles avec les technologies CMOS.

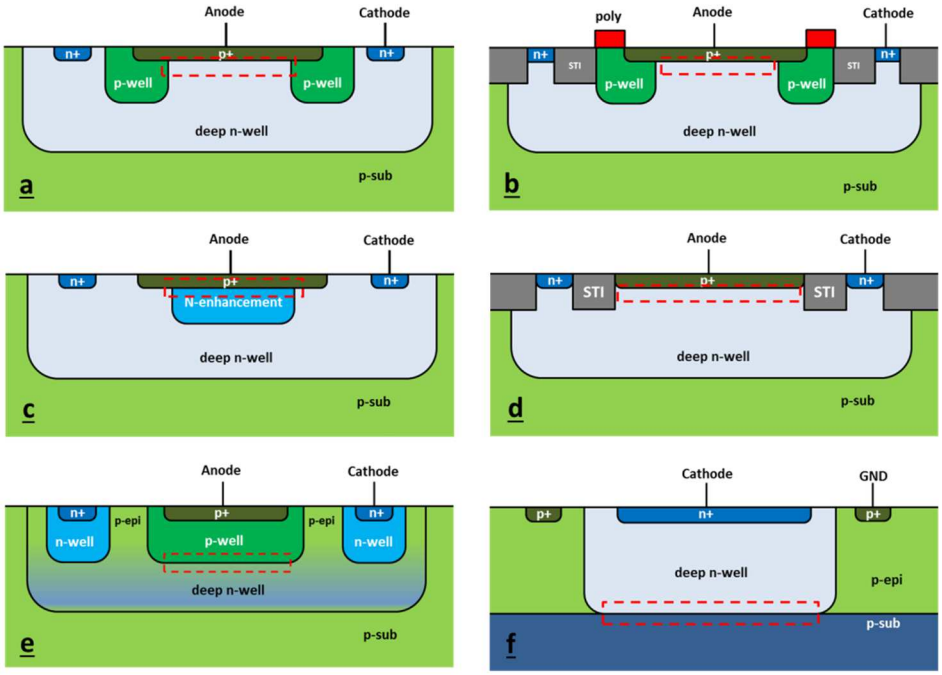

Figure 8: *Quelques exemples d'architectures de SPADs compatibles CMOS avec : a) anneaux de garde réalisés par diffusion, b) anneaux de garde et tranchées d'isolation, c) anneau de garde « virtuel », d) tranchées d'isolation, e) anneaux de garde par dopage rétrograde, f) région de multiplication enterrée.*

L'architecture du SPAD doit permettre d'obtenir une zone de multiplication bien homogène en évitant tout claquage prématuré sur les bords, ainsi l'utilisation de solutions telles que : anneaux de garde, tranchées d'isolation etc., est nécessaire.

2.2 Un nouveau détecteur : le 3D-SiCAD pour 3D Silicon Coincidence Avalanche Detector

2.2.1 Principe du 3D-SiCAD

Notre architecture de nouveau détecteur de particules chargées (*Figure 9*), appelé 3D-SiCAD pour 3D Silicon Coincidence Avalanche Detector, repose sur l'alignement vertical de pixels avec un mode de fonctionnement en coïncidence permettant d'obtenir les événements corrélés temporellement (associés au passage d'une seule particule chargée qui déclenche les deux niveaux). L'avantage attendu est de pouvoir discriminer les vrais événements, du bruit intrinsèque du simple pixel. A ce jour seulement, une réalisation non intégrée a été démontrée à base de photomultiplicateurs en silicium [3], et des travaux similaires sont menés à l'INFN en Italie [13].

2.2.2 Amélioration du bruit d'obscurité

Taux de faux événements de coïncidence

Le mode de détection en coïncidence dans une fenêtre temporelle donnée permet d'éliminer une grosse partie du bruit intrinsèque à chaque pixel, néanmoins il reste un taux de faux événements corrélés qui peut être estimé avec la relation suivante :

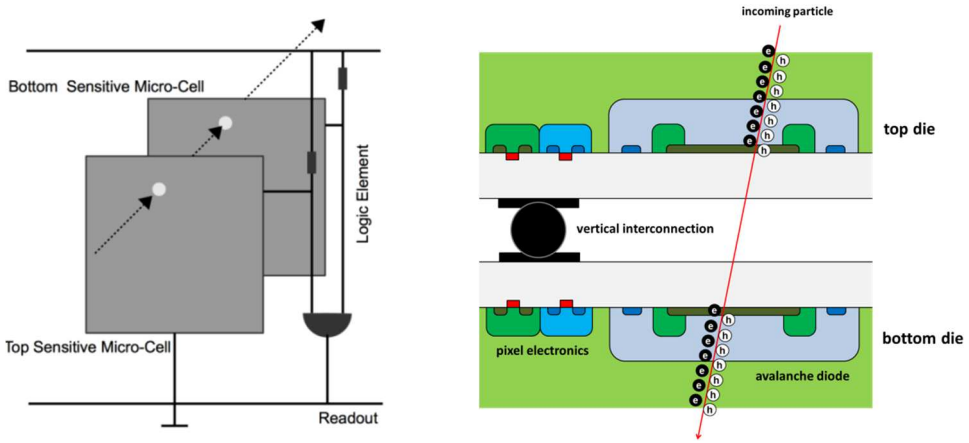

Figure 9: *Principe du détecteur à avalanche en coïncidence 3D (3D-SiCAD pour 3D Silicon Coincidence Avalanche Detector), illustration de gauche tirée de* [3].

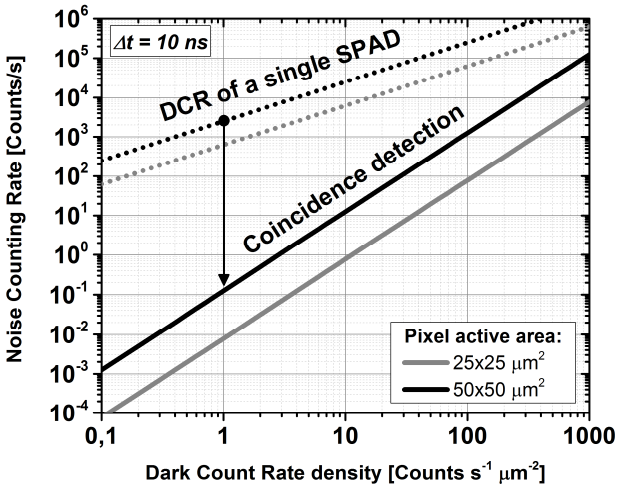

***Figure 10**: Taux de comptage corrélé de faux événements en fonction du bruit de chaque SPAD pour une fenêtre temporelle de $\Delta t = 10\,ns$ et deux surfaces: $25x25\mu m^2$ et $50x50\mu m^2$.*

$$FCR = 2 \cdot DCR_{top} \cdot DCR_{bottom} \cdot \Delta t$$

où DCR_{top} et DCR_{bottom} représentent DCR de chaque SPAD (dessus ou dessous) et Δt : la fenêtre temporelle d'observation.
La *Figure 10* présente l'amélioration attendue avec ce nouveau détecteur. Pour une technologie CMOS standard offrant un DCR intrisèque de l'ordre $1 - 10\,Hz$, le gain théorique attentu peut atteindre un facteur 10^4 pour une fenêtre temporelle d'observation $\Delta t = 10\,ns$.

Couplage - cross-talk optique

Dans une telle configuration 3D avec les surfaces actives dans une configuration face à face, le couplage optique peut être très présent (
*Figure **11***) et introduire une composante supplémentaire dans le taux de faux événements corrélés suivant l'équation :

$$NCR(\Delta t) = 2\,\Delta t\,\lambda_{top,0}\lambda_{bottom,0} + P_X\left(\lambda_{top,0} + \lambda_{bottom,0}\right)$$

où $\lambda_{top,0}$ et $\lambda_{bottom,0}$ représentent DCR de chaque SPAD sans cross-talk optique, Δt : la fenêtre temporelle d'observation, P_X : la probabilité de cross-talk optique.

En supposant que $\lambda_{top,0} = \lambda_{bottom,0} = \lambda_0$, la *Figure 12* représente le facteur de réjection du bruit (NCR/λ_0) prenant en compte les deux composantes : la corrélation du bruit intrinsèque de chaque SPAD dans une fenêtre temporelle Δt donnée et le cross-talk optique. Il apparait clairement le rôle néfaste du cross-talk optique même avec des probabilités faibles ($< 1\,\%$).

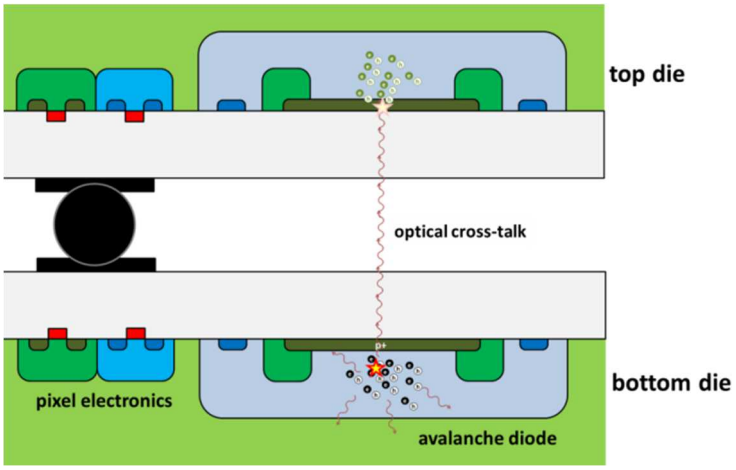

***Figure 11**: Présence du cross-talk optique dans le détecteur 3D.*

2.3 Conclusion partielle

Ce paragraphe a permis de rappeler les éléments essentiels du fonctionnent des diodes SPADs et de leurs caractéristiques. Ensuite, nous avons présenté le concept du nouveau détecteur de particules chargées, appelé 3D-SiCAD pour 3D Silicon Coincidence Avalanche Detector, reposant sur l'alignement vertical de pixels avec un mode de fonctionnement en coïncidence. La phase de conception est détaillée dans le paragraphe suivant.

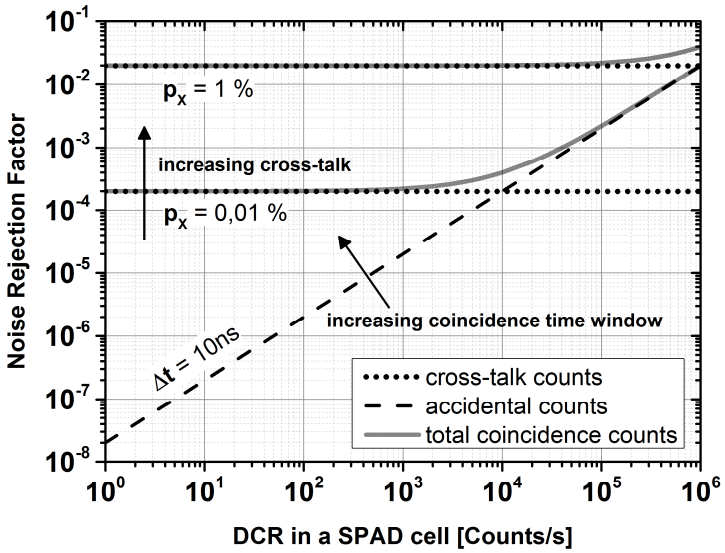

***Figure 12**: Taux de comptage corrélé de faux événements prenant en compte les deux composantes : corrélation du bruit intrinsèque de chaque SPAD dans une fenêtre temporelle donnée (ici* 10 ns*) et cross-talk optique (P_X = 0.1 % ou 1 %).*

3. Conception d'un prototype de 3D Silicon Coincidence Avalanche Detector

3.1 Conception de la surface active (SPAD)

Les choix de la technologie CMOS et de l'architecture du SPAD ont été guidés par l'objectif de démontrer la coïncidence dans un prototype 3D intégré à moindre coût. Ainsi, nous avons opté pour architecture de SPAD classique (minimum de risque) dans une technologie CMOS bas coût accessible via les runs multi-projets. La coupe schématique du SPAD en technologie CMOS haute tension 0,35 µ*m* de la société AustriaMicroSystem (AMS) est donnée sur la *Figure 13* (ainsi qu'une photo), l'architecture est basée sur une jonction p+ / nwell profond (ou deep n-tub) avec des anneaux de garde réalisés par des pwell (p-tub). Le pixel a une forme hexagonale de 50 µ*m* de côté.

3.2 Conception de l'électronique associée

3.2.1 Circuit d'étouffement et de réamorçage

Deux solutions sont envisageables pour disposer la diode (*Figure 14*) ; soit la « diode en haut » (le nœud de sortie est alors l'anode), soit la « diode en bas » (le nœud de sortie est alors la cathode). Nous avons choisi la configuration « diode en haut » (*Figure 14*a) car la capacité parasite entre cathode (nwell profond) et le substrat (de type p) n'a pas de rôle (la tension reste constante à ses bornes), de plus la sortie peut être lue directement sans découplage sur l'anode (la diffusion p+ dans notre cas).

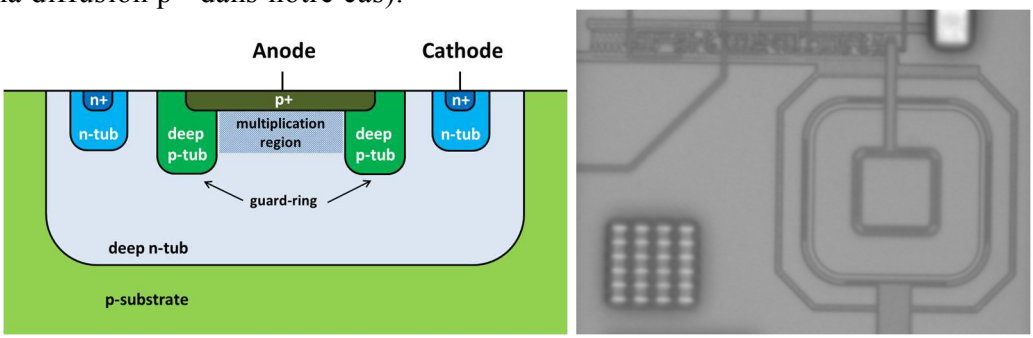

Figure 13: *à gauche : coupe schématique du SPAD réalisé en technologie CMOS haute tension* 0.35 µ*m de AMS, à droite) photo du SPAD et de l'électronique associée.*

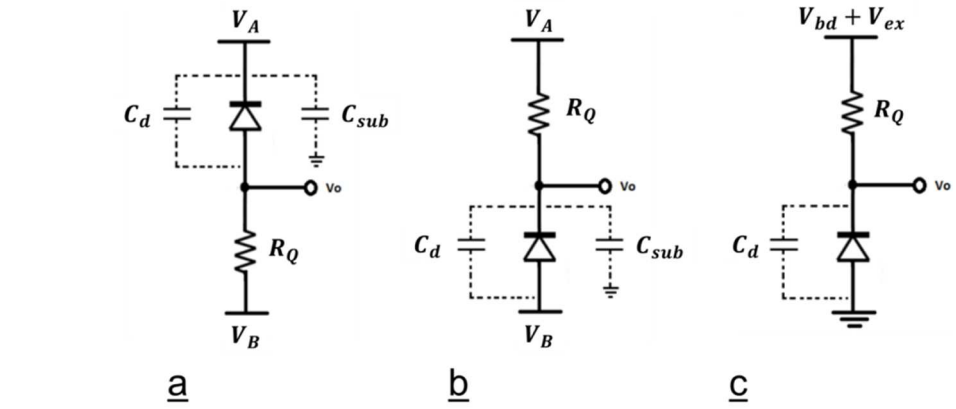

***Figure 14**: Différentes solutions pour réaliser le pixel.*

La *Figure 15* présente le circuit de gestion d'étouffement de l'avalanche (quenching) et de réamorçage (gestion du temps mort t_h avant réarmement), [14]. Le transistor Q1 (OFF au départ) se comporte comme une très forte résistance à travers laquelle nous allons sonder le déclenchement de la diode sur le nœud A (quenching passif). Avant toute avalanche, la tension inverse aux bornes de la diode est : $V_{bd} + V_{ex}$.

