POLITECNICO DI TORINO
GRENOBLE INP
EPFL

Master's Degree in Nanotechnologies for ICTs

Master's Degree Thesis

Analysis and test of a novel photodetector device

Supervisors

Prof. Edoardo CHARBON

Prof. Guido MASERA

Francesco GRAMUGLIA

Candidate

Carlo Alberto FENOGLIO

SEPTEMBER 2021

Summary

This thesis presents the test of an innovative 3D-stacked front-side illuminated (FSI) multi-channel digital silicon photomultiplier (MD-SiPM) fabricated in 0.18 µm CMOS technology for time-of-flight positron emission tomography (TOF-PET). During the master project, carried out at AQUA laboratory at EPFL, a complete testing system for the chip was designed and implemented. This work focuses on the implementation of the necessary firmware to interface with the unit under test, and the related software. The former is designed on a Opal Kelly XEM 7630 board, integrating a Kintex 7 FPGA from Xilinx. This board is fully supported by the FrontPanel SDK, a C++ class library to interface a software with the board. This is exploited to program a custom graphical user interface (GUI) to perform tests in a semi-automatic way. Its main purpose is to transfer data between PC and FPGA, so to provide an effective controllability and observability of the design. On the FPGA side of the interface, some FrontPanel's HDL modules work just as external pins, enabling the communication with the PC. The designed system was also employed to perform a preliminary characterisation of the chip, with the aim of verifying its basic functionalities and the successful outcome of this innovative, 3D-integrated design. On-chip time-to-digital converters' (TDCs) characterisation, dark count rate (DCR), photon counts and preliminary photon detection efficiency (PDE) measurements are presented. Some corruption of the TDCs' output bits will need further testing to be resolved, but, in general, they look active and working. The noise level results particularly higher than usual values and some artifacts might also be present in the counting system. The PDE, although preliminary and not complete, shows, instead, a very promising result, following the expected trend. A calibration with a photodiode will be necessary, to get a precise estimation. The chip seems to work correctly overall, apart from some minor issues. The 3D integration process looks successful and can be exploited for future designs. Nevertheless, a more detailed characterisation of the system is still ongoing and more accurate results will be available in the future.

Acknowledgements

I would first like to thank my supervisor, Prof. Edoardo Charbon, and Dr. Caludio Bruschini for the opportunity to study and work at AQUA and be part of its wonderful research group. Their plentiful expertise and insightful feedbacks were inspiring during these months.
My gratitude goes to Politecnico di Torino for the scholarship they granted me, and to supervisor Prof. Guido Masera for the careful review of this work and his valuable comments.
I am deeply grateful to Francesco Gramuglia and Dr. Emanuele Ripiccini, the best tutors and office partners, for their mentorship. Their precious guidance, kind advice and complete support helped me to successfully complete my thesis. Thank you for your friendship and the cheerful moments we shared together.
I would like to thank all the people of the AQUA laboratory group for the friendly welcome from the first day. Thanks for your patience and help.
My special thanks go to my friends and colleagues, sharing this period in Switzerland with me. You made this time marvellous and unforgettable.
All of this would not have been possible without the constant support, tremendous understanding, encouragement and extraordinary patience of my family. You are special.

Table of Contents

List of Tables

List of Figures

Acronyms

APD
Avalanche photodiode

ARC
Active recharge circuit

BGO
Bismuth germanium oxyde

CMOS
Complementary metal-oxide semiconductor

DCR
Dark count rate

FDG
Fluorodeoxyglucose

FIFO
First-in-first-out

FF
Fill factor

FPGA
Field-programmable gate array

FSI/BSI
Front-/Back-side illuminated

FSM

Finite state machine

LOR

Line of response

LYSO

Lutetium-yttrium oxyorthosilicate

MCP

Multi-channel plate

NEP

Noise equivalent power

OPU

Output processing unit

PCB

Printed circuit board

PET

Positron emission tomography

PMT

Photomultiplier tube

PDE

Photon detection efficiency

PDP

Photon detection probability

SEM

Scanning electron microscope

SiPM

Silicon photo-multiplier

SPAD

Single-photon avalanche diode

TDC

Time-to-digital converter

TSPC-FF

True single-phase clock flip-flop

TSV

Through-silicon via

Chapter 1

Introduction

In the last decades, an increasing number of applications required more and more precise detectors for very low light intensity measurements. Most common applications include medical physics, particle physics, Light Detection and Ranging (LiDAR) and quantum optics. As result, the performances of established detectors were pushed further towards the limit of one single photon detection while new technologies were developed with the same goal. The challenge was huge: being able to recognize discrete photons required unprecedented precision and incredibly low noise levels. Nonetheless, it was undertaken with great effort by the scientific community all over the world and several promising solutions were developed. In fact, many of these represented an evolution of already established technologies [1, 2, 3].

In particular, single-photon avalanche diodes (SPADs) gained increased attention thanks to their performances and interesting properties [4, 5]. They have almost completely replaced photomultiplier tubes (PMTs) for single photon detection thanks to their reliability and compactness, relatively low noise, high photon detection probability (PDP) and outstanding timing performances [4, 5, 6, 7, 8]. In the last years, the design of SPADs in commercial CMOS technology was developed, paving the way to more sophisticated circuits, functionalities, and reduced cost [6, 9, 10, 11, 12, 13].

In this context, analog silicon photomultipliers (A-SiPM) attracted great interest due to their robustness, low noise and high gain, especially for ToF-PET applications [14, 15, 16]. On the other hand, some years ago, the design of multi-channel digital silicon photomultipliers (MD-SiPM) introduced the possibility of an increased timestamp granularity and data pre-processing on chip [14, 17]. The implementation in CMOS technology allows cost reduction and the direct integration of circuits on chip SoC (System on Chip) implementation. By contrast, high-performance CMOS devices require advanced, nanometric technology nodes, not well suited for efficient SPADs. An innovative solution to fully overcome this

issue is the 3D-stacked design: a top-tier including only the photo-sensitive devices (SPADs) and one, (or more) bottom tier for the front-end circuitry [14, 18, 19, 20, 21]. The reason is that, in 3D-integrated sensors design, we can select the proper technology node for each tier of the detector, e.g. a CMOS image sensor (CIS) process for the top tier and a state-of-the-art, low power, high-performance, nanometric standard CMOS process for the bottom one. Additional benefits are the increased fill factor, reduction of the device pitch by eliminating the circuits next to the SPADs and the integration of additional, advanced functionalities (e.g. complex pixel circuits, TDCs, readout logic). Examples can be found in [22, 23, 24, 25, 26, 27]. The first examples were implemented in back-side illumination (BSI) approach [28, 22, 29]. This implementation presents several limitations, including a necessary back-thinning process and low sensitivity for wavelengths below 450 nm. In this work, we analyse an innovative 3D-stacked front-side illuminated (FSI) MD-SiPM called Blueberry, described in [14, 18] and first proposed in [20]. The chip is compatible with scintillator-based PET. The top tier of the chip is used for light detection and houses all the SPAD sensors. The bottom tier, on the contrary, includes all the electronics required by the system, such as readout logic, TDCs and counting functions. To our knowledge, this is the first chip of its kind, successfully implemented [30]. As such, its analysis aims to prove first its basic functionalities and to verify the successful outcome of this innovative design. The purpose of this work is to provide an initial and preliminary characterisation of the chip, and to develop the necessary structures for future, more advanced and quantitative tests. This first chapter provides the required theoretical background on single-photon detection, state-of the-art solutions and the target application: ToF-PET. The following one briefly describes the chip architecture, moving from the single SPAD pixel, to the top-level structure and the 3D integration process. The third chapter focuses on the developed test system, comprising a PCB, the firmware to perform the tests and communicate with the chip, and the relative GUI, to control and perform the test in a semi-automatic way. Finally, the last chapter presents the results of the tests, including the TDC preliminary characterisation, some dark count rate (DCR) measurements, photon count results under laser illumination, and, at last, an first promising PDE (Photon Detection Efficiency) evaluation.