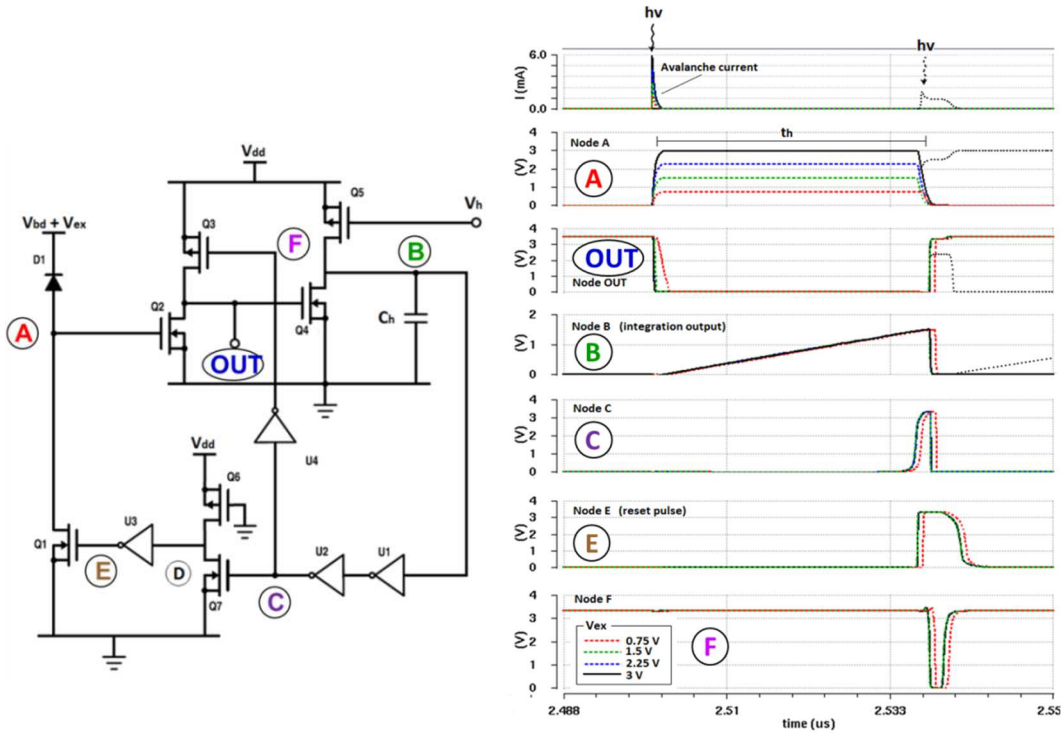

***Figure 15**: à gauche) schéma du circuit d'étouffement et réamorçage, à droit) simulations montrant les signaux lors d'une avalanche avec un temps mort $t_h = 35ns$.*

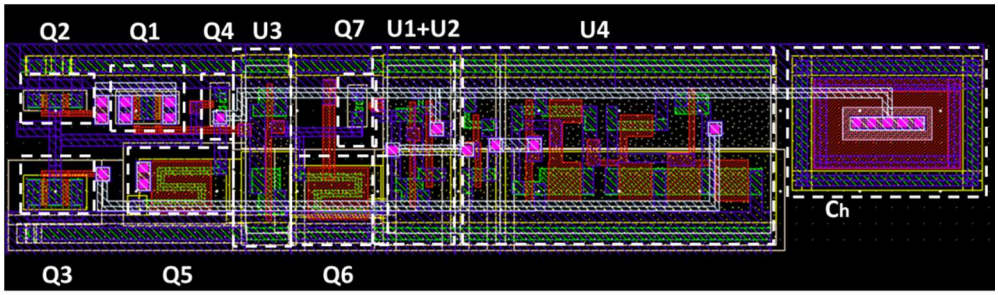

***Figure 16**: Layout de l'électronique intégrée dans le pixel.*

Dès qu'une avalanche se produit, le potentiel du nœud A monte à $V_A = V_{ex}$ (la tension inverse aux bornes de la diode redescend à : V_{bd}). Le transistor Q2 passe ON (le transistor Q3 est initialement OFF). Le signal OUT présente alors un front descendant caractéristique de la détection d'une avalanche. Le transistor Q4 passe alors OFF, la phase d'intégration commence avec le chargement de la capacité C_h via le transistor Q5 contrôlé par V_h (qui permet de contrôler le temps mort de l'extérieur avec la tension V_h). Dès que le nœud B atteint le seuil de la chaine d'inverseurs U1-U2, la phase de réarmement démarre en commutant Q1 ON (via Q7 et U3), en même temps la phase d'intégration est stoppée via U4 et Q3. Il faut s'assurer que Q1 reste suffisamment longtemps à l'état ON pour que le nœud A redescende bien à $0\,V$. Le layout de cette électronique est donnée sur la *Figure 16*.

3.2.2 Circuit de coïncidence

La coïncidence entre les événements issus des deux pixels alignés peut être détectée soit avec un circuit de coïncidence intégré directement dans le silicium, soit par un traitement extérieur via une carte FPGA. La
Figure 17 présente le principe de cette détection de coïncidence avec des pulses numériques qui sont réduits à la largeur de la fenêtre temporelle de coïncidence que l'on additionne (ET logique) dans un second temps.

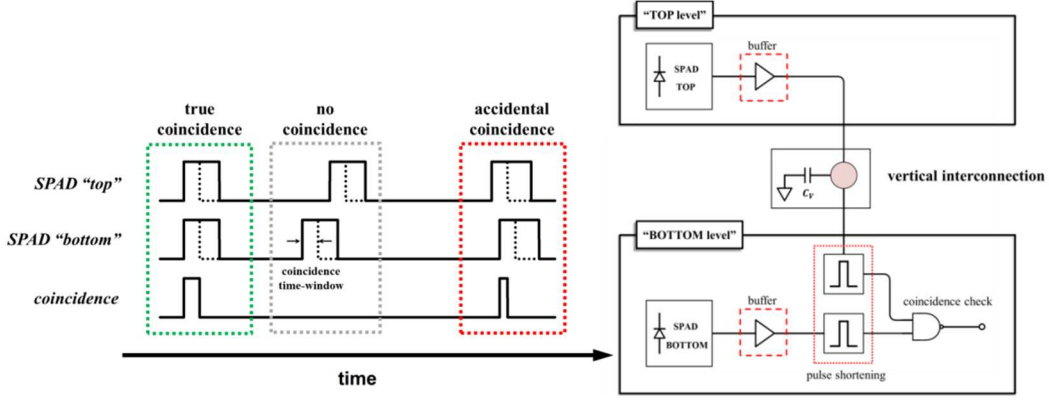

***Figure 17**:Principe du circuit de détection de la coïncidence intégré sur silicium.*

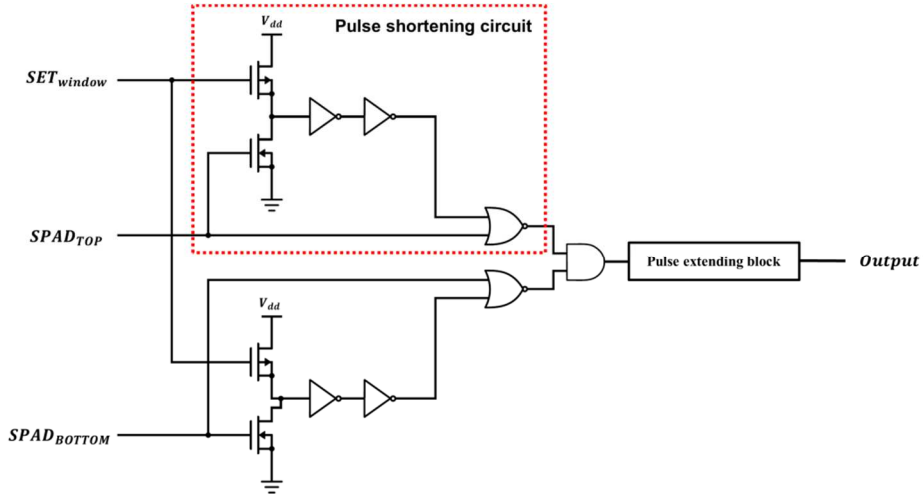

***Figure 18**: Vue détaillée du circuit de mise en forme des impulsions pour la détection de coïncidence.*

Le circuit de mise en forme des impulsions pour la détection de coïncidence est détaillé sur la *Figure 18*.

3.3 Assemblage 3D

Les technologies d'intégration 3D ont été très développées ces dernières années [15] avec l'apparition de briques importantes très sophistiquées telles que les vias traversant (TSV pour Through Silicon Via), le collage de wafer sur wafer (collage cuivre-cuivre ou collage hybride), les interconnexions verticales par micro-billes, micro-piliers etc. Dans notre cas, nous ne pouvons pas envisager de travailler au niveau du wafer pour une question de coût, une solution idéale serait de réaliser un prototype tel que décrit sur la
*Figure **19*** avec des micro-bumps entre les 2 niveaux et des TSV et des bumps pour interconnecter avec le monde extérieur (boitier, carte). Nous avons opté pour une solution plus simple et moins onéreuse avec une interconnexion par billes (de diamètre d'environ 70 μm) et des reprises par wire-bonding sur la puce inférieure (*Figure 20*).

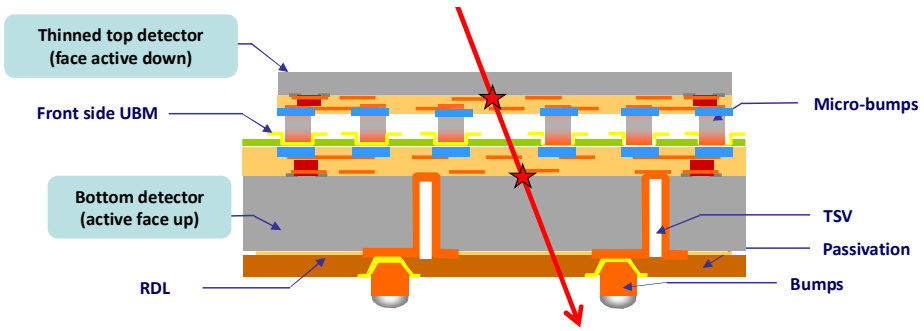

***Figure 19**: Possible intégration 3D de notre détecteur (proposition du LETI)*

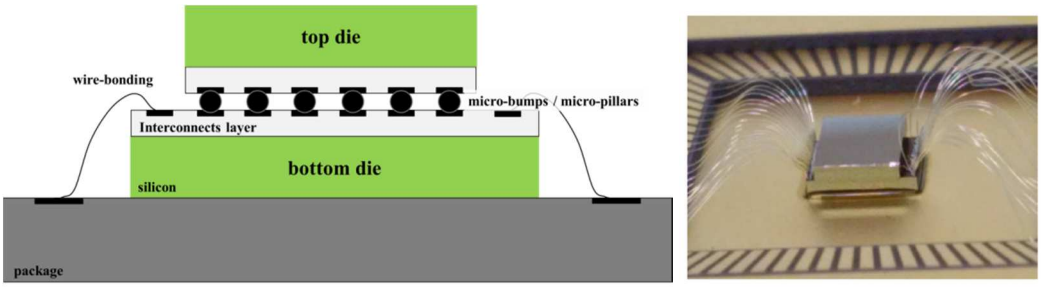

***Figure 20**: Schéma d'intégration 3D adopté pour notre prototype à gauche, photo du prototype à droite.*

De plus, pour minimiser les coûts nous avons décidé d'utiliser un layout identique pour les puces inférieure et supérieure avec l'astuce d'assemblage présenté sur la *Figure 21*. Afin d'utiliser la même puce dessous et dessus (après rotation et retournement), son dessin (*Figure 22*) présente un axe de symétrie qui permet la superposition des cellules verticalement dans le prototype final. Les connexions par wire-bondings ne se font que par les deux côtés opposés du circuit inférieur ; seulement la moitié des cellules est accessible au final mais l'avantage est de ne payer qu'un seul run silicium en technologie AMS CMOS haute tension 0,35 μm. Le centre du test-chip contient des diodes et circuits de quenching-recharge isolés. La taille de la puce est de $2{,}3 \times 3\ mm^2$. La *Figure 23* contient les photos de quelques éléments présents sur notre test-chip: deux pixels adjacents pour tester la coïncidence dans le plan, un pixel unique, une matrice 2×2, une diode seule avec ses connexions directes.

3.4 Conclusion partielle

Dans ce paragraphe, nous avons présenté les différentes phases de conception de notre prototype : l'architecture de la cellule active (diode SPAD), l'électronique associée pour l'étouffement et la recharge, l'électronique pour détecter la coïncidence puis la technique d'assemblage 3D adoptée. Le prochain paragraphe contient les résultats de mesure sur le pixel simple.