1.1 Single-Photon Detection

1.1.1 Light and photon detectors formalism

Two important parameters that characterize the light are the wavelength λ and the frequency ν. They are linked together by a universal constant, the light propagation

speed in vacuum c:

$$c = \lambda\nu = 2.998 \times 10^8 \,\mathrm{m/s} \tag{1.1}$$

At the beginning of the last century, Einstein brilliantly demonstrated that light is composed of discrete energy packets and that it can also behave like a flux of particles, as postulated by Planck few years before [31]. Energy and momentum of these electromagnetic energy quanta called photons depend on the wave frequency, as given in Eq. 1.2, 1.3

$$E_p = h\nu = \frac{hc}{\lambda} \tag{1.2}$$

$$P_p = \frac{E_p}{c} = \frac{h}{\lambda} \tag{1.3}$$

where $h = 6.63 \times 10^{-34} \,\mathrm{m^2 kg/s}$ is the Planck universal constant.
The key contribution of Einstein was to explain the photoelectric effect, i.e. electrons emission from a metal surface when irradiated by light. This same effect occurs also in semiconductors and it is the underlying principle for photon detection. In a semiconductor, electrons occupy almost completely the energy levels in the valence band whereas the conduction band is practically empty. When a photon, with an energy greater than the energy gap E_g between the bands, hits the material, it transfers its energy to an electron that will be excited from valence to conduction band. If $h\nu$ is greater than E_g the excess energy is observed as kinetic energy of the electron that can conduct current. The Einstein relation between these quantities for a semiconductor is given in Eq.1.4.

$$K_e = h\nu - E_g \tag{1.4}$$

where K_e is the electron kinetic energy. The energy gap defines a threshold frequency for the incoming photons below which no electrons are excited to the conduction band: each material absorbs photons only above a given frequency. On the other hand, light intensity only determines the rate of photoelectric emission. These considerations are necessary to understand the underlying physics but a simpler, macroscopic model of light-matter interaction is sufficient to discuss their main characteristics.
Consider an incident light ray of definite power P_0 hitting a surface or, more in general, a strong discontinuity of the refraction index. This discontinuity results in a high reflection coefficient R defined as the ratio between the reflected power P_r and the incident one:

$$R = \frac{P_r}{P_0} \tag{1.5}$$

Only part of the incident power is transmitted, namely equal to

$$P_{T0} = (1 - R)P_0 \tag{1.6}$$

Inside the material, photons are gradually absorbed so that the optical power decreases following an exponential trend. The transmitted power at position x follows from Eq 1.7

$$P_T(x) = P_{T0} \cdot e^{-\alpha x} = (1 - R)P_0 \cdot e^{-\alpha x} \tag{1.7}$$

where α is the *optical absorption coefficient* describing the exponential decrease of optical power: it expresses the probability for a photon to be absorbed per centimeter of propagation in the material. Its inverse is the *optical absorption depth*, L_a. A visual representation is shown on the graph in Fig. 1.1

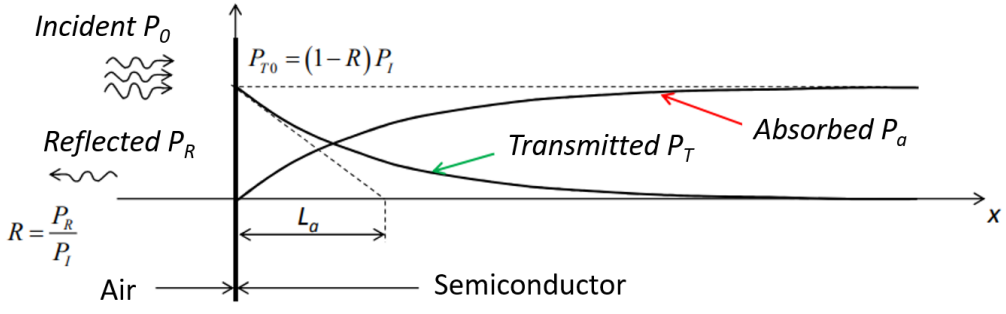


Figure 1.1: Reflection, transmission and absorption of optical power in presence of a strong discontinuity of refraction index (e.g. air-semiconductor). From [32]

.

Many photodetectors exploit semiconductors to detect photons, generating electron/hole pairs by photoelectric effect. The charge carriers generation rate at a given distance is computed from Eq. 1.7: the number of photons at distance x must be multiplied by the probability that they are absorbed given by α.

$$g(x) = \frac{P_T(x)}{h\nu} \cdot \alpha = \frac{P_0}{h\nu} \cdot (1 - R) \cdot e^{-\alpha x} \cdot \alpha \tag{1.8}$$

Quantum photodetectors properties

Quantum photodetectors, or simply *photon detectors*, are a class of detectors employing photoelectric effect to detect light. The generated charge carriers can produce a current under the correct bias. This transduction from optical to electrical signal can be characterized with few parameters useful to define and compare the performance of this class of detectors.

- **Quantum detection efficiency, η**
  Also called *photon detection efficiency* or just *quantum efficiency*, it describes the ability of the detector to convert photons in electron/hole pairs [33]. It is defined as the ratio between the photo-generated electrons (or holes) and the number of photons incident on the detector [34]:

$$\eta = \frac{\text{number of photo-generated electrons}}{\text{number of photons on the detector}} = \frac{N_e}{N_{ph}} \tag{1.9}$$

  From the definition we can deduce that photon reflection on surface, electron scattering losses and recombination reduce the quantum efficiency. By contrast this parameter does not take into account any internal gain considering only the primary electrons/holes pairs.

- **Responsivity, R**
  Also called *radiant sensitivity*, it focuses on the transduction from optical power to electrical current [35]:

$$R = \frac{\text{output current}}{\text{optical power on the detector}} = \frac{I_S}{P_S} \ \ [A/W] \tag{1.10}$$

  Eq. 1.9 can be rewritten as

$$\eta = \frac{I_S/q}{P_s/h\nu} \tag{1.11}$$

  with q the electron charge. From this and Eq. 1.1 and 1.10 follows

$$R = \eta\frac{q}{h\nu} = \eta\frac{q}{h}\frac{\lambda}{c} \tag{1.12}$$

- **Signal-to-noise ratio and Noise Equivalent Power, NEP**
  Every measurement is affected by noise entailing a defined uncertainty. This noise is evaluated, for example at the output of the detector, as a current fluctuation I_N from the mean value I_S. It is necessary to link this noise with the input signal determining their ratio as in Eq 1.13. The signal-to-noise ratio allows us to assess the ability of the detector to discriminate between the signal and noise. In general a signal is detectable if this ratio is greater than one.

$$\frac{S}{N} = \frac{I_S}{I_N} \tag{1.13}$$

  It follows that the Noise Equivalent Power (NEP) is the input optical power that produce a signal-to-noise ratio of one and represents the smallest detectable signal [36]. For an input power $P_S = NEP$ the output signal from Eq. 1.10 is

$$I_S = R \cdot NEP \tag{1.14}$$

By definition if the signal-to-noise ratio is one, $I_N = I_S$, then

$$NEP = \frac{I_N}{R} \tag{1.15}$$

Photon shot noise

Shot noise results from the discretisation/quantisation of physical phenomena and their random fluctuations considered in terms of number of events. Photon shot noise is an important parameter to assess because it has relevant consequences on the performances of photodetectors, for example limiting their detection. A complete discussion can be found in [37] and [38]
Optical radiation is formed by photons arriving randomly in time so that their number in a given time interval is a statistical variable. Consider the optical power in terms of average number of photons $\langle n \rangle$ arriving in a given interval τ:

$$P = \frac{h\nu}{\tau}\langle n \rangle \tag{1.16}$$

The associated rms noise can be expressed by its variance as

$$\langle \Delta P_{shot}^2 \rangle = \left(\frac{h\nu}{\tau}\right)^2 \cdot \langle \Delta n^2 \rangle \tag{1.17}$$

Considering photons as independent events, their statistics is usually well approximated by the Poisson distribution so that the variance is given by

$$\langle \Delta n^2 \rangle = \langle n \rangle \tag{1.18}$$

Replacing in Eq.1.17 we obtain

$$\langle \Delta P_{shot}^2 \rangle = \left(\frac{h\nu}{\tau}\right) \cdot P = 2 \cdot h\nu \cdot P \cdot \Delta f \tag{1.19}$$

where Δf denotes, for convenience, only the positive frequencies of the noise bandwidth. The factor 2 is added to take into account also the negative ones. Dividing this value for the bandwidth we get the unilateral noise spectral density:

$$S_p = 2h\nu P = 2\frac{hc}{\lambda}P \tag{1.20}$$

There is no dependency on the noise frequency, i.e. shot noise is *white noise*: its spectrum is essentially flat over a large frequency range. On the other hand, for a given optical power P, the photon shot noise is smaller for photons at lower frequency. At infrared wavelength for example, other processes and noise sources

become dominant. We can intuitively understand this dependency considering that if the wavelength decreases, each photon will have lower energy. This means that in this case, for a given power, the signal will contain more photons thus reducing the Signal-to-Noise ratio. The effect is the same we observe when integrating the signal longer to increase the number of events.
Shot noise is not limited to photons but is common to every discrete quantum phenomenon like current itself but the computation are similar anyway. In any case this intrinsic noise defines the lowest limit of detection of every photon detector.