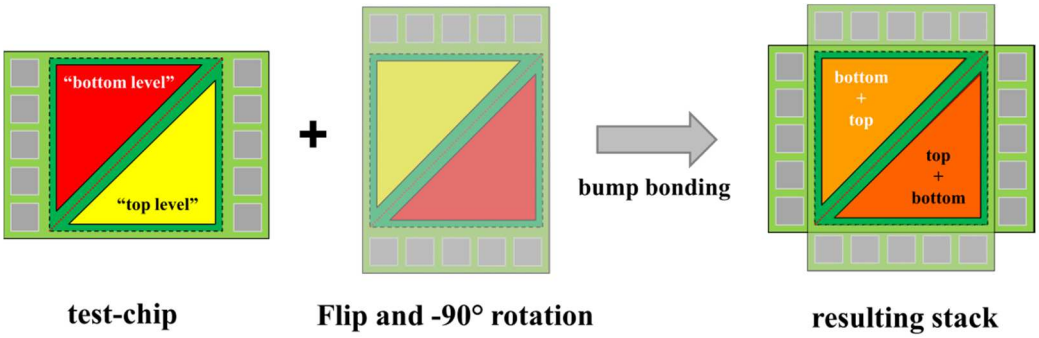

***Figure 21**: Assemblage de deux puces identiques pour réaliser le prototype 3D.*

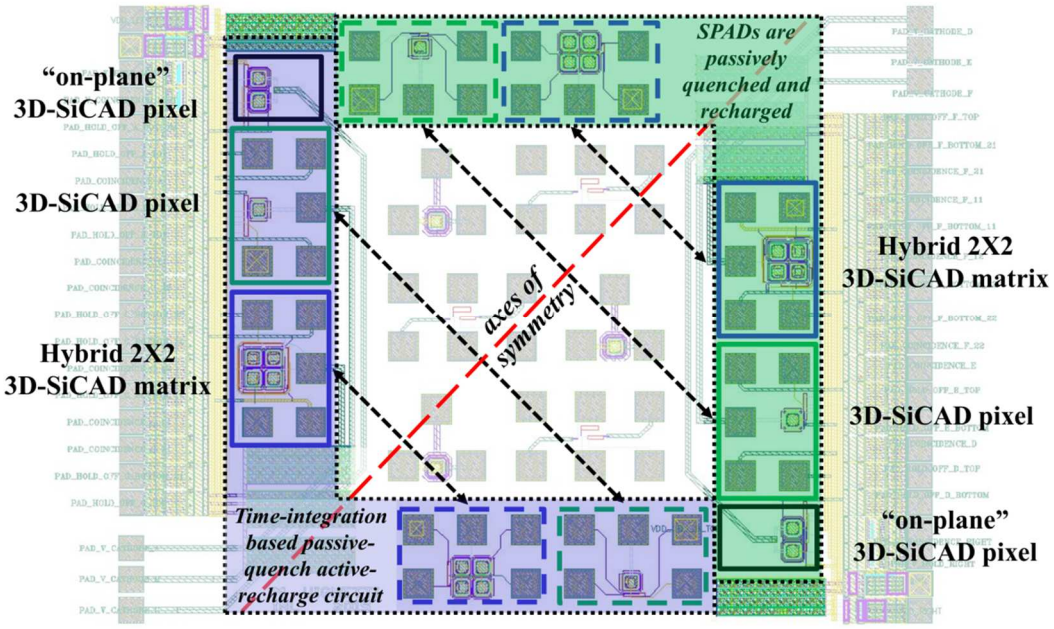

***Figure 22**: Layout de notre test-chip.*

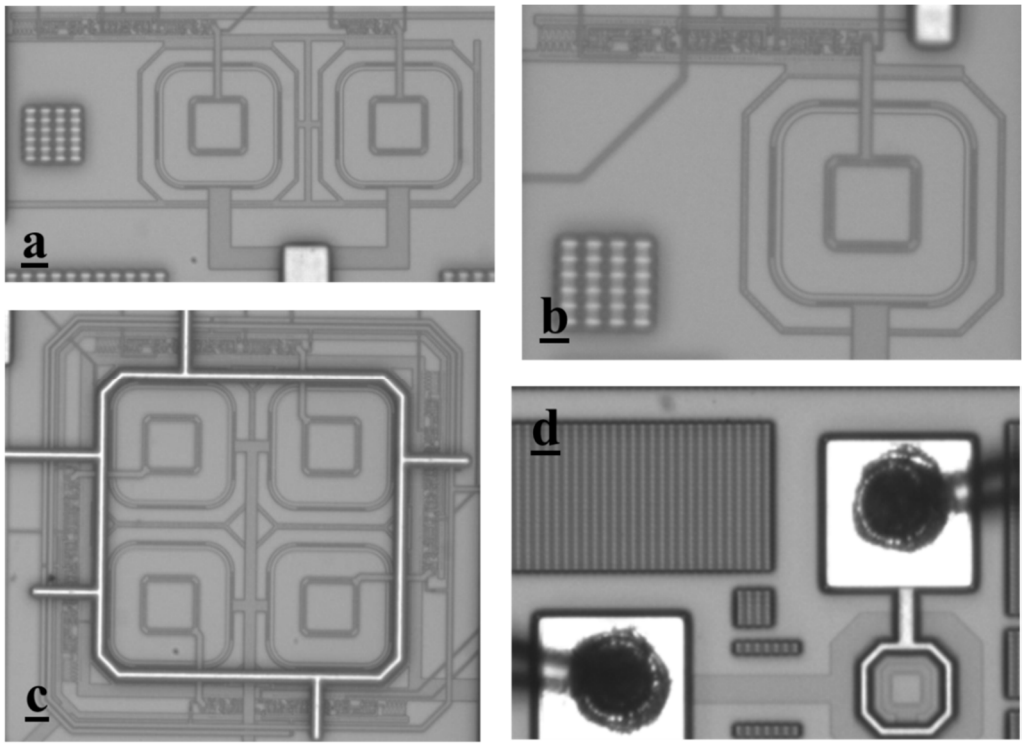

***Figure 23**: Photos de quelques éléments présents sur notre test-chip: a) 2 pixels adjacents pour tester la coïncidence dans le plan, b) pixel unique, c) matrice 2x2 d) diode seule avec ses connexions directes.*

4 Caractérisations du simple pixel

4.1 Caractérisation des diodes

4.1.1 Caractérisation statique en direct

Les premières mesures ont été réalisées sur les diodes simples isolées (au centre du test-chip, cf. *Figure 22*) avec tout d'abord la caractérisation statique en polarisation directe. Ainsi les mesure des courbes $I-V$ sur une large plage de températures de $233\,K$ à $313\,K$ $(-40°C$ à $+40°C)$ ont permis de vérifier le bon fonctionnement de ces diodes dans le mode direct (*Figure 24*).

4.1.2 Caractérisation statique en inverse : tension de claquage et courant d'obscurité

Ensuite, nous avons mesuré la caractéristique statique inverse des diodes dans l'obscurité et sous faible éclairement (lumière parasite) afin d'extraire la tension de claquage (*Figure 25*). Notre tension de claquage extraite est de $V_{bd} = 18{,}77 \pm 0.13\,V$. Nous avons reproduit ces mesures sur une large plage de températures ($233\,K$ à $313\,K$) comme illustré sur la *Figure 26* pour déterminer la variation de tension de claquage en fonction de la température. Le coefficient directeur extrait est de : $dV_{bd}/dT = 20.4\,mV/°C$, celui-ci est conforme à la littérature et indique une variation inférieure à quelques % de la tension de claquage autour de la température ambiante. Par la suite, nous avons représenté sur la *Figure 27* (dans un plot de type Arrhenius) le courant d'obscurité mesuré pour différentes tensions inverses ($13\,V, 15\,V, 17\,V$) en fonction de $1/T$.

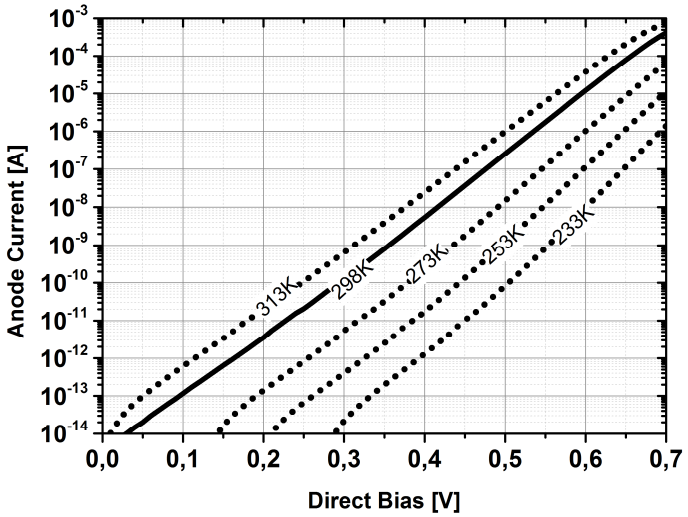

***Figure 24**: Caractéristiques $I-V$ en direct (entre $233\,K$ et $313\,K$).*

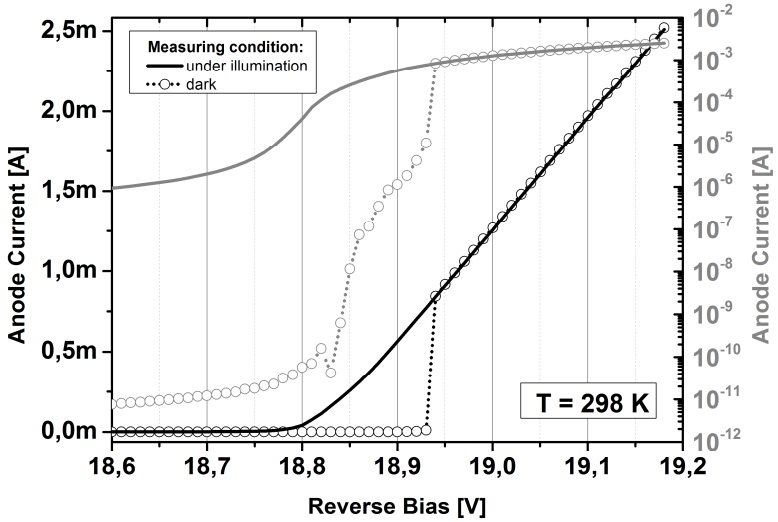

***Figure 25**: Caractéristiques $I-V$ en inverse (dans l'obscurité et sous faible lumière parasite), échelles linéaire à gauche et logarithmique à droite.*

A partir de la pente (ou des pentes) de ce tracé, il est possible d'extraire une (des) énergie(s) d'activation qui sont des signatures du(des) mécanisme(s) de créations de paires électron-trous responsables du courant d'obscurité en inverse [16]. Pour des températures supérieures à $+10°C$, les énergies d'activation déterminées sont de $E_A^{13V} = 0.383\ eV$, $E_A^{15V} = 0.336\ eV$ et $E_A^{17V} = 0.337\ eV$; c'est à dire assez proches de $E_g/2$ indiquant un processus plutôt de type génération – recombinaison de la théorie Shockley-Read-Hall. En revanche, il semble que pour une tension inverse proche de la tension de claquage et à plus basse température (inférieure à $10°C$), on puisse extraire une énergie d'activation plus faible $E_A^{17\,V} = 0.212\ eV$ qui serait la signature d'un mécanisme de génération de porteurs assistée par le champ électrique. Des mesures complémentaires seraient nécessaires pour confirmer cette hypothèse.

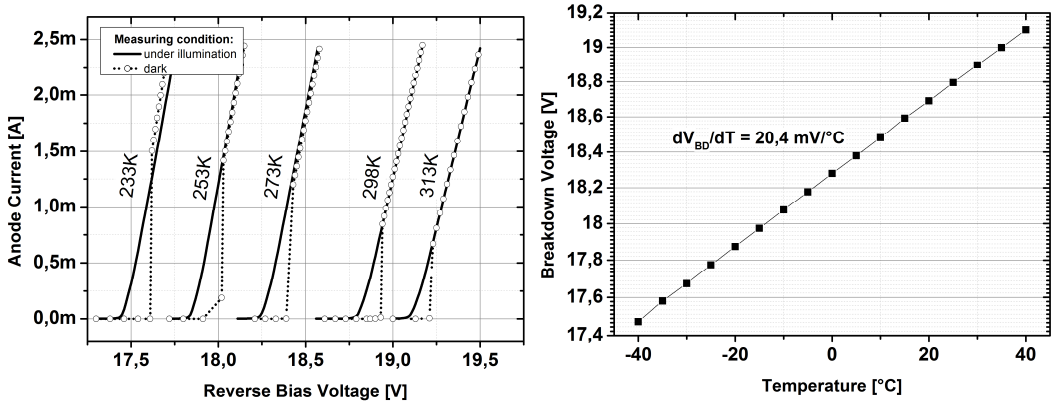

***Figure 26**: A gauche : Caractéristiques $I-V$ en inverse pour (entre 233K et 313K). A droite : variation de la tension de claquage extraite en fonction de la température.*

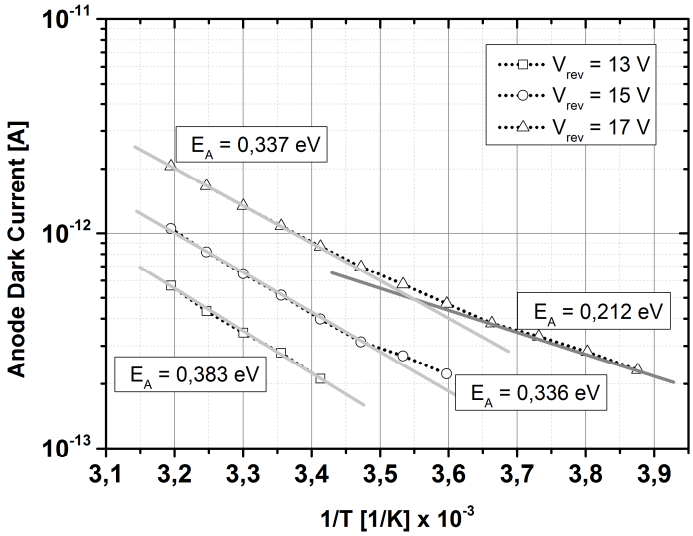

***Figure 27**: Courant d'obscurité mesuré pour différentes tensions inverses (*$13\ V, 15\ V, 17\ V$*) en fonction de* $1/T$.

4.1.3 Electroluminescence

Une image de la lumière émise par électroluminescence lorsque la diode est polarisée en inverse au-dessus de sa tension de claquage est donnée sur la *Figure 28*, avec $V_{rev} = 20V$, soit. $V_{ex} = V_{rev} - V_{bd} = 1.2\ V$. Ce type d'image permet de vérifier si la diode est bien conçue ; à savoir si l'électroluminescence est bien homogène (dans la zone de multiplication) sans faire apparaitre de « points chauds » qui pourraient être le signe d'un champ électrique excessif par exemple sur un coin (c'est-à-dire un claquage favorisé sur un bord ou un angle).

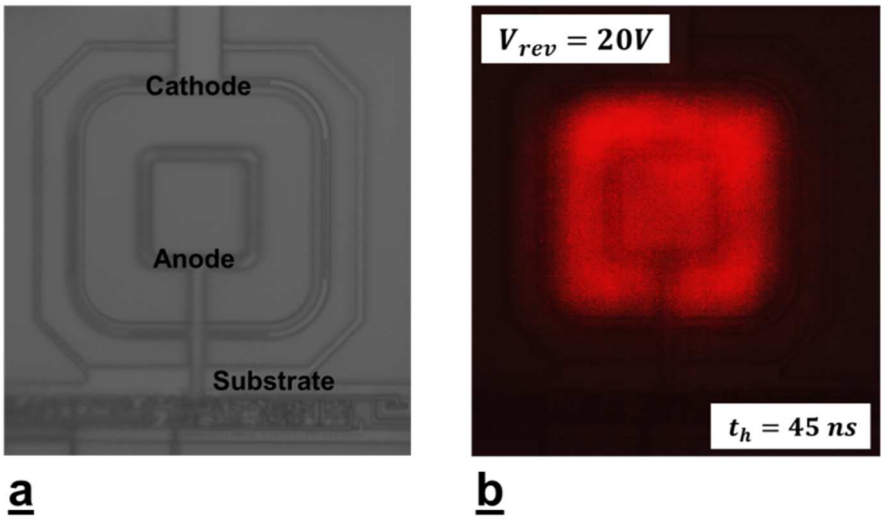

***Figure 28**: A gauche: photo de la diode à avalanche, et à droite: image d'électroluminescence lorsque la diode est polarisée en inverse à* $20\ V$.

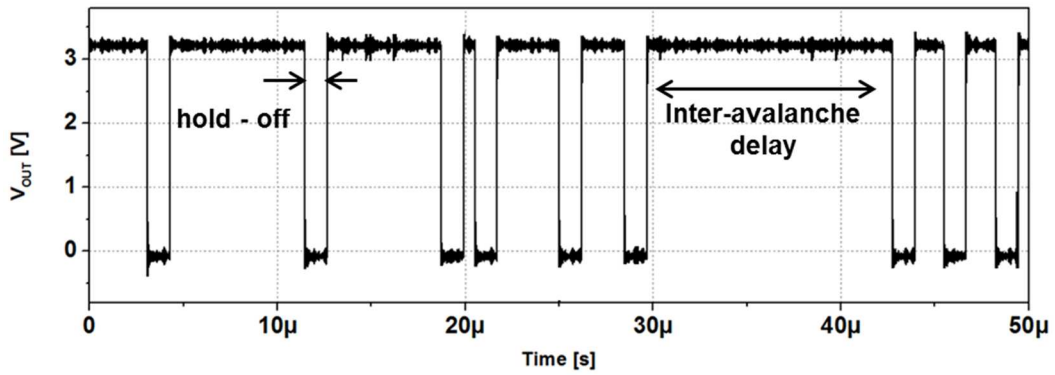

Figure 29: *Exemple de relevé temporel de la sortie du SPAD.*

La luminescence provient des porteurs chauds présents en raison du très fort champ électrique présent dans la jonction, son spectre est assez large dans le visible [17][18]. L'électroluminescence reste un sujet encore étudié de nos jours [17]. L'image d'électroluminescence de la *Figure 28* est satisfaisante ; laissant conclure que la diode a été correctement dessinée (le contour central sombre sur cette image est vraisemblablement associé au métal qui contacte l'anode et qui introduit un effet d'ombrage).