1.1.2 Single-photon detectors: state of the art

Photomultiplier tube

Photomultiplier tubes (PMT) are one of the first established technologies for single-photon detection and they have been employed for many years in several applications ranging from biology, astronomy and nuclear physics [39], [3].
They derive from vacuum tubes (or photo-tubes) and are based on the same principle of light detection by photoelectric effect on a metallic cathode. The great difference is that an amplification effect is obtained between the cathode and the anode by an electron multiplication process. This amplification of the detector signal overcomes the electronic circuit noise that is a major limitation in vacuum tubes.
The process occurs by secondary electron emission in vacuum and current amplification in a dynode chain inside the tube: the photogenerated electron (primary electron) is emitted from the cathode with low kinetic energy ($E_k < 1\,\text{eV}$) to be attracted by a high potential difference (hundreds of volts) towards an electrode (dynode). By penetrating in the material, the electron loses its kinetic energy transferring it to the electrons of the dynode and some of them can be emitted in vacuum (secondary electrons). The yield of secondary electrons per primary ones is indicated by the parameter $\delta_i > 1$, also called secondary emission ratio. The same process occurs on every dynode leading to a final huge current amplification whose overall gain is given by the product of the single yields:

$$G = \delta_1 \cdot \delta_2 \cdot ... \cdot \delta_N = \delta^N \textit{ (with N equal dynodes)} \tag{1.21}$$

Figure 1.2 shows a sketch of a PMT working principle.

To understand the emission characteristic of the electrodes and its primary energy dependency, it is necessary to analyze the the interaction process of the electrons in the material. When the electrons penetrate in the dynode, they transfer their kinetic energy to generate secondary electrons; the higher the energy, the more secondary electrons are produced. On the other hand, primary electrons with

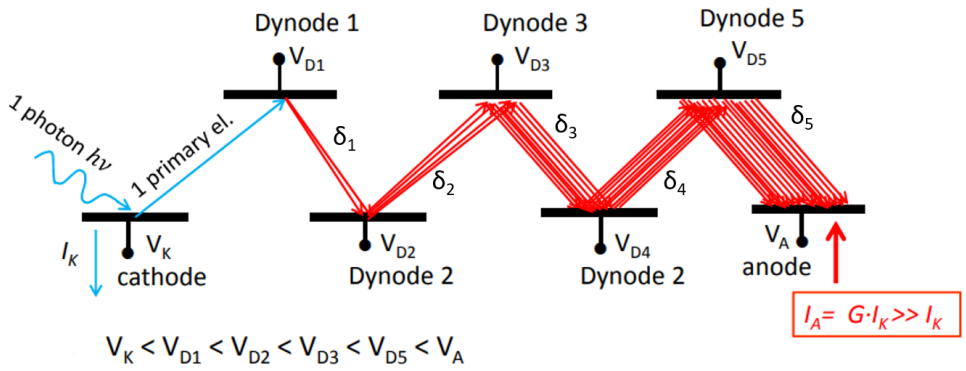


Figure 1.2: Sketch of the dynodes' working principle inside a PMT, from [32]

higher energy penetrate deeper in the material. In this case the generated particles have to go through a longer path before reaching the surface. Part of them can be absorbed, others lose energy through collisions and fall below the vacuum level. All these mechanisms represent losses that decrease the secondary emission yield [40]. Therefore an optimal voltage value exists as clearly shown in the graph of Fig. 1.3

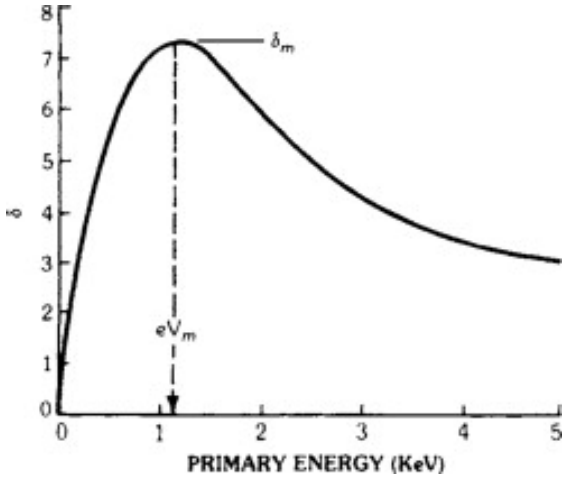


Figure 1.3: Secondary emission yield curve vs primary energy [40]

The output of a PMT can be imagined as a superposition of single pulses corresponding to single electron emissions from the photo-cathode, also called (Single Electron Response, SER). These pulses under weak illumination can easily be separately observed allowing us to detect single photons on the cathode. One way to reduce the effect of noise sources is to define a range of amplitude in which the pulses should fall to be considered valid. The lower limit cuts off all the pulses

due to electrons' thermal generation in intermediate dynodes; the upper one rejects the pulses due to high energy events like cosmic rays.
Unfortunately, the basic assumption that all the valid pulses have a characteristic or at least similar height, is eventually not true. The amplification gain G is not constant at all because it depends on the secondary electron emission, a process of statistical nature with random variations itself. The statistical distribution of this gain is an issue because it also introduces an additional noise factor, even if it is usually negligible for high-quality PMT.
Main limitations of these devices are linked to its rather fragile structure that limits the practical use like the vacuum-requirement inside the tube and the size of the detector itself, especially of the dynode chain. Another drawback is the high DC voltage (several hundreds volts) required to bias the dynodes [3], [41].
On the other hand, PMTs performances are enhanced thanks to their internal gain that makes external noise sources negligible with respect to the one associated with the quantized nature of light. Another remarkable feature is the very low dark current due to thermionic emission from the cathode and its shot noise. This is why, although PMTs remain fragile detectors not suitable for many applications in harsh environment like space, they still represent a good choice for single-photon detection [3].

Micro-channel plate

Micro-Channel Plates represent an evolution in the structure of a PMT to make it simpler, more compact and robust [42]. The idea behind is to reduce the complexity of the device by replacing the series of multiplying dynodes with a thin glass tube with micro-metric diameter. Thousands of these structures are embedded together to form a matrix in a MCP as shown in Fig. 1.4. The inner walls of the channels are treated and converted to semiconductor able to emit secondary electrons. A high voltage is applied in parallel to the plate via metal electrodes on the two opposite faces. The final gain depends on the number of impacts along the tube so it is strictly related to the aspect-ratio L/D: for a high ratio the number of impacts increases but the yield drop because electrons energy becomes lower. The best aspect-ratio for the maximum gain is usually around $L/D \approx 50$.

Photo-electrons can be focused on the plate simply by applying a high voltage to the multiplier input, then no other focusing is required during amplification inside the channels. At the other end of the plate, electrons can be collected either by anode or by a phosphorescent screen. In the latter case the structure can be coupled with other detectors, for example a CCD or CMOS camera, capturing the image on the screen through a system of lenses or fibre optics. The MCP would work as an intensifier allowing detection in low light intensity conditions.