4.2 Caractérisation des pixels

La *Figure 29* illustre un relevé réalisé à l'oscilloscope (LeCroy HDO6104-MS) qui permet d'observer, de mesurer et d'établir des statistiques sur : la durée du temps mort, et le temps inter-avalanches. Chaque événement d'avalanche se caractérise par un front descendant, comme expliqué dans le paragraphe « Circuit d'étouffement et de réamorçage ». Sur ce chronogramme, un événement d'avalanche secondaire (after-pulsing) est observable autour de 20 µs.

4.2.1 Caractérisation du circuit de recharge

Des mesures du temps mort ont été réalisées sur 28 pixels différents et les résultats comparés aux simulations Monte Carlo prenant en compte les variations du procédé technologique, le mismatch (appariement des composants) et les parasites extraits du dessin physique (layout). La *Figure 30* indique un temps mort d'environ $t_h = 45\,ns$ au minimum jusqu'à $t_h = 6{,}5\,\mu s$ au maximum pour une tension de contrôle de la recharge (hold-off, cf. *Figure **15***) $V_h = 1{,}6V$ à $V_h = 2{,}4V$. Sur cette dernière figure, l'écart type $\pm\sigma$ est également représenté sur les résultats de simulations. Nous observons que le temps mort mesuré rentre dans le domaine de variabilité à $\pm\sigma$, et que très probablement toutes les puces mesurées proviennent du même wafer de silicium.

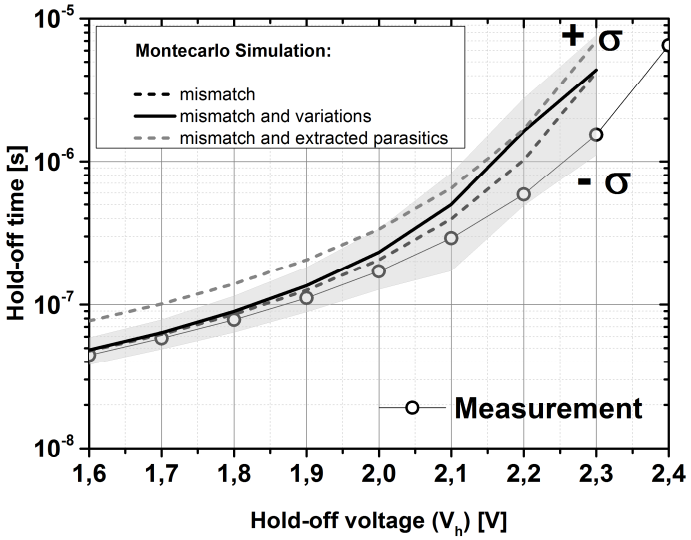

***Figure 30**: Caractérisation du circuit de recharge (mesure du temps mort t_h en fonction de la tension de contrôle extérieure notée V_h).*

4.2.2 Taux de comptage dans l'obscurité des SPADs

Le comportement dynamique des SPADS dans l'obscurité a été mesuré à $25°C$ afin de déterminer le taux de comptage dans le noir, soit directement à l'oscilloscope LeCroy (HDO6104-MS) en mesurant le nœud OUT (cf. *Figure **15***), soit via la carte mezzanine FPGA couplée à notre carte test permettant un comptage déporté des événements.
En reprenant le chronogramme de la *Figure 29*, en l'absence d'after-pulsing (par exemple en imposant un temps mort très grand), la distribution statistique du temps inter-avalanche doit suivre une loi de Poisson.
Le taux de comptage apparent dans l'obscurité (Dark Count Rate : DCR) a été extrait en mesurant le temps moyen entre deux avalanches consécutives $< t_{av} >$ (t_{av} étant le temps inter-avalanche). L'inverse du temps moyen entre deux avalanches consécutives $< t_{av} >$ donne directement le DCR sans avoir besoin de faire de correction liée au temps mort.
Ce DCR a été mesuré en fonction du temps mort t_h, pour plusieurs tensions d'excès V_{ex} (*Figure 31*). Nous rappelons que la tension de claquage est $V_{bd} = 18{,}8\,V$. Comme décrit dans la section « Les figures de mérite du SPAD », nous observons une augmentation du DCR avec la tension d'excès V_{ex} en raison d'une augmentation de la probabilité de déclenchement avec la tension inverse mais aussi en raison de l'augmentation de la génération de porteurs par les mécanismes assistés par le champ électrique élevé (voir paragraphe « Caractérisation statique en inverse : tension de claquage et courant d'obscurité »). L'augmentation des événements secondaires (after-pulsing) apparait clairement pour les faibles temps morts ($t_h < 100\,ns$), et est d'autant plus présente que la tension d'excès V_{ex} est forte (augmentation de la charge piégée avec la tension inverse).

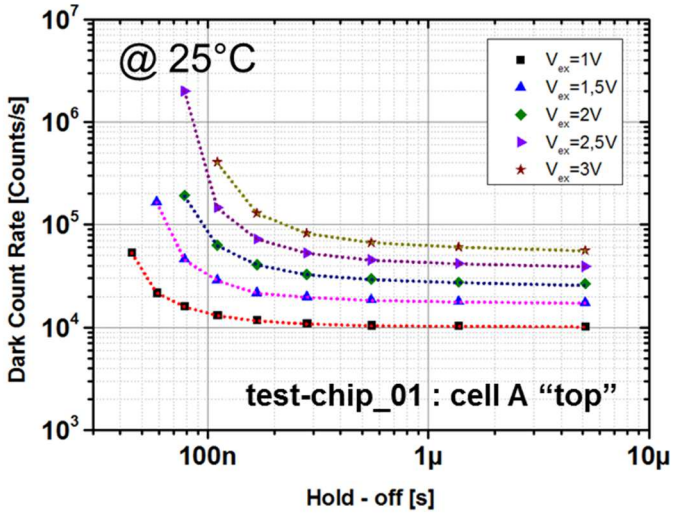

***Figure 31**: Taux de comptage (DCR) en fonction du temps mort t_h.*

Sur la *Figure 32*, le *DCR* mesuré a été tracé en fonction de la tension inverse pour deux diodes SPAD. Nous observons que le *DCR* croit quasi linéairement avec la tension inverse. En extrapolant le *DCR* aux faibles tensions d'excès, nous retrouvons bien la tension de claquage de nos diodes SPAD autour de $18{,}8\,V$. Des mesures intensives de *DCR* en fonction de la tension d'excès ont été faites sur 35 diodes SPAD issues de 5 puces différentes (*Figure 33*). Ces mesures, à température ambiante, ont été réalisées avec un temps mort de $5\,\mu s$ pour garantir un très faible after-pulsing. Nous pouvons conclure que, pour une tension d'excès de $1\,V$, nous obtenons une valeur médiane $70\,Hz/\mu m^2$ (c'est-à-dire que la moitié des diodes SPAD présentent un *DCR* plus faible que cette valeur). Ces résultats sont conformes aux résultats présentés par Vilella et al. pour une architecture similaire de SPAD [19].

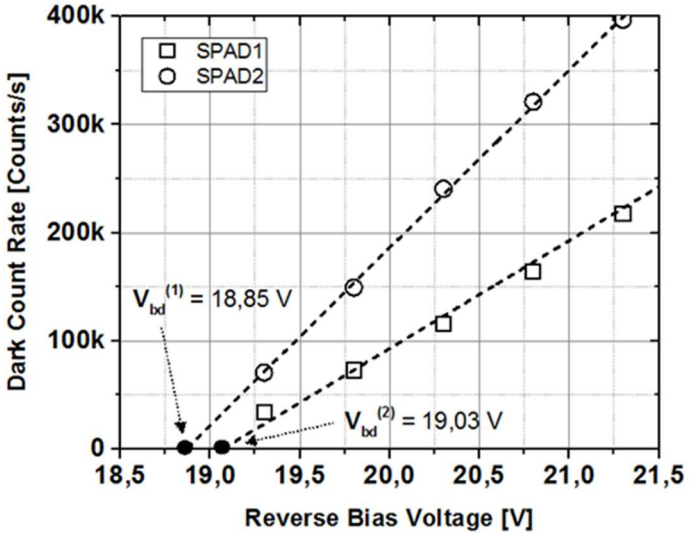

***Figure 32**: Extraction de la tension de claquage à partir de mesures de taux de comptage dans l'obscurité.*

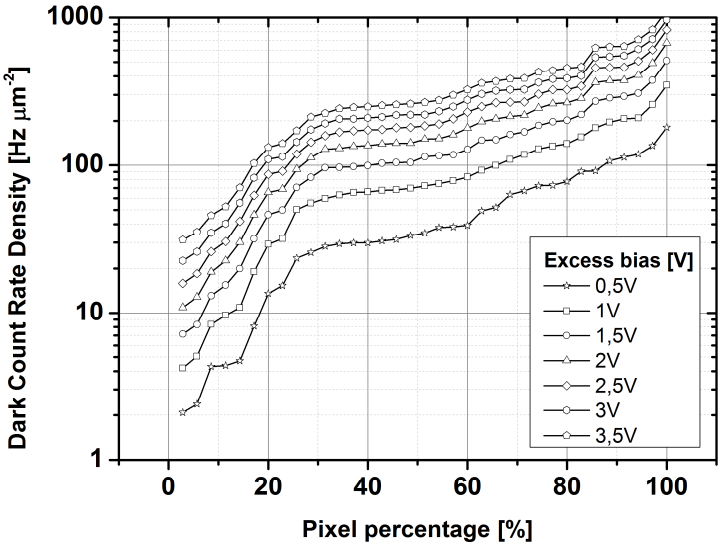

Figure 33 *: Distribution cumulée du DCR sur 35 SPADS avec un temps mort* $t_h \approx 5$ µ*s (after-pulsing négligeable).*

4.2.3 Déclenchements secondaires (after-pulsing)

La probabilité de déclenchements secondaires (after-pulsing) P_{ap} pour différentes tensions d'excès (V_{ex}) avec $t_h = 150$ ns a été représentée sur la *Figure 34*. Ces valeurs ont été extraites par une méthode « dite rapide » avec les relations suivantes :

$$\lambda^* \approx \frac{\lambda_0}{1-P_{ap}}$$

où λ^* représente le DCR observé avec $t_h = 150$ ns, et λ_0 le DCR du SPAD sans after-pulsing (mesuré avec un t_h élevé).

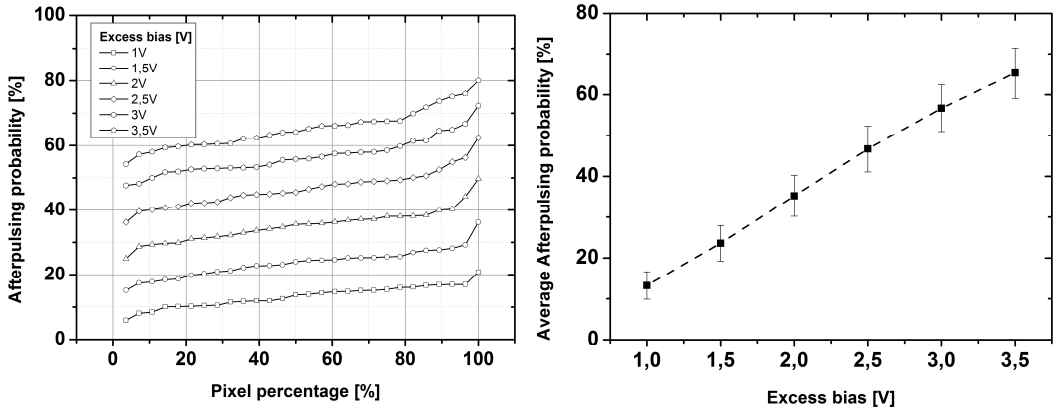

Figure 34*: Probabilités de déclenchement secondaires (after-pulsing) pour différentes tensions d'excès (*V_{ex}*) avec* $t_h = 150$ *ns.*

Nous obtenons alors la probabilité de déclenchements secondaires (after-pulsing) P_{ap} avec al relation :

$$P_{ap}(t_h = 150\,ns) \approx 1 - \frac{\lambda_0}{\lambda^*} = 1 - \frac{DCR(long\ t_h)}{DCR(t_h{=}150\,ns)}$$

La probabilité d'after-pulsing peut-être extraite de manière plus précise en exploitant la distribution des temps inter-avalanches t_{av} (voir chronogramme de la *Figure 29*). La densité de probabilité d'after-pulsing est donnée par :

$$p_{av}(t) = \big(\lambda_0 + f(t)\big)e^{-\int_0^t \left(\lambda_0 + f(t')\right)dt'}$$

Cette relation décrit un processus de Poisson avec un paramètre caractéristique $\lambda_0 + f(t)$: λ_0 est associé au taux de comptage en absence d'after-pulsing (pulses primaires de probabilité : $P_d(t) = 1 - e^{-\lambda_0 t}$), tandis que la fonction $f(t)$ est relative aux phénomènes d'after-pulsing (déclenchements secondaires).

La probabilité d'avalanche $P_{av}(t)$ est décrite par la relation : $P_{av}(t) = P_{ap}(t) + P_d(t) - P_{ap}(t)P_d(t)$, c'est-à-dire la somme des probabilités de pulses primaires $P_d(t)$ ou secondaires $P_{ap}(t)$, moins les probabilités combinées. Ainsi la probabilité d'after-pulsing est donnée par :

$$P_{ap}(t) = \frac{P_{av}(t) - P_d(t)}{1 - P_d(t)}$$

La *Figure 35* représente un histogramme expérimental normalisé des temps inter-avalanches t_{av} (exemple sur le chronogramme de la *Figure 29*) pour $V_{rev} = 19.8\,V$, à température ambiante, pour quatre configurations de temps morts ($t_h = 76{,}7\,ns$, $160\,ns$, $506\,ns$, $4{,}34\,\mu s$).

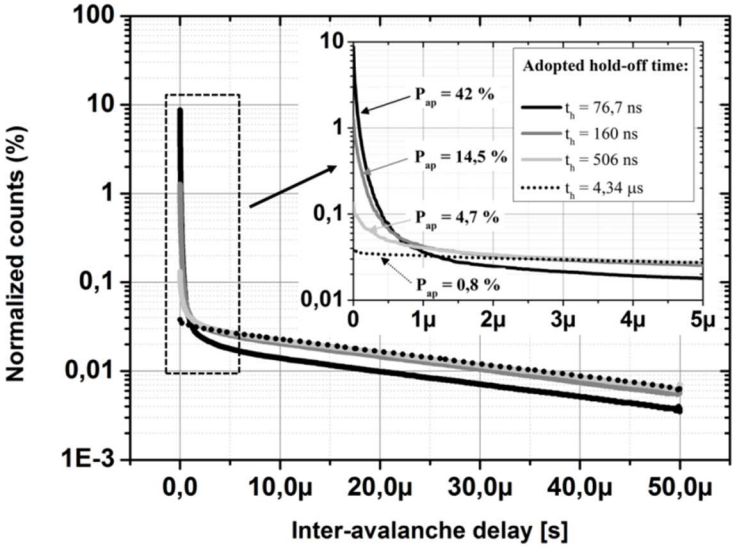

***Figure 35**: Histogramme de la répartition des événements en fonction du temps entre avalanches ($V_{rev} = 19{,}8\,V$).*

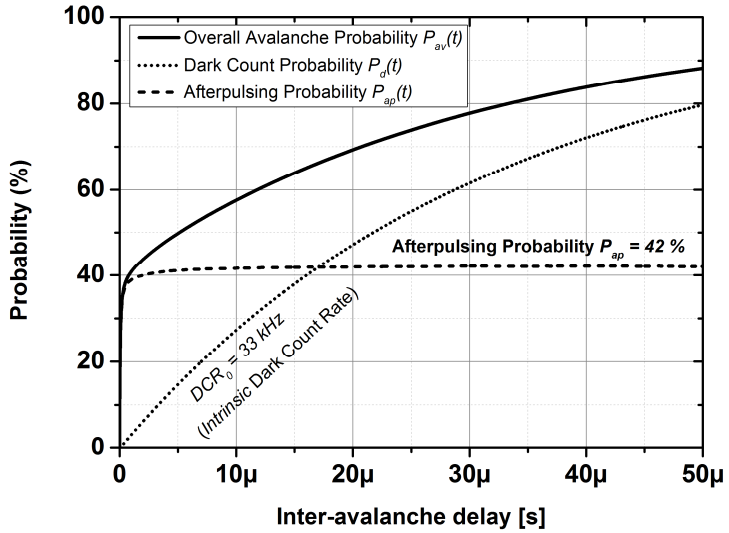

Figure 36 *: Probabilité d'avalanche pour $t_h = 76{,}7\ ns$ ($V_{rev} = 19{,}8\ V$). Les deux composantes sont également tracées : probabilité de déclenchement de pulses primaires et secondaires (after-pulsing)*

En intégrant la courbe noire de la *Figure 35* obtenue pour $t_h = 76{,}7\ ns$, nous obtenons la probabilité d'avalanche $P_{av}(t)$ tracée sur la *Figure 36*. Sur cette même figure, nous représentons également

- le niveau de probabilité de pulses primaires $P_d(t)$ telle que $P_d(t) = 1 - e^{-\lambda_0 t}$ (λ_0 est extrait en utilisant un long temps mort, exemple $t_h = 4{,}34\ \mu s$ sur la *Figure 35*),
- la probabilité de pulses secondaires $P_{ap}(t)$ en utilisant l'équation précédente.