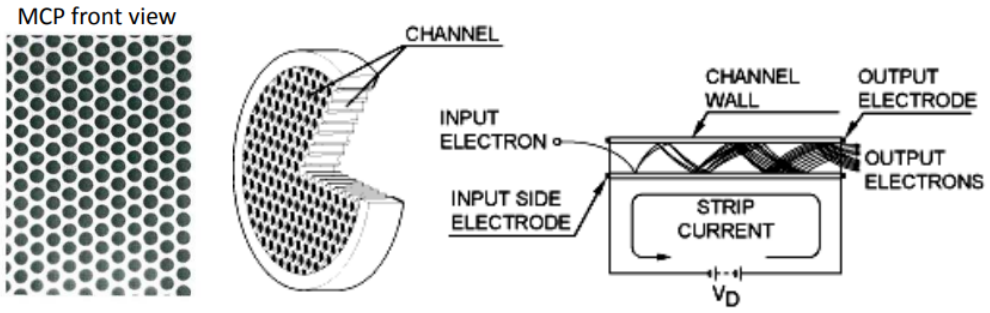


Figure 1.4: Schematic view of an MCP multiplier. Adapted from [2]

1.2 Single-Photon Avalanche Diodes (SPAD)

Single-Photon Avalanche Diodes are solid-state detectors capable of photon counting measurement with considerable time resolution [43]. This property makes them attractive for several applications, especially for time-of-flight measurements and in the field of bio-imaging, as discussed later in section 1.5. The structure of these devices is the same of an avalanche photo-diode (APD) generally based on p^+-n/p^+-i-n junction under reverse bias.
In a diode in blocking mode, under no illumination the current is given by the thermal generation of carriers, mainly in the depletion region. Detection of light occurs in these structure in a two-step process: absorption of light by the semiconductor with the resulting photo-generation of electron-hole pairs; collection of these optical generated charge carriers by means of the electric field in the junction. The generation rate can be expressed again as in Eq.1.8.
The charge carrier generation and collection can occur in three different cases as shown schematically in Fig.1.5. In the ideal case, photon is absorbed in the depletion region and carriers are extracted directly by the electric field in that region: electrons drift towards the n-region, holes towards the p-region. But photons can also be absorbed in one of the two diffusion region at the sides of the junction. In these cases the minority carrier of the photo-generated pair diffuse towards the junction. Once it reaches the depletion region, it can be extracted by the electric field. One first problem in this scenario is that, during diffusion, the carriers can recombine in the neutral region without being collected. Furthermore, the diffusion process takes a relatively long time introducing delay on the current response, so it is desirable to have absorption in the depletion region for fast photo-diodes. For this reason the structure of the junction can be modified to increase the depletion region for example introducing an intrinsic (or very low-doped) layer between the two doped ones.

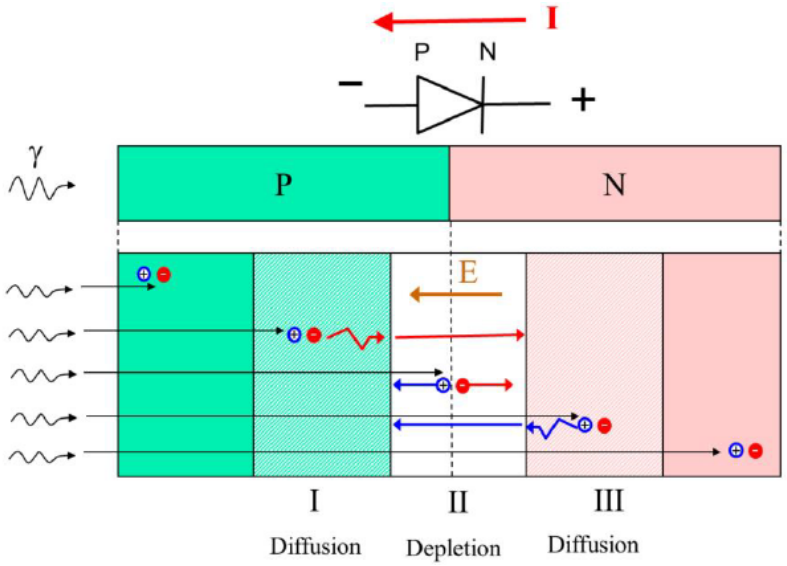


Figure 1.5: Schematic view of a pn junction in blocking mode highlighting the three cases of photon absorption. The ideal one is when absorption occurs in the depletion region and carriers are extracted directly by the electric field. Absorption in the diffusion region results in a slower response of the detector

Avalanche effect

When a strong reverse voltage is applied on the diode, the electric field in the depletion region increases and the carriers are strongly accelerated. Drifting in the electric field, electrons and holes gain kinetic energy that is in part transmitted to the lattice upon scattering events. If the electric field is strong enough, the kinetic energy might be sufficient to generate ionizing collision events with the lattice, resulting in additional electron/hole pair [34]. These carriers are accelerated themselves and can undergo the same scattering events. The cascade of these collision generates an avalanche effect. The process is bidirectional because also the holes are accelerated (in the opposite direction), producing an additional contribution to the current (Fig.1.6). This results in a positive feedback loop when holes generate more electrons by impact ionisation and vice-versa in a self sustaining process [3]. Two parameters called ionisation coefficients, α for electrons and β for holes, define the probability of carriers to produce impact ionisation.
The avalanche of n and p carriers create a dipole-like mobile space charge that generates an electrical field opposite to the drift one across the junction. Its effect is to limit the cascading process and avoid the divergence of the current. The usual bias, just below the breakdown voltage, produces, indeed, a linear current amplification.

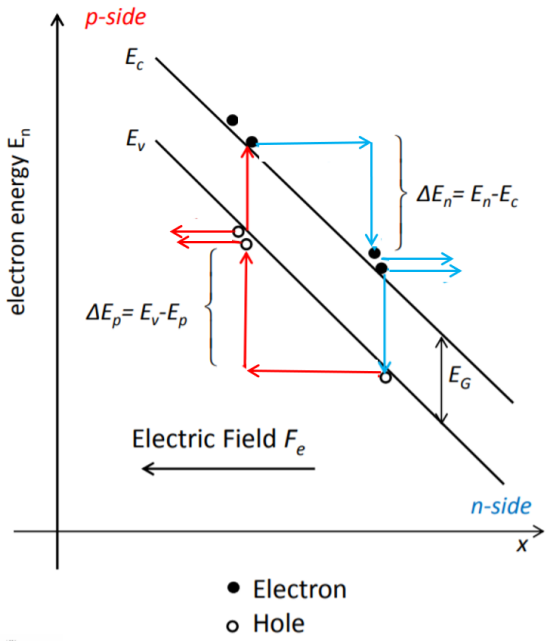


Figure 1.6: Schematic representation of an avalanche process in the depleted region of a p-n junction [32]

SPAD working principle

SPADs are essentially avalanche PDs working under strong reverse bias, above the breakdown voltage V_{bd}, exploiting the huge current amplification in this region. Under these conditions, even a single photon hitting the photo-sensitive area can trigger an avalanche in the device from the photo-generated primary electron-hole pair. The huge gain allows a macroscopic current to flow in the device with a very short rise time (the avalanche build-up time is usually of few picoseconds [5, 43]). This operation regime is known as Geiger mode, recalling the similar pulsed output of the more famous radiation detector.
The avalanche process must be controlled and stopped with a quenching mechanism to prevent the high current flow from destroying the diode itself. This is obtaining lowering the bias voltage V_{op} of the SPAD and leaving the avalanche to run out for example connecting a resistor in series with the device [9]. After the quenching the voltage must be restored to the initial value before the next photon detection is possible [44]. During this period, called *dead time*, the SPAD is almost insensitive but it gradually regains its sensitivity until its nominal value when the recharge is complete. Fig. 1.7 shows on the I-V curve the three step performed in the Geiger mode. A SPAD sensor operating in this region works basically as a photon-triggered switch, resulting in a binary output where each pulse correspond to

photon detection. The binary output is eventually independent on the number of absorbed photons, so the magnitude of the instantaneous flux is not available.

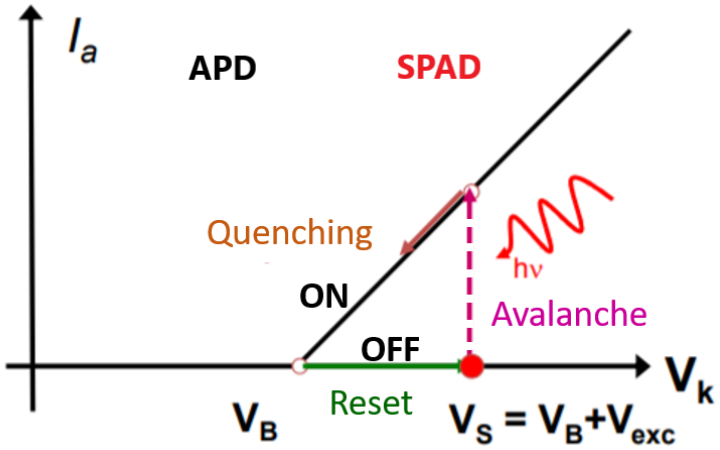


Figure 1.7: SPAD operation in the reverse diode I–V characteristics. The APD regime, or linear-mode operation, is right before the breakdown voltage. Above it the diode enters the Geiger mode, i.e. the SPAD operation. Adapted from [32]

The simplest quenching mechanism consist of a resistor in series with the SPAD. The problem with this passive circuit is that the recharge process takes some time to recover from the breakdown to the bias voltage. This dead time causes several count losses because the detection efficiency must be gradual recovered itself. These reasons make a passive quenching circuit not very suitable for fast photon counting applications, where an active quenching is usually preferred [44].

In general SPADs' performances can be assessed with several parameters. One of the most significant is the Photon Detection Probability (PDP), representing the probability to trigger an avalanche due to the absorption of a photon at a defined wavelength [9]. Current CMOS SPAD technology present a peak PDP in the visible range that can reach up to 70% for isolated, optimized devices [43]. Another important figure of merit is the Dark Count Rate (DCR), i.e. the detection rate in absence of light. This parameter represents the most relevant source of noise in this kind of devices and can be caused by several processes. For example, thermal generation of carriers in the depletion region is a major contribution. Carriers generation can also be greatly favoured by the strong electric field across the junction (field-enhanced generation) and the resulting relevant band bending. Electrons (and holes vice-versa) can reach directly the conduction band by tunnel effect (Band-to-Band Tunneling, BTBT) or by Trap-Assisted Tunneling (TAT) processes [45].

Afterpulsing is another parameter affecting the performance of the device. It introduces false detection events correlated in time with previous one. The explanation is that, during the avalanche, some carriers can be captured by some traps, i.e. defects in the silicon lattice structure. These carriers can be released with a delay of up to several ns and re-trigger the avalanche, generating an additional, false output pulse.
The ratio between the photo-sensitive area and the total area of the device is the fill factor, another significant figure of merit. When multiplied for the PDP it gives the overall Photon Detection Efficiency (PDE).
The deviation in time of rising edges of the SPADs' fast pulses defines the so-called timing jitter. This represent generally the timing resolution of the detector and, for SPADs, it generally ranges between tens and hundreds of ps [46]. Table 3.1 summarizes the main figures of merit of SPAD with the corresponding usual values in recent CMOS technology (130 nm and below), taken from literature [43], [21].

SPAD pixel parameter	Value range
Dead time [ns]	10-100
DCR [cps/µm^2]	0.3-100
PDP (peak) [%]	10-50
Spectral range (PDP > 1%) [nm]	350–1000
Fill factor [%]	1-60
Timing resolution [ps]	30-100
Afterpulsing probability [%]	0.1-10

Table 1.1: Summary of the main SPAD parametrs and their standard values as in [21] and [43]

.

1.3 SPAD arrays

Single SPAD devices can be integrated in arrays of various dimension to create imagers. Depending on the desired dimensionality and target application, several architecture are possible.
The simplest solution is a linear 1D array of pixels. In this case, the requested electronics can be placed outside the detector and the different pixels are separated only by the diode guard ring (Fig. 1.8a). This type of array allows a truly parallel

operation of the pixels while maintaining the highest possible fill factor (only the detectors are present on the imager area). In case the target application requires a 2D image, an optical or mechanical scanning mechanism is mandatory.
This limitation can be overcome by two-dimensional SPAD arrays that require, by contrast, more complex designs (Fig. 3.14b). In general the presence of the read-out electronics significantly decreases the fill factor of the detector so that microlenses are often employed to focus the incoming light on the photo-sensitive area of the array. This type of imagers usually present some sort of resource sharing among multiple pixels that prevent a fully parallel operation of the detector (Fig. 1.8c).

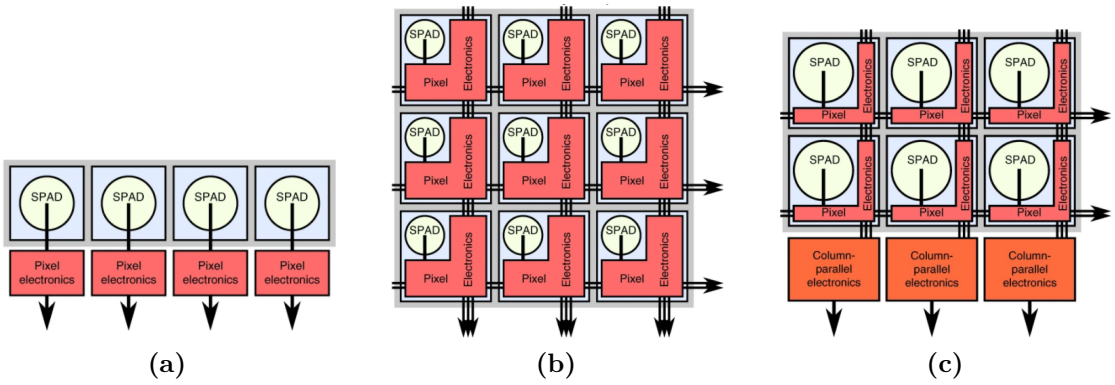


Figure 1.8: (a): 1D array with electronics outside the pixel area; (b): 2D array with electronics inside the pixel itself; (c): 2D array with electronics at the array periphery, in this case column-based. Taken from [43]

1.3.1 Analog SiPM

Analog Silicon Photo-Multiplier are large area photo-detectors implementing a dense array of independent SPAD sensors with their own quenching resistor usually called *microcells* [47], [48], [15], [49]. All the microcells detect photons independently and are connected in parallel to a node so that the sum and combination of the photo-currents create an analog output, carrying information also on the magnitude on the instantaneous photon flux. In particular, the amplitude of the output pulse is proportional to the number of firing microcells, thus overcoming the limitation of single SPAD detectors. Fig. 1.9 illustrates the simplified electric circuit of an analog SiPM architecture and shows the schematic structure of a microcell.
Some SiPMs present an additional terminal for a fast output signal [49] as shown in Fig. 1.9. This capacitively coupled output allows ultra-fast timing measurements

and can be beneficial in general for its lower capacitance.