Dans cette configuration de mesures ($V_{rev} = 19.8\ V$, $t_h = 76{,}7\ ns$), nous extrayons alors une probabilité d'after-pulsing atteignant 42 %.

Le Tableau 1 donne une comparaison des probabilités d'after-pulsing (pulses secondaires) obtenues par les méthodes dites « précise » et « rapide ». Les valeurs concordant plutôt biens ; ainsi les résultats de la *Figure 34*, obtenus avec la méthode dite « rapide », peuvent être considérés comme très fiables

Tableau 1 : Comparaisons des probabilités de déclenchement obtenues par les méthodes précise et rapide.

Temps mort	P_{ap} (méthode précise)	P_{ap} (méthode rapide)
76,7 ns	42 %	42,2 % ($DCR = 49\ kHz$)
106,8 ns	26,3 %	25,5 % ($DCR = 38\ kHz$)
160 ns	14,5 %	16,4 % ($DCR = 33{,}9\ kHz$)
506 ns	4,7 %	5,1 % ($DCR = 29{,}85\ kHz$)
4,34 μs	0,8 %	0 % ($DCR_{ref} = \lambda_0 = 28{,}3\ kHz$)

4.2.4 Efficacité de détection (Photon Detection Efficiency)

L'efficacité de détection (ou PDE pour photon detection efficiency ou PDP pour photon detection probability) caractérise la capacité d'une diode SPAD à détecter un photon. D'un point de vue pratique, le PDE est obtenu en faisant le rapport d'un taux de comptage observé moins le DCR divisé par le flux de photons (nombre de photons par seconde) arrivant sur le pixel [20]. Pour mesurer correctement le PDE, il faut s'assurer que le flux de photons est suffisamment faible pour avoir en moyenne un photon absorbé à chaque déclenchement de la diode. Si le flux de photons est trop élevé, les photons arrivant pendant le temps mort vont être systématiquement perdus quelle que soit l'efficacité de détection. Le taux de comptage mesuré R_m peut s'écrire :

$$R_m = \frac{R_{ph}PDE}{(1-P_{ap})} + \frac{DCR}{(1-P_{ap})} = \frac{R_{ph}PDE}{(1-P_{ap})} + DCR^*$$

avec R_{ph} : le flux de, photons incident, P_{ap} : la probabilité d'after-pulsing, PDE : l'efficacité de détection du pixel, DCR^* : le taux de comptage dans l'obscurité.
La présence du temps mort va introduire un correctif : diminution du taux de comptage effectif, comme décrit dans l'équation suivante [21]:

$$R_m(observé) = \frac{R_m}{1+R_m t_h}$$

Ainsi le flux de photons réellement détectés suit la relation suivante :

$$R_{ph}PDE = \left(\frac{R_m(observed)}{1-R_m(observed)t_h} - \frac{DCR^*(observed)}{1-DCR^*(observed)t_h}\right)\left(1 - P_{ap}\right)$$

Le banc de mesures électro-optiques (du laboratoire ICube) est équipé principalement : d'une source modulable de lumière blanche, d'un monochromateur, d'une sphère intégrante qui homogénéise le faisceau vers deux sorties : vers une photodiode calibrée (permettant de déterminer le flux de photons), et vers notre diode SPAD. La *Figure 37* illustre les premières mesures qui ont consisté à tracer le taux de comptage brut observé en fonction du flux de photons avec un temps mort $t_h = 178\,ns$ et une longueur d'onde $\lambda = 600$ nm,. Sur cette figure, nous retrouvons la valeur du DCR dans l'obscurité (à flux de photons nul), ensuite le taux de comptage croit linéairement avec le flux de photons puis sature (photons perdus pendant le temps mort ou la diode SPAD est « aveugle »). Si nous utilisons le calcul correctif établit précédemment, nous retrouvons bien un comptage corrigé linéaire. En pratique, nous allons régler la puissance de la source pour toujours rester dans le domaine linaire afin de ne pas avoir de correctif lié au temps mort à faire.

Avec les conditions optimales en termes de densité de puissance optique et de taux de comptage, nous avons mesuré et tracé le PDE :

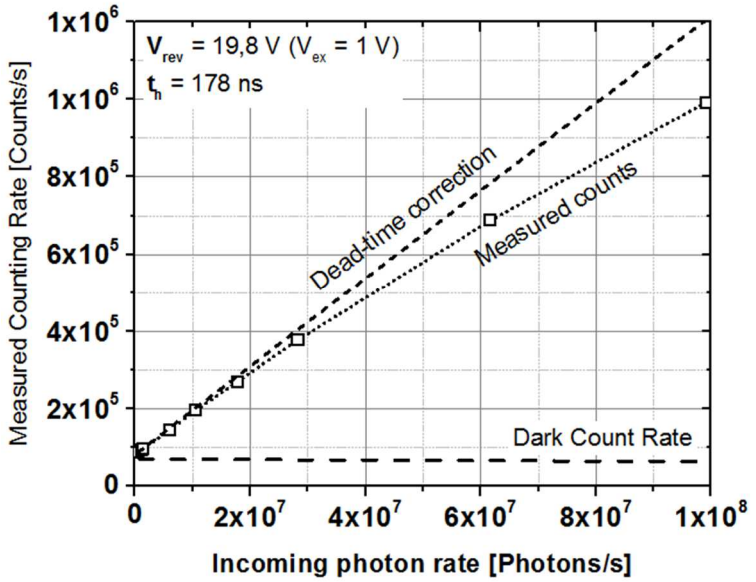

***Figure 37**: Etude de la linéarité du SPAD en mode comptage de photons avec $t_h = 178\,ns$ ($\lambda = 600\,nm$).*

- sur la *Figure 38* : en fonction de la tension inverse (à $\lambda = 600\,nm$), pour deux pixels (en incluant le PDE brut et le PDE corrigé de l'after-pulsing),

- sur la *Figure 39* : en fonction de la longueur d'onde (pour deux tensions inverses : $19.8\,V$ et $20.8\,V$).

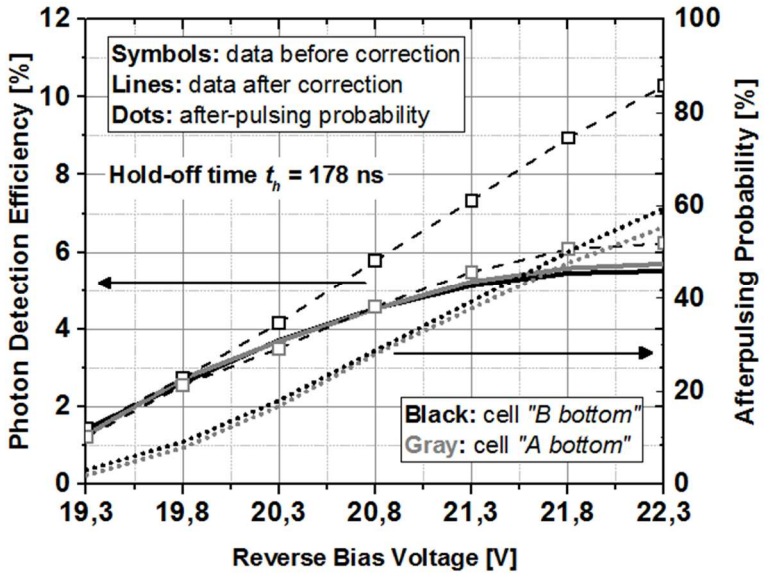

***Figure 38**: Efficacité de détection du SPAD à $\lambda = 600\,nm$ en fonction de la tension en inverse avec $t_h = 178\,ns$.*

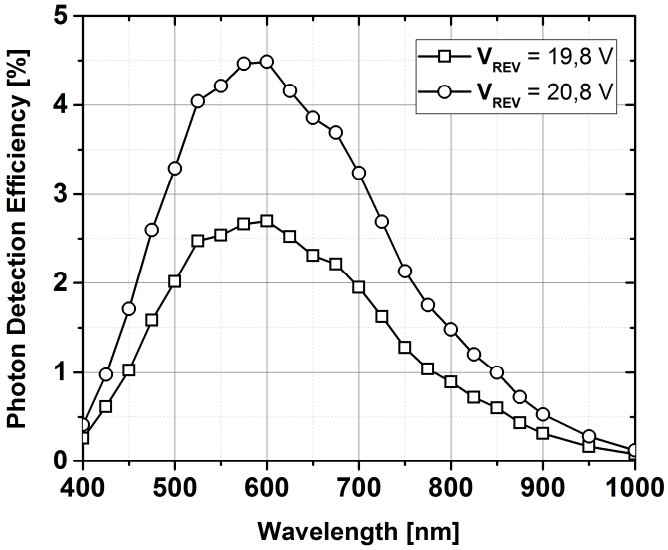

***Figure 39**: Efficacité de détection du SPAD pour deux tensions inverses ($V_{rev} = 19{,}8\,V$ et $V_{rev} = 20{,}8\,V$), avec $t_h = 178\,ns$.*

Nous notons que le PDE augmente quasi linéairement avec la tension inverse comprise $19.3\,et\,20.8\,V$ (grâce à l'augmentation de la probabilité de déclenchement), ensuite le PDE sature au-dessus de $21\,V$ (probablement car la probabilité tend vers 100%). Le maximum d'efficacité de détection autour de 5 % (pour $V_{rev} = 20.8\,V$) est atteint vers $\lambda = 600$ nm. Cette valeur maximale du PDE est relativement faible, de plus elle est décalée vers le milieu du spectre visible alors qu'habituellement le maximum de sensibilité est atteint en début du spectre visible (proche UV). Nous expliquons ce résultat par la présence d'une couche de passivation du test-chip ($2\,\mu m$ de matériau de type polyamide) avec un rôle absorbant non négligeable dans la gamme 400 à 600 nm.

4.3 Conclusion partielle

Dans ce paragraphe, nous avons présenté les principales caractéristiques de notre pixel SPAD, en commençant par la diode elle-même (tension de claquage, courant d'obscurité), puis le circuit d'étouffement (contrôle du temps mort), puis le pixel (bruit dans l'obscurité, probabilité de déclenchements secondaires, efficacité de détection).
Dans le prochain paragraphe, nous présentons les caractérisations expérimentales de notre nouveau détecteur de particules chargées, appelé 3D-SiCAD pour 3D Silicon Coincidence Avalanche Detector.

5. Caractérisation du prototype 3D-SiCAD

Dans ce paragraphe, nous présentons les résultats de caractérisation notre prototype 3D-SiCAD avec : *i)* les performances en bruit dans l'obscurité et la capacité à abaisser le niveau de bruit intrinsèque grâce à la coïncidence, *ii)* le comportement du détecteur pour la détection de particules ionisantes (utilisation d'une source radioactive dont nous déterminerons l'activité).

5.1 Performances en bruit

Les performances en bruit ont été mesurées en utilisant le même protocole expérimental que celui décrit au paragraphe 4. Dans une première phase, les mesures ont été réalisées sur deux diodes SPAD adjacentes dans le plan (*Figure 40* - gauche), pour ensuite caractériser le vrai prototype de 3D-SiCAD obtenu par l'empilement 3D et l'alignement de deux circuits (*Figure 40* - droite).

5.1.1 Mesures préliminaires en mode coïncidence sur deux pixels adjacents

Les premières mesures, présentées sur la *Figure 41*, concernent deux pixels adjacents sur lesquels les faux événements corrélés (Fake Coincidence Rate – FCR) ont été décomptés, grâce à la carte extérieure à base de FPGA, en fonction de la fenêtre de coïncidence temporelle avec un temps mort $t_h = 5{,}6\,\mu s$. Sur cette figure, les symboles représentent les mesures tandis que les lignes donnent les valeurs attendues calculées avec l'équation $FCR = 2 \cdot DCR_{top} \cdot DCR_{bottom} \cdot \Delta t$.

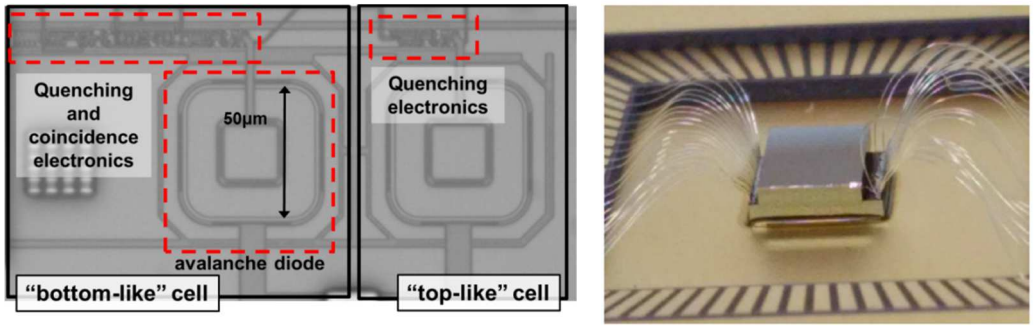

Figure 40: *à gauche) Photo de deux diodes SPAD adjacentes (et de leur électronique associée pour le quenching et la détection de coïncidence), à droite) photo du prototype 3D-SiCAD avec 2 puces empilées et les connections par wire-bondings sur les côtés gauche et droit.*

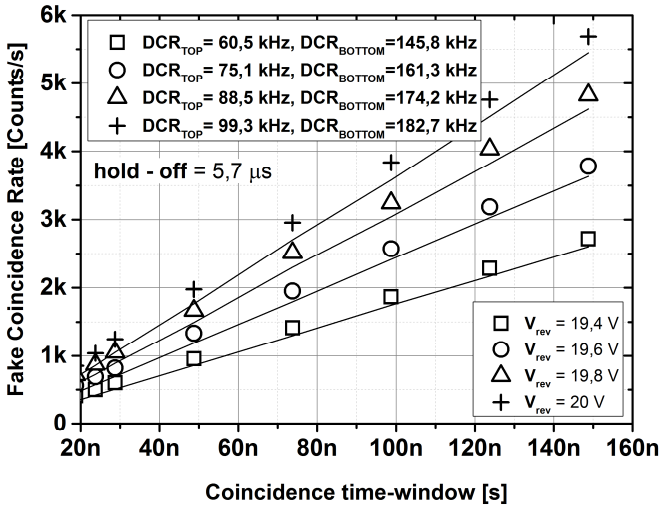

***Figure 41**: Taux de faux événements mesurés sur deux pixels adjacents en fonction de de la fenêtre de coïncidence temporelle pour différentes tensions inverses de* $19.4\,V$ *à* $20\,V$ *(temps morts* $t_h = 5{,}7\ \mu s$*).*

Nous relevons un bon accord théorie – expérience avec un taux de comptage de faux événements qui croit linéairement avec la largeur de la fenêtre temporelle de coïncidence. Également, le FCR augmente avec la tension inverse, comme attendu, en raison de l'augmentation du DCR. Des mesures complémentaires sont présentées sur la *Figure 42*, avec le taux de faux événements FCR en fonction de la fenêtre de coïncidence pour différents temps morts $t_h = 165\ ns$ à $t_h = 5{,}7\ \mu s$, avec une tension inverse $V_{rev} = 19{,}8\,V$. L'augmentation du temps mort permet de diminuer le DCR de chaque SPAD et par conséquent d'abaisser le taux de faux événements.

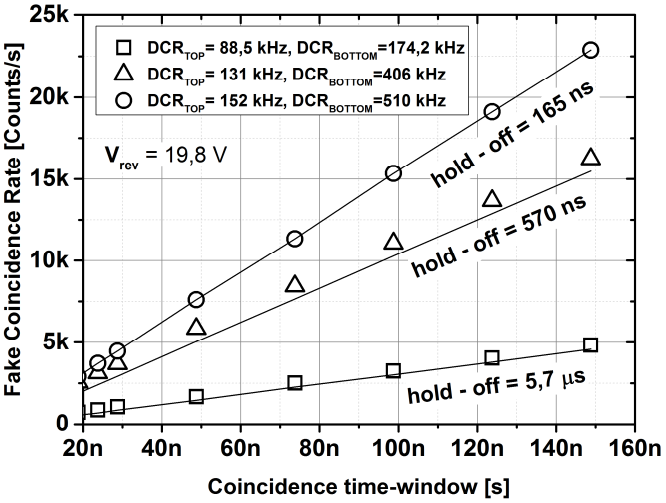

***Figure 42**: Taux de faux événements mesurés sur deux pixels adjacents en fonction de de la fenêtre de coïncidence temporelle pour différents temps morts* $t_h = 165\ ns$ *à* $t_h = 5{,}7\ \mu s$ *(avec une tension inverse* $V_{rev} = 19{,}8\,V$*).*

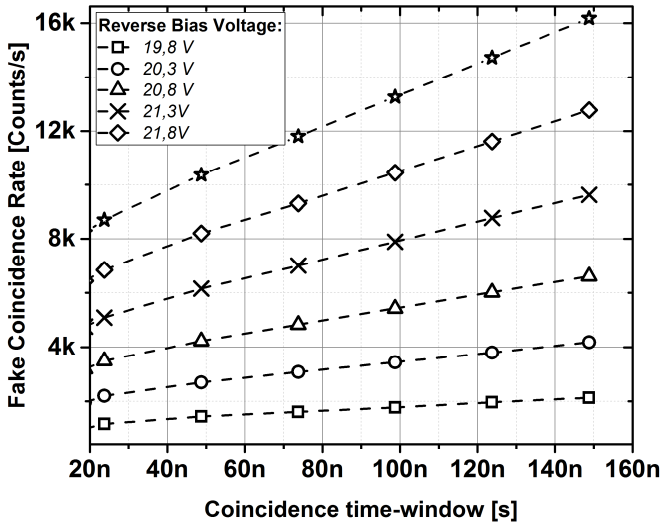

***Figure 43**: Taux de faux événements mesurés sur le prototype 3D-SICAD en fonction de la fenêtre de coïncidence pour différentes tensions inverses de* $19.8\,V$ *à* $21.8\,V$ *(temps mort* $t_h = 2.5\,\mu s$*).*

Sur ces deux figures, nous pouvons constater, par extrapolation, que si la fenêtre de coïncidence tend vers zéro, le taux de faux événements tend lui aussi vers zéro ; ce qui permet d'apprécier le grand potentiel de ce mode de détection pour abaisser le niveau de bruit. Les DCR mesurés sur nos SPAD sont dans la plage $100\,kHz - 500\,kHz$ (en fonction de la tension inverse et du temps mort) alors que, pour une fenêtre de coïncidence de $20\,ns$, nous mesurons un taux de faux événements entre $500\,Hz$ et $5\,kHz$ ce qui signifie 2 ordres de grandeur en dessous de leurs DCR respectifs. Un taux de réjection du bruit encore supérieur pourrait être obtenu en diminuant la fenêtre de coïncidence (et aussi en utilisant une technologie silicium moins bruyante).

5.1.2 Mesures sur le prototype 3D-SiCAD

De la même manière, les taux de faux événements FCR ont été mesurés sur le prototype 3D-SiCAD, les mesures sont reportées sur la *Figure 43* (pour différentes tensions inverses). Le Tableau 1 reprend les valeurs mesurées de DCR sur les SPADs constituant le prototype 3D-SiCAD. Comme attendu, nous constatons que le FCR augmente avec la fenêtre de coïncidence et la tension inverse, cependant les taux sont plus élevés que ceux prévus. En effet, avec une fenêtre de coïncidence de $20\,ns$, le FCR mesuré est dans la gamme $1\,kHz - 10kHz$, ce qui est un facteur 10 plus grand que les valeurs théoriques ($100\,Hz - 1kHz$). Contrairement aux mesures précédentes réalisées sur les deux SPADs adjacents, nous observons que les courbes ne tendent pas vers zéro par extrapolation aux très faibles fenêtres de coïncidence en raison d'un décalage sur les courbes. Une autre source d'événements corrélés est présente dans le prototype 3D-SiCAD. La section suivante contient des mesures approfondies et leur analyse.

Tableau 1: Valeurs mesurées de DCR sur les SPADs (bas et haut) constituant le prototype 3D-SiCAD.

Tension inverse	DCR SPAD du haut	DCR SPAD du bas
19,8 V	44 kHz	83 kHz
20.3 V	65 kHz	118 kHz
20.8 V	85 kHz	147 kHz
21.3 V	104 kHz	174 kHz
21.8V	123 kHz	200 kHz
22.3V	140 kHz	222 kHz

5.1.3 Mesures approfondies et analyse

Pour mieux comprendre le comportement du prototype 3D-SiCAD, l'électronique intégrée au pixel a été utilisée car elle permet (après étalonnage) d'atteindre des fenêtres de coïncidence temporelles plus courtes. Un étalonnage sur un pixel (avec une confrontation à la simulation) est présenté sur la *Figure 44*. En faisant varier la tension de contrôle externe $V_c = 0$ V à $V_c = 2{,}3$ V, la fenêtre de coïncidence varie de $\Delta t = 490\,ps$ à $\Delta t = 106\,ns$ avec un relativement bon accord entre mesures et simulations. En utilisant le circuit de coïncidence interne (intégré dans le pixel), le taux de faux événements en fonction de la fenêtre de coïncidence est représenté sur la *Figure 45* en échelle logarithmique. Nous constatons qu'au-dessus d'une fenêtre de coïncidence de $1.5\,ns$, le FCR subit un saut (le décalage déjà constaté).
Nous incriminons ce comportement à l'existence de couplage optique, tel qu'illustré sur la *Figure 11*.

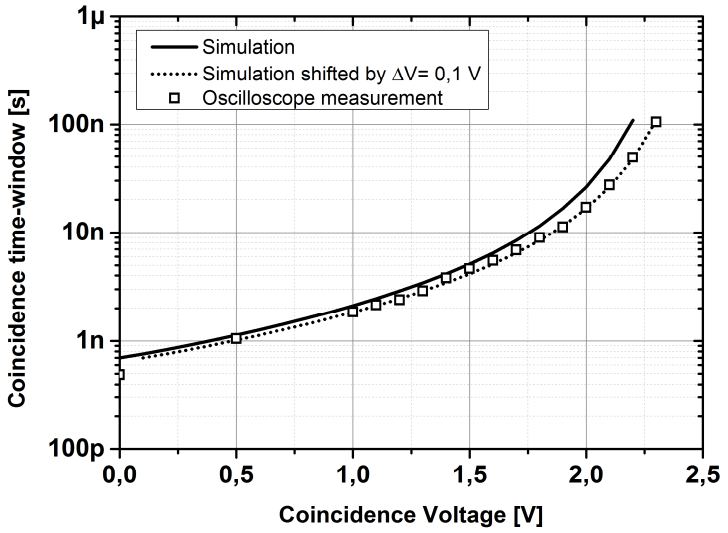

***Figure 44**: Mesure et simulation de la fenêtre de coïncidence imposée par l'électronique intégrée en fonction de la tension de contrôle externe.*

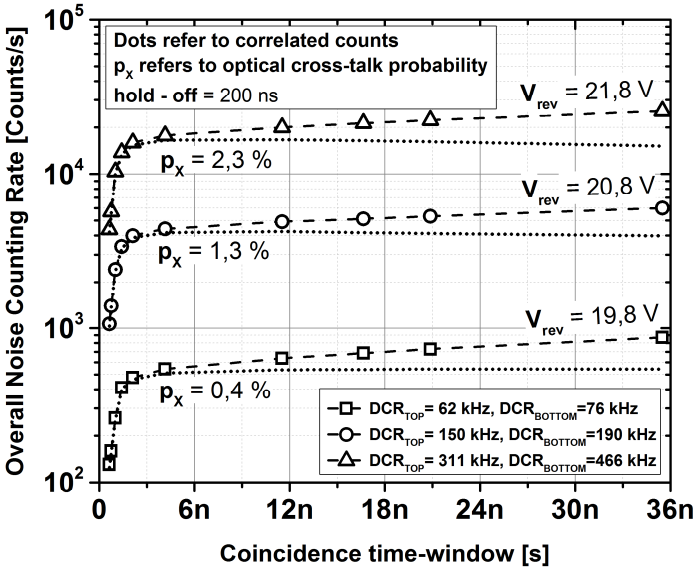

***Figure 45**: Taux de faux événements mesurés sur le prototype 3D-SiCAD en fonction de la fenêtre de coïncidence imposée par l'électronique intégrée pour différentes tensions inverses de* $19.8\,V$ *à* $21.8\,V$ *(temps morts* $t_h = 200\,ns$*).*

Nous rappelons que le taux de faux d'événements corrélés est donné par la relation suivante où le second terme représente la composante liée au couplage optique :

$$NCR(\Delta t) = 2\,\Delta t\,\lambda_{top,0}\lambda_{bottom,0} + P_X\left(\lambda_{top,0} + \lambda_{bottom,0}\right)$$

avec $\lambda_{top,0}$ et $\lambda_{bottom,0}$ les DCR de chaque SPAD (dessus ou dessous) sans cross-talk optique, Δt : la fenêtre temporelle d'observation, P_X : la probabilité de cross-talk optique. Le Tableau 2 donne les valeurs FCR mesurées et idéales pour différentes tensions inverses, ainsi que la probabilité de couplage optiques extraite de l'équation précédente. La *Figure 12* représente le facteur de réjection défini comme $NRF = \frac{NCR}{\lambda_0} = 2(P_X + \Delta t\lambda_0)$ en supposant que $\lambda_{top,0} = \lambda_{bottom,0} = \lambda_0$. Nous constatons qu'une très faible probabilité de couplage optique (même inférieure au %) introduit une composante non négligeable dans le taux de faux événements corrélés mesuré. Dans notre prototype 3D-SiCAD, la réduction du couplage optique pourrait être réalisée en introduisant une couche absorbante ou réfléchissante entre les deux circuits ; par exemple avec les niveaux de métaux disponibles dans la technologie CMOS.

Tableau 2: Probabilité de couplage optique extraite.

Tension inverse	$FCR_m\ @\Delta t = 4ns$	$FCR_{id}\ @\Delta t = 4ns$	P_X
19,8 V	542 Hz	39 Hz	0,4 %
20,8 V	4,4 kHz	237 Hz	1,3 %
21,8 V	17,7 kHz	1,2 kHz	2,3 %

5.2 Capacité du prototype 3D-SiCAD à détecter les particules chargées

Nous avons finalement testé la capacité du prototype 3D-SiCAD à détecter les particules chargées avec une source commerciale radioactive de Strontium-90 présentant une activité $\lambda_{90_{Sr}} = 37 \pm 11{,}1\ MBq$ (
). Nous avons mesuré le taux de comptage et tracé la loi inverse carrée en fonction de la distance pour retrouver l'activité théorique de la source. Le spectre de la source radioactive de Strontium-90 se caractérise par deux branches d'émissions de β^- associées aux ^{90}Sr et ^{90}Y (*Figure 47*), issu de [22].

5.2.1 Méthode

Le taux de comptage d'événements liés au passage des particules chargées suit une loi inverse carrée de la forme suivante :

$$CR_m = FCR + \lambda_{90_{Sr}} \alpha(t_{sub}) \frac{A_{pixel}}{4\pi(d_0+d)^2} \eta$$

avec FCR : taux de faux événement (bruit), $\lambda_{90_{Sr}}$: activité de la source radioactive, $\alpha(t_{sub})$: fraction des électrons beta qui atteignent la zone active du pixel après une épaisseur de silicium notée t_{sub}, A_{pixel} : surface du pixel 3D-SiCAD, d_0 : distance minimale entre source et pixel 3D-SiCAD, η : efficacité de détection, d : distance entre source et pixel 3D-SiCAD que nous faisons varier.

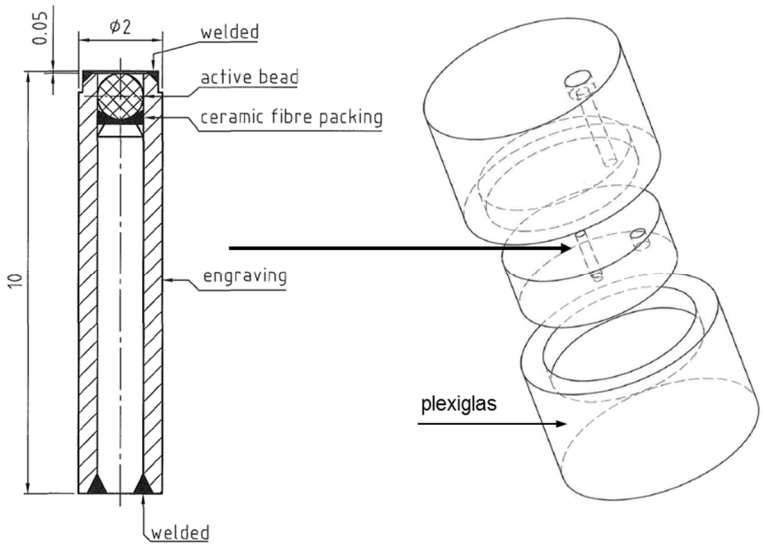

***Figure 46**: à gauche) la source de strontium 90Sr, à droite) le support en plexiglass dans laquelle elle est positionnée (réalisation : IPNL)*

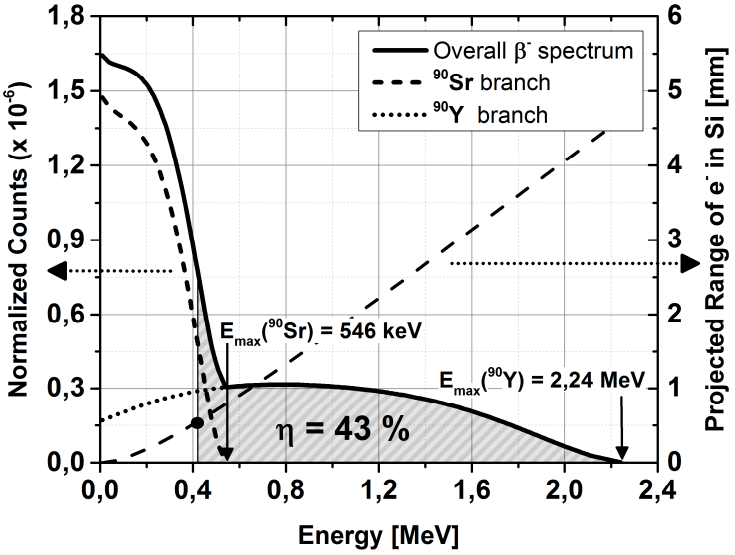

***Figure 47**: à gauche) spectre d'émission de la source de strontium avec les deux branches ^{90}Sr et ^{90}Y, à droite) profondeur de pénétration des β^- dans le silicium*

Afin de déterminer la fraction des électrons beta qui atteignent réellement la zone active du pixel, nous avons utilisé la loi empirique de Katz and Penfold [23], valable pour les électrons d'énergie inférieure à 2.5 MeV :

$$R_{Si} = \frac{1}{\rho_{Si}} 0{,}412\, E^{1{,}265-0{,}0954\ln(E)}$$

avec E : l'énergie des électrons, ρ_{Si} ; la densité du silicium.
Seuls les électrons avec une énergie supérieure à $\sim 0{,}42\ MeV$ sont susceptibles d'atteindre la zone active (*Figure 47*). Ainsi seulement une fraction de 43% des particules émises par la source radioactive, est détectable par notre prototype 3D-SICAD : $\alpha(t_{sub}) = \int_{E_0(t_{sub})}^{\infty} S_{90_{Sr}}(E)dE = 43\%$.