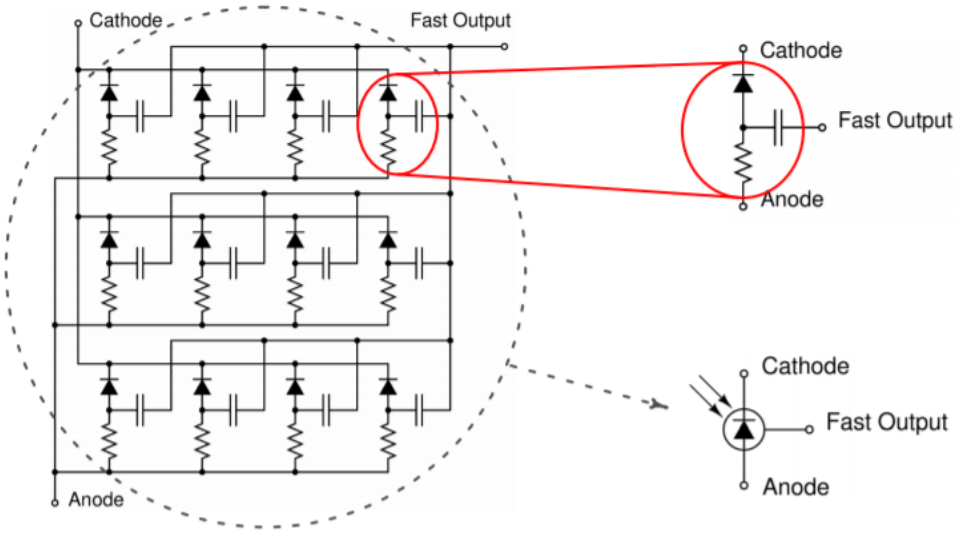


Figure 1.9: Simplified circuit schematic of an analog SiPM composed of an array of microcells (SPAD, resistor and output capacitor) as found in [49]

A precise description and discussion on the main parameters of these sensors like breakdown voltage, gain, PDE, dark count rate (noise), optical crosstalk, afterpulsing and temperature dependency can be found in [15] and [49].
Several advantages, such as very low bias voltage (especially if compared to PMT), high gain and PDE combined with compactness, robustness and magnetic insensitivity, make these devices the ideal choice for photon detection when fast timing is required. These properties, together with ongoing research to improve photon time resolution, will help SiPM to become a valid alternative to PMT in most applications [15].
Another architecture however might be the device of the future, with its ability to process each photon independently and potentially resolving the position of each impinging particle with a camera-like operation: the digital SiPM.

1.3.2 Digital SiPM

In fully-digital or multi-digital SiPM, every SPAD pixel is connected to its own front-end electronics, allowing a completely parallel operation of the sensor. The microcell is in this case the source of the binary digital signal from the SPAD output. In fact the information for each single microcell consists of one bit: a photon dectection correspond to "one" whereas no detection correspond to "zero".

One and zero signals are well separated thanks to the high gain of the detector and can be well resolved by the following electronics. The signal can then be directly be processed by the readout circuitry. Fig. 1.10 presents a sketch of the detector's structure.

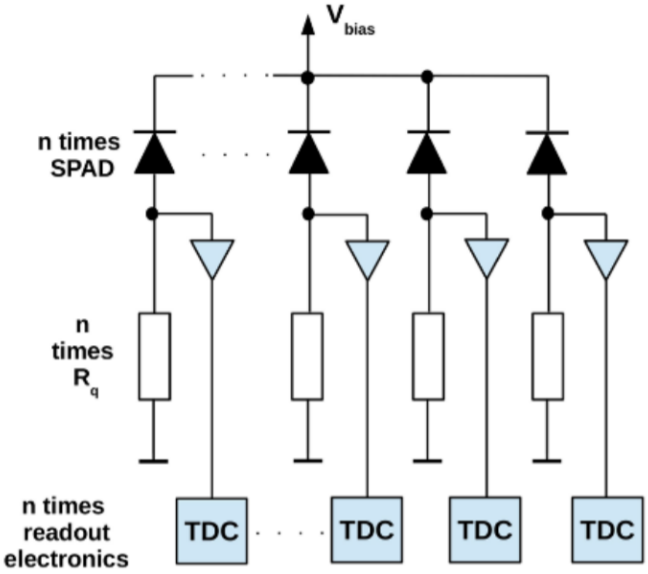


Figure 1.10: Simplified sketch of a multi-digital SiPM composed of an array of microcells, each one with its readout electronics in this case represented by buffers and Time-to-Digital Converters (TDC) as found in [50]

First commercial digital SiPM were released by Philips for Time-of-Flight PET (ToF-PET) [15]. The field of medical imaging is one of the most relevant for these devices employed for photon counting and ToF measurements [50].
One great advantage with respect to analog SiPM is that digital ones do not use any analog signal processing, resulting in faster and more accurate information. Another benefit of the digital approach is the possibility to greatly reduce the DCR by masking some particularly noisy microcells. This is possible only with the direct access to the single pixels offered in this detectors. One of the main characteristic features of the digital SiPM is its ability to provide the spatial distribution of the detected photons, working as an imager.
The current trend for SiPM is towards a 3D integration in CMOS technology. As briefly discussed in the following section, this would allow a local, on-chip, fast data processing while improving the detection efficiency and the fill factor, one of the main limitation of planar devices [15], [50].

Comparison criteria	Analog-SiPM	Digital-SiPM
Output	Analog signal (must be amplified)	Digital signal
Electronic noise	Significant	Irrelevant to resolve single photon
Noisy pixels	Problematic	Can be turned-off one-by-one with a proper masking circuit
Quenching and recharge	No control	Can be tuned
Afterpulsing and crosstalk	No control	Can be limited with strategies limiting the avalanche
Static power consumption	From auxiliary and readout circuits	From auxiliary circuits and leakage currents

Table 1.2: Comparison between analog- and digital-SiPM, with some significant criteria[1]. More information can be found in [15, 30]

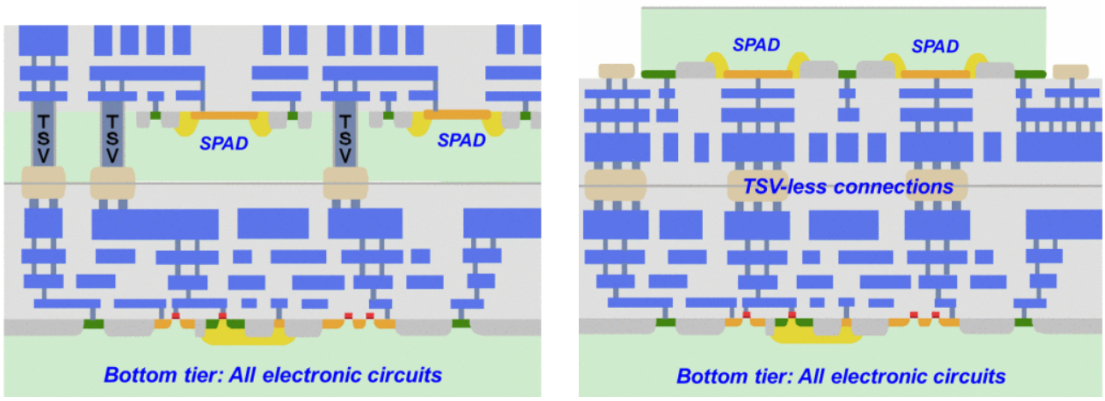
.

1.4 3D Architecture for Digital SiPM

Combined integration of readout electronics with the sensitive SPAD devices in one wafer limits the flexibility of the CMOS processes, and usually results in more noisy pixels than in the analog counterpart. The limited geometrical factor is another drawback of the presence of both electronics and sensor on the same tier.
For these reasons, the research on digital SiPM is now focusing towards a 3D assembling of the detector, supported by the development in semiconductor technology proposed in few standard CMOS facilities [15, 50]. The structure usually includes two-three different dies, or tiers, stacked one onto the other to separate the sensitive area with the SPAD from the processing circuitry [21]. However,

[1]It is hard to compare the power consumption of such different systems. For the analog-SiPM, several components like the ADC are continuously consuming power, whether there are events or not. In digital SiPM based on CMOS readout circuits, in absence of events, the power consumption can be very low and it only becomes significant when a SPAD is triggered. The idle power consumption is set by auxiliary circuits and transistors' leakage currents, which can be considerable [30].

sensor arrays always occupy completely the topmost one improving dramatically the detection efficiency and the fill factor. Through-Silicon-Vias (TSVs) connect each microcell to the second tier, dedicated to the quenching circuit and to the processing electronics. Several structures were proposed in the last years, each developing advanced on-chip functionalities depending on the target application, e.g. in [8, 14, 20, 22, 23, 24, 25, 26, 27].



(a) **(b)**

Figure 1.11: Cross section of (a) Front-side illuminated and (b) Back-illuminated 3D-stacked SPAD sensor, from [19]

A fundamental advantage of this structures is that each chip can be optimized independently with distinct processes, e.g. CMOS image sensor (CIS) process for the top tier and a state-of-the-art, low power, standard CMOS process for the bottom one. It also allows the assembly of large SPAD cameras with accurate timing, thanks to the low skew even across large chips. Additional benefits lay in the possibility to implement advanced on-chip functionalities like histograms[51], neural networks and deep learning engines [19].
Apart from performance enhancement, this approach can open new opportunities in the NIR and IR spectrum by using different semiconductor, like InGaAs for the top sensitive layer. Table 1.3 summarises the main differences in this approach.
Fig. 1.11 shows the cross section of two alternative structures for of a 3D-stacked SPAD image sensor. Both present a top sensitive array on a CMOS chip but some differences are evident. Fig. 1.11a represent a Front-side illuminated (FSI) configuration. In this case each device must be connected vertically to the underlying circuit through all the top tier with TSV, requiring a delicate additional process. On the other hand, the Back-side illuminated (BSI) configuration depicted in Fig. 1.11b shows the two tiers face-to-face, with the SPADs directly connected to the

bottom electronics. Unfortunately, this solution requires a backside thinning of the wafer to ensure a sufficient PDP.
In these two cases the junction is at different depth from the surface of the image sensor, resulting in distinct absorption spectra: FSI 3D image sensors perform best at near-UV and blue wavelength while BSI ones are more suitable for IR and near-IR applications [21].

	2D-SiPM	3D-SiPM
Fill factor	Trade-off with electronics and logic functions	Optimal
On-chip functionalities	Limited, trade-off with photosensitive area	Higher design freedom and customisation
Reading node capacitance	Minimal	Higher, due to the vertical connections
Technology	Same for SPADs & electronics (no optimization)	Can be chosen to optimize each tier
Process/cost	Standard process, leading to low cost and fast prototyping	Different processes, higher cost and time.
Material	Silicon	Other materials can be used for the top layer (InGaAs/InP and Ge [15, 52, 53])

Table 1.3: Comparison between analog- and digital-SiPM, with some significant criteria. More information and state-of-the-art examples can be found in [15, 19, 30]

.

1.5 PET and applications

SPAD detectors are attractive in general for applications requiring fast timing resolution and potentially single photon sensitivity such as Positron Emission Tomography (PET) [54], Fluorescence Lifetime Imaging Microscopy (FLIM) [55], Raman Spectroscopy [56], Near-infra-red Optical Tomography (NIROT) [57] and Light Detection and Ranging systems (LiDAR) [26]. Our main focus will be on the

first one, the target application of the chip analyzed in this work.
PET is a nuclear bio-imaging technique first introduced in the early 1970s [58]. It has a whole set of different clinical and research applications, with the main goal to visualise and characterise biological processes. In particular, it is extensively used for diagnosis and monitoring of cancer [59], [60]. This highly sensitive imaging technique allows three dimensional mapping of metabolic activity of the target tissue, measuring the distribution of a radio-labeled pharmaceuticals previously administered. The most common radio-tracer is ^{18}F, a positron emitting isotope of fluorine, usually inserted in an analogue of glucose, the fluorodeoxiglucose ([^{18}F]FDG). This molecule represents the standard for PET radio-pharmaceuticals and helps study the glucose metabolism, particularly enhanced in cancer cells. FDG accumulation in tissues is in fact related to increased consumption of glucose, characteristic of most cancers. The radiation emission will then be concentrated in these critical regions that can be easily detected and studied.
The isotope decay process originates from a β^+ decay, i.e. the conversion of a proton in a neutron resulting in the emission of a positron (the anti-particle of the electron) and a neutrino as expressed by the nuclear reaction 1.22.

$$\mathrm{p} \longrightarrow \mathrm{n} + \beta^+ + \nu_e \tag{1.22}$$

The emitted positron (β^+) interacts almost immediately with an electron (its antiparticle) in an annihilation event producing two energetic photons (or gamma rays) of 511 keV, simultaneously emitted. Energy and momentum conservation implies that these two photon are emitted in (almost) opposite direction and are detected by coincidence detectors of the PET scanner. When two detections occur within the coincidence window (of few ns generally), a coincidence event is recorded. The positron annihilation is then assumed to have occurred along the so-called line of response (LOR), i.e. the virtual line connecting the two triggered detectors. The collection of multiple LOR allows a reconstruction of the emitting tissue in the body, as sketched in Fig.1.12.

A good spatial resolution is a key feature for PET scanners. Unfortunately it can be limited by several factors. The first is the small distance the positron travels before annihilating, of the order of few mm for ^{18}F [61]. This results in a shift of the reconstructed LOR from the true radioactive source. A second factor to consider is that the positron can have a small residual momentum at the time of annihilation. As a consequence, the two gamma rays are not emitted exactly at 180°.
A significant improvement in the tissue reconstruction can be achieved with the extraction of time-of-flight (ToF) information from the detected photons. More precisely, the location of the positron can be assessed with greater accuracy measuring the difference in the arrival time between the two annihilation gammas. The ToF

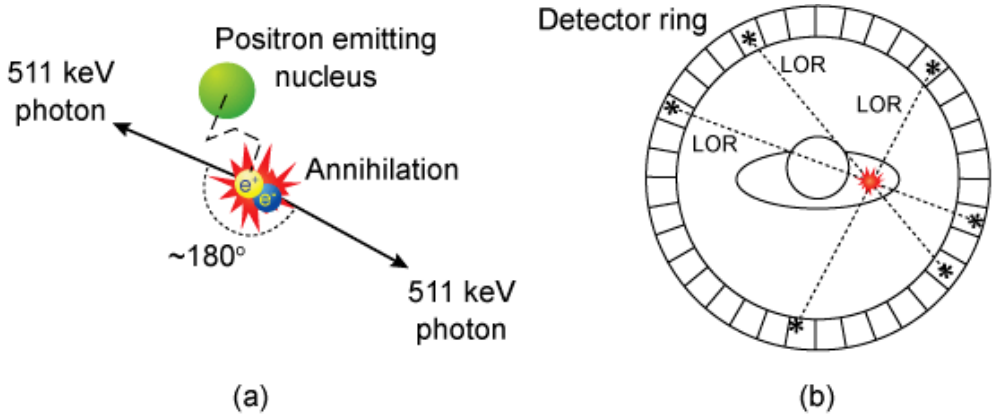


Figure 1.12: Principle of PET: (a) a positron annihilates close to the originating nucleus with an electron and two 511 keV gamma rays are emitted in (almost) opposite directions; (b) if two events are detected within the coincidence window by the ring of detectors surrounding the patient, a coincidence event is recorded. The tomographic reconstruction is performed recording many LORs (taken from [61])

.

information provides a measurement of the annihilation point with an accuracy given in Eq. 1.23. Without ToF the annihilation point is assumed to be uniformly distributed along the line of response, as sketched in Fig. 1.13.

$$\Delta x = \frac{c}{2}\Delta t \tag{1.23}$$

where Δx and Δt are respectively the errors in the position along the LOR and in the timing measurement [62]. To get sub-centimeter precision, the required timing precision should be less than 50 ps. This explains the importance of having very performing detectors with high timing resolution.
Besides the related improvement in spatial resolution, ToF-PET can also significantly reduce the statistical noise and improve the SNR [62, 63] as showed by Eq.1.24

$$SNR_{gain} \propto \sqrt{\frac{2D}{c\Delta t}} \tag{1.24}$$

where D is the size of the emission source. The difference in the arrival times corresponds to a position on the LOR, with an uncertainty linked to the measurement errors.