Le Tableau 3 récapitule les principaux paramètres. Le setup expérimental est schématiquement représenté sur la *Figure 48*. Finalement le taux réel mesuré d'événements liés aux particules chargées est obtenu avec : $CR_\beta = CR_m - FCR$ avec une acquisition sur 10 s.

Tableau 3: Principaux paramètres.

Paramètre	Valeur
$\alpha(t_{sub})$	43 %
A_{pixel}	2500 μm^2
d_0	5 mm
$\lambda_{90_{Sr}}$ (nominal)	37 $\pm$ 11,1 MBq

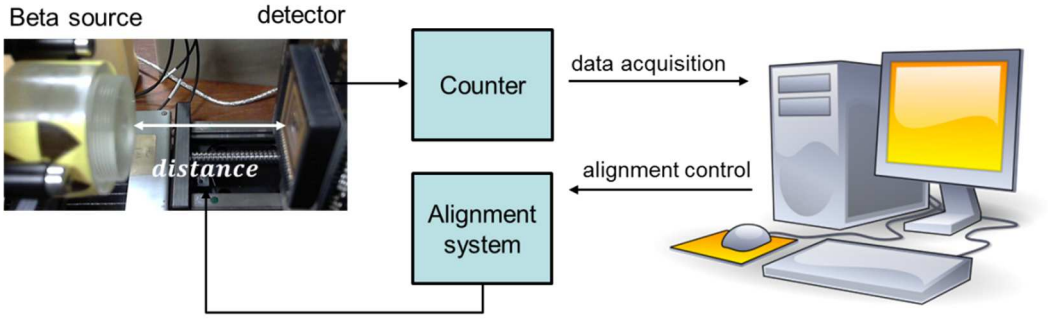

***Figure 48**: Vue schématique du setup expérimental installé à l'IPNL.*

5.2.2 Détermination des meilleures conditions de mesures

La première étape est la détermination des meilleures conditions expérimentales pour détecter les particules émises par la source radioactive de strontium. Ainsi nous reportons :

- Le taux de comptage en fonction du temps mort (tension V_h contrôlant le temps mort) sur la *Figure 49*. Nous constatons que la tension doit rester inférieure à $2.2\,V$ sinon le temps mort devient trop grand rendant le pixel « aveugle » trop longtemps. Une tension $V_h = 2\,V$ (c'est à dire un temps mort $t_h \sim 200\,ns$) sera utilisée car cela garantit une perte négligeable dans la détection de vrais événements (moins de 2% en se basant sur la valeur de DCR et t_h) tout en conservant une probabilité d'événements secondaires acceptable (inférieure à 10 %).

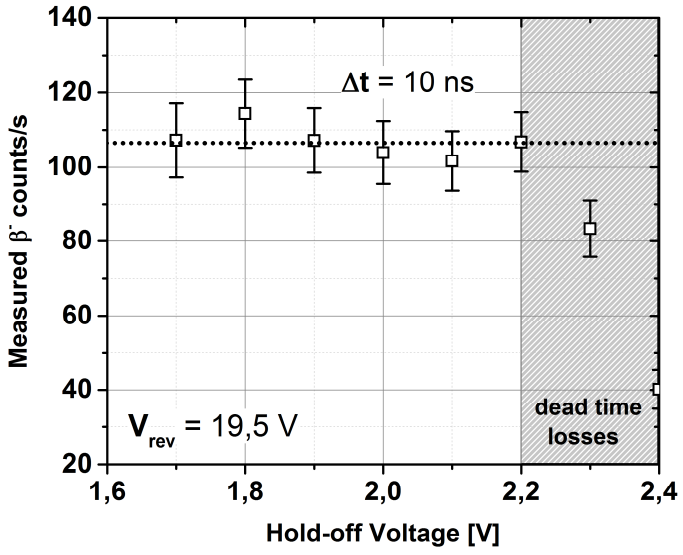

***Figure 49**: Taux de comptage des électrons en fonction de la tension contrôlant le temps mort ($V_{rev} = 19.5\,V$ et $\Delta t = 10\,ns$).*

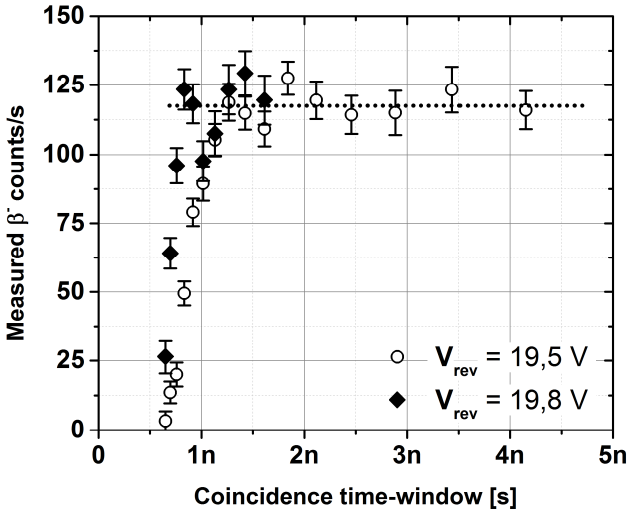

***Figure 50**: Taux de comptage des électrons en fonction de la fenêtre de coïncidence ($V_{rev} = 19.5\,V$ et $19.8\,V$).*

- Le taux de comptage en fonction de la fenêtre de coïncidence sur la *Figure 50*. Une fenêtre de coïncidence $\Delta t = 2\,ns$ est considérée comme suffisamment large pour ne pas perdre de vrais événements corrélés et suffisamment courte pour rejeter le bruit intrinsèque des SPADs
- Le taux de comptage en fonction de la tension inverse sur la *Figure 51*. La tension inverse doit être supérieure à $19.3\,V$ pour obtenir une tension d'excès V_{ex} supérieure à $0{,}5\,V$. Cette limitation est causée par le seuil de notre comparateur autour de $0.5\,V$.

Un calcul rapide de la capacité à détecteur les particules chargées peut être mené avec la relation :

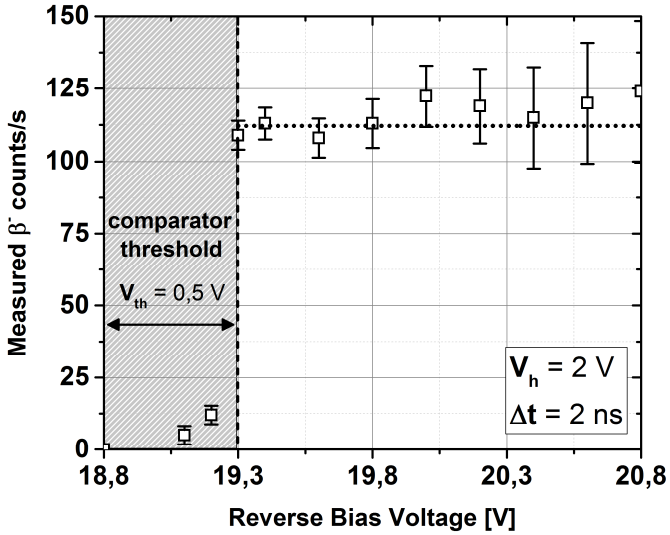

***Figure 51:** Taux de comptage des électrons en fonction de la tension inverse (V_h = 2 V, et $\Delta t = 2\,ns$).*

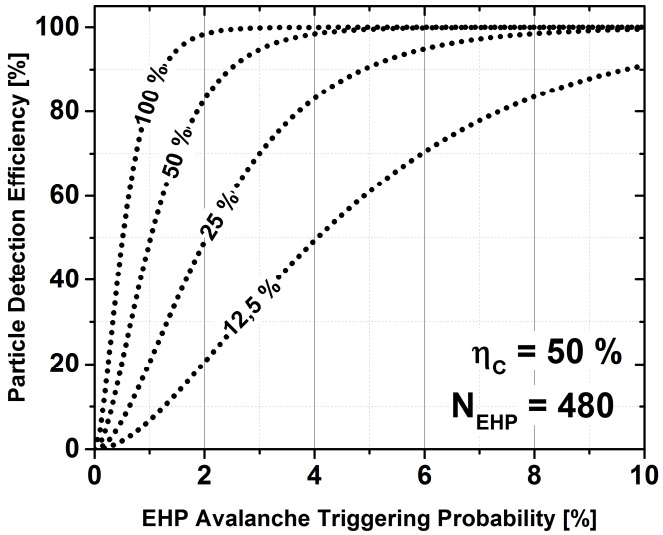

***Figure 52**: Efficacité de détection de particules chargées.*

$$\eta_{MIP_{3D-SiCAD}} \approx [1-(1-\eta_c P_{tr})^{N_{EHP}}]^2$$

avec $\eta_{MIP_{3D-SiCAD}}$: l'efficacité de détection d'une particule chargée, N_{EHP} : le nombre de paires électron-trou créés, η_c : la probabilité de collecter ces paires, P_{tr} : la probabilité de déclencher une avalanche. La *Figure 52* permet de comprendre que l'efficacité de détection tend rapidement vers 100 % pour une probabilité de détection supérieure à 10 % en raison de la valeur élevée de N_{EHP} (ici estimée à 480 c'est-à-dire 80 paires électrons-trous par micron sur une taille de zone de collecte estimée à $W_{SPAD} \approx 6\ \mu m$). Ainsi, il est inutile de travailler avec une forte tension inverse (impliquant un fort DCR), une tension inverse de 19.5 V a été choisie.

5.2.3 Mesure de la loi inverse carrée et détermination de l'activité de la source radioactive

En adoptant les conditions optimales de mesures déterminées précédemment ($V_h = 2\ V$ et $V_{rev} = 19{,}5\ V$), les mesures du taux de comptage en fonction de la distance d sont reportées sur la *Figure 53* pour trois fenêtres de coïncidence $\Delta t = 2, 4, 10$ ns. Un fit a permis d'extraire l'activité de la source qui est reportée dans le Tableau 4 pour les différentes valeurs de la fenêtre de coïncidence. L'activité de la source extraite par nos mesures est 13 % inférieure à l'activité nominale donnée par le constructeur mais celle-ci est donnée avec une incertitude de ± 30 % ; en conséquence nos mesures semblent tout à fait satisfaisantes. Il est bien entendu qu'une telle mesure n'aurait pas été possible avec un seul SPAD présentant un bruit intrinsèque (DCR) supérieur à l'activité de la source à mesurer. Le mode de coïncidence utilisé dans le détecteur 3D-SiCAD permet d'abaisser le niveau de bruit pour atteindre de faibles comptages ~400 coups par seconde sur notre pixel de $50 \times 50\ \mu m^2$.

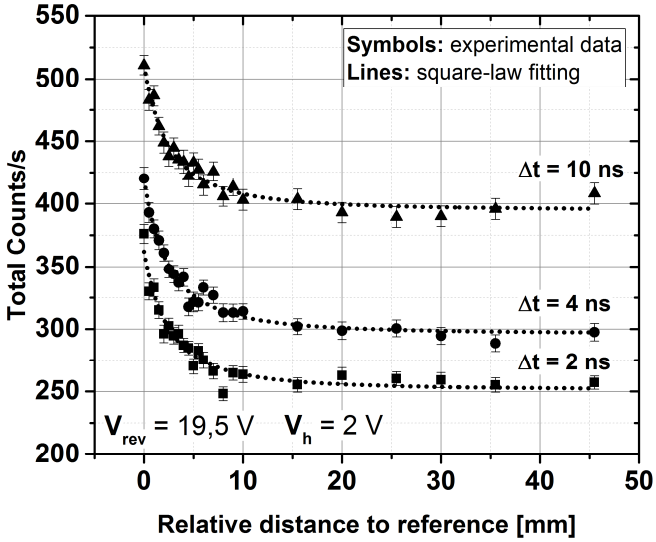

***Figure 53**: Taux de comptage des β^- en fonction de la distance entre source radioactive et détecteur 3D-SiCAD pour 3 fenêtres de coïncidence $\Delta t = 2, 4, 10\ ns$ ($V_h = 2\ V$ et $V_{rev} = 19{,}5\ V$).*

5.3 Conclusion partielle

Dans ce paragraphe, nous avons présenté les résultats de caractérisation de notre prototype 3D-SiCAD. Tout d'abord, le bruit sur deux pixels adjacents opérant en mode de coïncidence a été reporté, démontrant une diminution du bruit d'un facteur 10^2 à 10^3. Les mesures sur le pixel 3D-SiCAD n'ont pas reproduit ce très bon résultat avec seulement une réjection d'un facteur 10 à 10^2 environ en raison de la présence présumée d'un couplage optique entre les deux SPADS en configuration face à face. Cette limitation pourra être corrigée en intégrant une couche absorbante ou réfléchissante entre les deux niveaux. Finalement, le prototype 3D-SiCAD a été utilisé pour mesurer l'activité d'une source radioactive de strontium-90. Les mesures, par une approche de loi inverse carrée, ont permis d'extraire une activité de la source tout à fait en accord avec les données du constructeur. Une telle mesure n'aurait pas été possible avec un simple SPAD dans cette technologie en raison du niveau de bruit intrinsèque trop élevé.

Tableau 4: Activité de la source extraite de la loi inverse carrée.

Δt	d_0	$\alpha(t_{sub})$	$FCR\ @\ V_{rev} = 19{,}5\ V$	Activité extraite
2 ns	5 mm	43 %	$251 \pm 2\ Hz$	$32{,}3\ \pm 1{,}8\ MBq$
4 ns	5 mm	43 %	$296 \pm 2\ Hz$	$35{,}9\ \pm 1{,}4\ MBq$
10 ns	5 mm	43 %	$395 \pm 2\ Hz$	$33{,}8\ \pm 1{,}5\ MBq$

6. Conclusion générale et Perspectives

Ce travail de recherche s'est focalisé sur la conception, la réalisation et la caractérisation d'un premier prototype de nouveau détecteur nommé 3D-SiCAD pour 3D Silicon Coincidence Avalanche Detector.
Nous rappelons les points principaux présentés dans ce manuscrit de thèse (et particulièrement dans ce long résumé en français)

- Le deuxième paragraphe a permis de rappeler les éléments essentiels sur les diodes SPADs (physique, principe de fonctionnement, figures de mérites, réalisations en technologie CMOS). Ensuite le fondement du nouveau détecteur 3D-SiCAD est détaillé.
- Le troisième paragraphe a présenté toute la phase de conception de la zone active du détecteur à l'intégration 3D en passant par l'électronique intégrée au niveau pixel pour l'étouffement - quenching et la coïncidence. Dans la technologie CMOS haute tension 0,35 μm de la société AustriaMicroSystem (AMS), la diode SPAD est de type diffusion p+ / nwell profond. L'électronique associée permet de régler un temps mort entre $50\,ns$ et $5\mu s$, ainsi que la fenêtre de coïncidence entre $0{,}5\,ns$ et $50\,ns$.
- Les résultats de caractérisation du pixel simple sont inclus dans le quatrième paragraphe. La tension de claquage du pixel est de $18{,}77 \pm 0.13\,V$. Les images en électroluminescence ont révélé une bonne uniformité. Le bruit dans l'obscurité (« dark count rate » DCR), pour une tension d'excès de $1\,V$, est extrait à une valeur médiane $70\,Hz/\mu m^2$ (c'est-à-dire que la moitié des diodes SPAD présentent un DCR plus faible que cette valeur). Ces résultats sont conformes aux résultats présentés par Vilella et al. pour une architecture similaire de SPAD [19]. Le maximum d'efficacité de détection PDE autour de 5% (pour $V_{rev} = 20.8\,V$) est atteint vers $\lambda = 600$ nm. Cette valeur maximale du PDE est relativement faible, probablement en raison de la présence d'une couche de passivation du test-chip (2 μm de matériau de type polyamide).
- Les résultats de caractérisation du prototype 3D-SiCAD ont été présentés dans le cinquième paragraphe. Le bruit sur deux pixels adjacents opérant en mode de coïncidence a été reporté, démontrant une diminution du bruit d'un facteur 10^2 à 10^3 par rapport au bruit intrinsèque de chaque SPAD. Les mesures sur le pixel 3D-SiCAD n'ont pas reproduit ce très bon résultat avec seulement une réjection d'un facteur 10 à 10^2 environ en raison de la présence présumée d'un couplage optique entre les deux SPADS en configuration face à face. Cette limitation pourra être corrigée en intégrant une couche absorbante ou réfléchissante entre les deux niveaux. Finalement, le prototype 3D-SiCAD a été utilisé pour mesurer l'activité d'une source radioactive de strontium-90. Les mesures, par une approche de loi inverse carrée, ont permis d'extraire une activité de la source tout à fait en accord avec les données du constructeur. Une telle

mesure n'aurait pas été possible avec un simple SPAD dans cette technologie en raison du niveau de bruit intrinsèque trop élevé.

Les perspectives à court-terme concernent :

- La poursuite des mesures avec la source radioactive afin d'optimiser davantage les conditions expérimentales, notamment descendre la fenêtre de coïncidence au minimum pour abaisser le plancher de bruit.
- L'amélioration de la plateforme expérimentale afin d'automatiser les mesures pour tirer davantage de données statistiques.
- La mesure sous faisceau clinique de protons (Centre Lacassagne à Nice, faisceau 65 MeV) afin de déterminer le taux de comptage maximal du prototype 3D-SiCAD (et sa linéarité en énergie).
- La conception d'un nouveau prototype 3D-SiCAD avec les améliorations suivantes à implémenter : une électronique permettant de travailler avec de plus faibles tensions d'excès, l'intégration de matrices de plus grande taille, le passage à une technologie CMOS moins bruyante (mais cela nécessite de revalider la cellule SPAD de base), l'utilisation d'une technique d'intégration 3D plus performante telle que celle présentée sur la
- *Figure **19***.

L'amélioration du facteur de remplissage est également nécessaire pour augmenter l'efficacité de détection de notre détecteur 3D-SiCAD dans le cas d'une matrice. Une première solution serait de passer de deux à trois couches en empilant deux niveaux de zones sensibles et un niveau pour l'électronique. Cette solution, au premier abord attrayante, est délicate : d'une part par sa complexité et son coût, et d'autre part cela augmenterait l'épaisseur de matériau (« material budget »), la rendant incompatible avec les applications en physique des particules. Ainsi en parallèle de ce travail de thèse, nous avons mené une pré-étude sur la faisabilité d'une intégration de diode SPAD dans une technologie CMOS FDSOI (Fully Depleted Silicon-On-Insulator) avancée [10][24]. Comme illustré sur la *Figure **54***, l'originalité vient du fait d'intégrer la diode SPAD sous le box en utilisant les couches présentes pour le « back-biasing » des transistors.

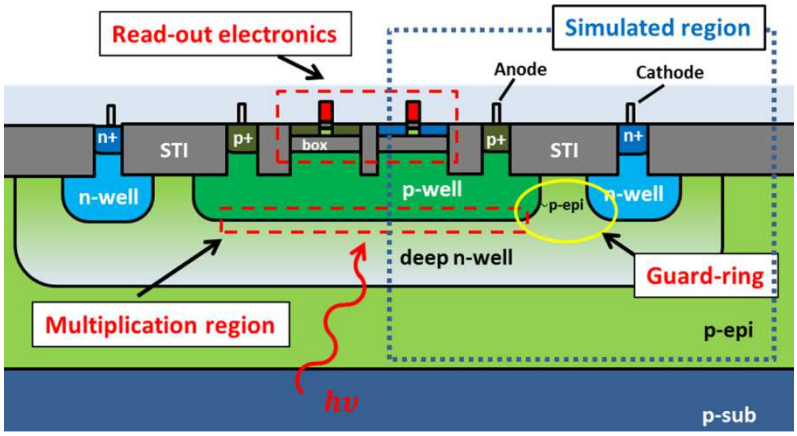

***Figure 54**: Intégration de SPAD dans une technologie CMOS FDSOI avancée.*

La diode SPAD est réalisée avec la jonction entre caisson p et nwell profond. Une étude TCAD a été menée permettant d'optimiser l'architecture de l'anneau de garde (dopage rétrograde du nwell profond) afin d'éviter tout claquage prématuré sur les bords. Ainsi la *Figure 55* illustre une carte de champ électrique 2D optimisée avec, en insertion, le tracé du champ dans une coupe verticale au milieu du pixel.
Cette nouvelle architecture, très prometteuse, permet d'obtenir un pixel naturellement 3D avec le niveau de la zone de détection (diode SPAD) sous la couche de transistors (dans le film silicium) pour réaliser les circuits d'étouffement, d'adressage des pixels etc. Cette étude devra se poursuivre car elle permettra de réaliser des détecteurs de particules chargées optimisées du type 3D-SICAD à deux niveaux avec de bien meilleurs facteurs de remplissage mais elle adresse aussi la détection de photons avec un éclairage en face arrière (après affinement du substrat).

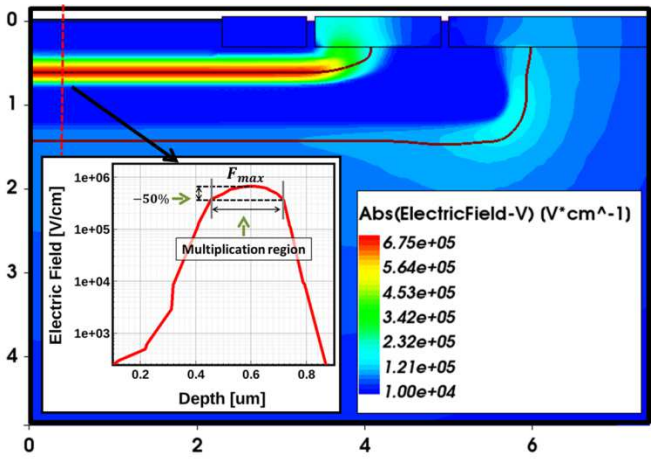

***Figure 55**: Carte de champ électrique du SPAD intégré en technologie CMOS FDSOI $V_{rev} = 16.5\,V$ (échelles en µm).*

Références bibliographiques

[1] "The Large Hadron Collider at CERN." [Online]. Available: http://home.cern/topics/large-hadron-collider.

[2] J. Christiansen, "Outline and requirements of Phase 2 Pixel system and Read-Out Chip," 2014.

[3] N. D'Ascenzo, P. S. Marrocchesi, C. S. Moon, F. Morsani, L. Ratti, V. Saveliev, A. S. Navarro, and Q. Xie, "Silicon avalanche pixel sensor for high precision tracking," *J. Instrum.*, vol. 9, no. 3, p. C03027, 2014.

[4] C. Golnik, F. Hueso-González, A. Müller, P. Dendooven, W. Enghardt, F. Fiedler, T. Kormoll, K. Roemer, J. Petzoldt, A. Wagner, and G. Pausch, "Range assessment in particle therapy based on prompt γ -ray timing measurements," *Phys. Med. Biol.*, vol. 59, no. 18, p. 5399, 2014.

[5] J. Krimmer, L. Balleyguier, D. Dauvergne, N. Freud, J. Hérault, J. M. Létang, H. Mathez, M. Pinto, E. Testa, and Y. Zoccarato, "Prompt-gamma detection towards absorbed energy monitoring during hadrontherapy," *ANNIMA Conf.*, 2015.

[6] V. Saveliev, "Avalanche Pixel Sensor and Related Methods," US patent 8269181, 2012.

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

[8] F. Zappa, S. Tisa, a. Tosi, and S. Cova, "Principles and features of single-photon avalanche diode arrays," *Sensors Actuators, A Phys.*, vol. 140, no. 1, pp. 103–112, 2007.

[9] R. J. McIntyre, "On the avalanche initiation probability of avalanche diodes above the breakdown voltage," *IEEE Trans. Electron Devices*, vol. 20, no. 7, 1973.

[10] M. M. Vignetti, F. Calmon, P. Lesieur, and A. Savoy-Navarro, "Simulation study of a novel 3D SPAD pixel in an advanced FD-SOI technology," *Solid. State. Electron.*, vol. 128, pp. 163–171, 2017.

[11] A. Rochas, A. R. Pauchard, P. A. Besse, D. Pantic, Z. Prijic, and R. S. Popovic, "Low-noise silicon avalanche photodiodes fabricated in conventional CMOS technologies," *IEEE Trans. Electron Devices*, vol. 49, no. 3, pp. 387–394, Mar. 2002.

[12] M. a Karami, H. J. Yoon, and E. Charbon, "Single-photon Avalanche Diodes in sub-100nm Standard CMOS Technologies," *Intl. Image Sens. Work.*, 2011.

[13] L. Pancheri, P. Brogi, G. Collazuol, G. F. Dalla Betta, A. Ficorella, P. S. Marrocchesi, F. Morsani, L. Ratti, and A. Savoy-Navarro, "First prototypes of two-tier avalanche pixel sensors for particle detection," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers,*

Detect. Assoc. Equip., pp. 1–4, 2016.
[14] M. M. Vignetti, F. Calmon, R. Cellier, P. Pittet, L. Quiquerez, and A. Savoy-Navarro, "A time-integration based quenching circuit for Geiger-mode avalanche diodes," in *New Circuits and Systems Conference (NEWCAS), 2015 IEEE 13th International*, 2015, pp. 1–4.
[15] "3DinCities." [Online]. Available: http://www.3dincites.com/3d-incites-knowledge-portal/what-is-3d-integration/.
[16] S. M. Sze and K. Ng, *Physics of Semiconductor Devices*, 3rd Editio. 2006.
[17] Sidi Aboujja, "Électroluminescence en avalanche des jonctions p-n à base de silicium et d'arséniure de gallium, et effet d'irradiation," Université de Sherbrooke, 2000.
[18] A. L. Lacaita, F. Zappa, S. Bigliardi, and M. Manfredi, "On the bremsstrahlung origin of hot-carrier-induced photons in silicon devices," *IEEE Trans. Electron Devices*, vol. 40, no. 3, pp. 577–582, Mar. 1993.
[19] E. V. Figueras, "Feasibility of Geiger-mode avalanche photodiodes in CMOS standard technologies for tracker detectors Feasibility of Geiger-mode avalanche photodiodes in CMOS standard technologies for tracker detectors," (PhD Thesis), University of Barcelona, 2013.
[20] F. Villa, D. Bronzi, Y. Zou, C. Scarcella, G. Boso, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, U. Paschen, and W. Brockherde, "CMOS SPADs with up to 500 μm diameter and 55% detection efficiency at 420 nm," *J. Mod. Opt.*, vol. 61, no. 2, pp. 102–115, 2014.
[21] E. Sciacca, A. C. Giudice, D. Sanfilippo, F. Zappa, S. Lombardo, R. Consentino, C. Di Franco, M. Ghioni, G. Fallica, G. Bonanno, S. Cova, and E. Rimini, "Silicon planar technology for single-photon optical detectors," *IEEE TED*, vol. 50, no. 4, pp. 918–925, 2003.
[22] "Laboratoire National Henri Becquerel." [Online]. Available: http://www.nucleide.org/.
[23] L. Katz and A. S. Penfold, "Range-energy relations for electrons and the determination of beta-ray end-point energies by absorption," *Rev. Mod. Phys.*, vol. 24, no. 1, pp. 28–44, 1952.
[24] M. M. Vignetti, F. Calmon, P. Lesieur, F. Dubois, T. Graziosi, and A. Savoy-Navarro, "A novel 3D pixel concept for Geiger-mode detection in SOI technology," in *2016 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, 2016, pp. 166–169.

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : **VIGNETTI** DATE de SOUTENANCE : **09/03/2017**

Prénoms : **Matteo Maria**

TITRE :

Development of a 3D Silicon Coincidence Avalanche Detector (3D-SiCAD) for charged particle tracking

NATURE : Doctorat Numéro d'ordre : 2017LYSEI017

Ecole doctorale : EEA (Électronique, Électrotechnique et Automatique)

Spécialité : Electronique, micro et nanoélectronique, optique et laser

RESUME :

L'objectif de cette thèse est de développer un détecteur innovant de particules chargées, dénommé 3D Silicon Coincidence Avalanche Detector (3D-SiCAD), réalisable en technologie silicium CMOS standard avec des techniques d'intégration 3D. Son principe de fonctionnement est basé sur la détection en « coïncidence » entre deux diodes à avalanche en mode « Geiger » alignées verticalement, avec la finalité d'atteindre un niveau de bruit bien inférieur à celui de capteurs à avalanche standards, tout en gardant les avantages liés à l'utilisation de technologies CMOS ; notamment la grande variété d'offres technologiques disponibles sur le marché, la possibilité d'intégrer dans un seul circuit un système complexe de détection, la facilité de migrer et mettre à jour le design vers une technologie CMOS plus moderne, et le faible de coût de fabrication. Le détecteur développé dans ce travail se révèle particulièrement adapté au domaine de la physique des particules de haute énergie ainsi qu'à la physique médicale - hadron thérapie, où des performances exigeantes sont demandées en termes de résistance aux rayonnements ionisants, « material budget », vitesse, bruit et résolution spatiale. Dans ce travail, un prototype a été conçu et fabriqué en technologie HV-CMOS 0,35µm, en utilisant un assemblage 3D de type « flip-chip » avec pour finalité de démontrer la faisabilité d'un tel détecteur. La caractérisation du prototype a finalement montré que le dispositif développé permet de détecter des particules chargées avec une excellente efficacité de détection, et que le mode « coïncidence » réduit considérablement le niveau de bruit. Ces résultats très prometteurs mettent en perspective la réalisation d'un système complet de détection CMOS basé sur ce nouveau concept.

MOTS-CLÉS : détecteur particules chargées, 3D, technologie CMOS standard, détection en coïncidence, mode Geiger, SPAD, physique des particules de haute énergie, physique médicale.

Laboratoire (s) de recherche : Institut des Nanotechnologies de Lyon (INL) – UMR CNRS 5270

Directeur de thèse: Francis CALMON
Co-directrice de thèse : Aurore SAVOY-NAVARRO

Président de jury :

Composition du jury : Lucio PANCHERI (rap.), Wilfried UHRING (rap.), Denis DAUVERGNE, Patrick PITTET, Aurore SAVOY-NAVARRO, Francis CALMON.

Invités : Dominique GOLANSKI, Gabriel PARES, Alexis ROCHAS.