Typical detectors used in PET scanners are made of high density, inorganic, scintillating crystal arrays, instrumented with PMTs or SiPM for higher granularity

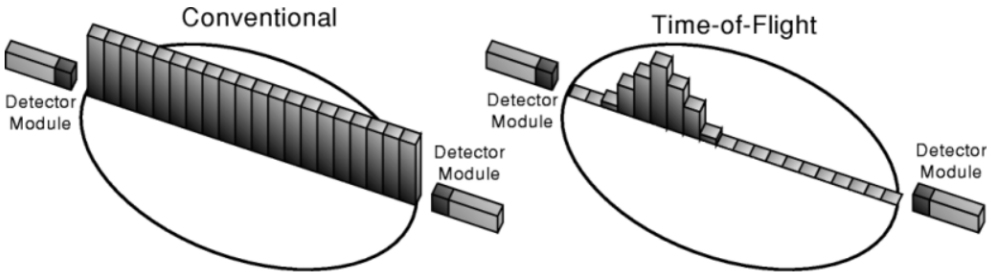


Figure 1.13: Principle of ToF reconstruction. Left: conventional reconstruction, all the pixel on the chord (LOR) are increased of the same amount. Right: ToF reconstruction, the increment depends on the the probability that the source is on that pixel. From [62]

.

and better timing performance. Two of the most used crystals are the LYSO (Lutetium-Yttrium Oxyorthosilicate) and BGO (Bismuth Germanium oxyde). The first presents major advantages like a higher light output, better energy resolution and fast scintillation that can reduce the detector dead time [58, 64]. These properties made it the reference scintillator for TOF-PET [65]. On the other hand, BGO received renewed interest lately, thanks to the possibility to detect Cherenkov radiation [66].

Chapter 2

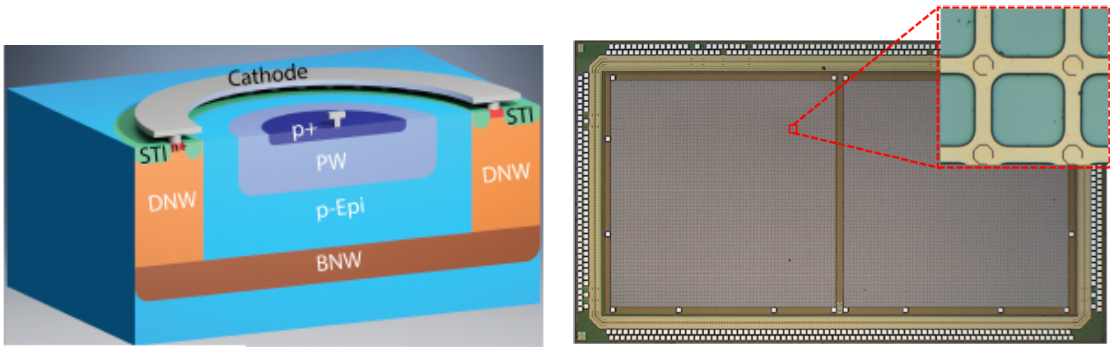
Blueberry chip architecture

In the last years, time-of-flight positron emission tomography (ToF-PET) demanded and pushed the development and design of innovative detectors [14]. In this context, the design of multi-channel digital silicon photomultipliers (MD-SiPM) allows an increased timestamp granularity and data pre-processing on chip [17]. The implementation in CMOS technology allows cost reduction and the direct integration of circuits on chip SoC (System on Chip) implementation. On the other hand, high-performance CMOS devices require advanced, nanometric technology nodes, not well suited for efficient SPADs. 3D-integrated sensors help to overcome this problem, because we can select the proper technology node for each tier [14, 18].
In this chapter, we briefly describe the architecture of Blueberry, a 3D-stacked frontside-illuminated (FSI) MD-SiPM, as in [14, 18] and first proposed in [20]. The chip is compatible with scintillator-based PET, and it is suitable, unlike BSI solutions, for application where the wavelength range of interest is below 450 nm. The top tier of the chip is used for light detection and houses all the SPAD sensors. The bottom tier, on the contrary, includes all the electronics required for the system, as will be discussed in the following.
A brief but complete analysis on different levels of the system is provided in this chapter: we start form the single pixel structure, to finish with the whole chip's 3D-stacked architecture.

2.1 SPAD sensor and pixel circuit

The single pixel detector implemented on Blueberry, is based on high-performance SPADs designed in 180 nm CMOS technology. In general, CMOS SPAD are less performing than those fabricated in custom epitaxial technology in terms of both PDP and noise level [6, 67]. The benefits, in turn, of easier on-chip integration

and cost reduction, thanks to mass production, are significant. The SPADs rely on a p-i-n structure, to ensure a wide range of sensitivity and low noise [11]. The pixel has a pitch of 50 µm and its square-with-rounded-corners profile ensures a a geometrical fill factor around 67 %. Fig. 2.1a shows the cross section of a circular SPAD with the same structure of the ones in Blueberry [6]. The only difference is the shape. The picture on the right, instead, (Fig. 2.1b) is a top view of the whole chip, with an inset displaying the true shape of the SPAD.



(a) **(b)**

Figure 2.1: (a): circular SPAD cross-section; in our case the pixel is square-shaped but, apart from that, the structure is the same [6]. (b): micrograph of Blueberry from the top [18]; the two halves of the chip are identical and formed by 4096 pixels each. The inset shows the square SPAD with rounded corners.

The SPAD is substrate isolated, with the p-well (PW) layer forming the anode of the SPAD and the buried n-well (BNW) layer working as cathode contact. A deep-n-well (DNW) connects this layer to the high voltage. All the SPADs of an array (one half of the chip) also share this common high voltage cathode. This allows us to apply two different excess bias voltages on the two arrays independently. The SPADs anodes, in turn, are all independent from the others, and they are directly connected to the bottom tier and the front-end circuit with their own TSV. The anode and the cathode are separated by an epi layer that creates a large high-field region, to achieve an increased sensitivity spectrum.
The breakdown voltage of the SPAD is about 22 V at room temperature, with a variation in temperature of around 1.8 mV/°C, according to [6].

Fig. 2.2 represents the SPAD front-end circuit, that follows the idea proposed in [29]. In this structure, thick-oxide transistors M_1 and M_2 form a cascode resistive divider, to increase the excess bias voltage up to 11 V, improving sensitivity and jitter performance [29, 68]. V_{cas} is supplied externally to fix the gates of the two

transistors. When an event is detected and an avalanche is triggered, the voltage at the source of M_1 rises and the overdrive decreases. If the voltage reaches $V_{cas} - V_{th}$, M_1 enters the sub-threshold regime and practically turns off, presenting a high impedance at the SPAD's anode.

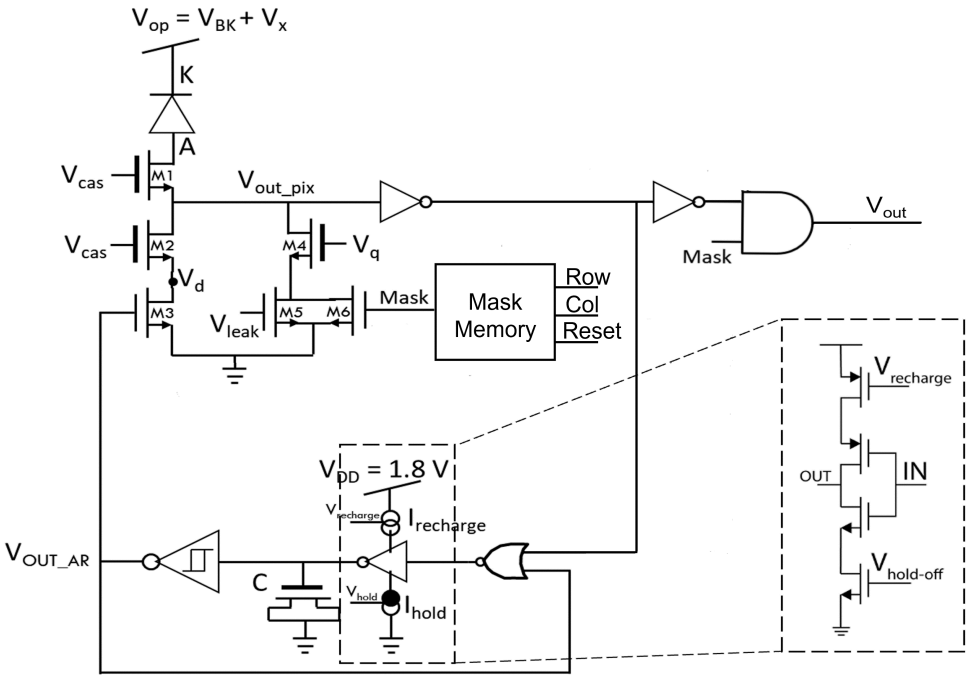


Figure 2.2: Schematic of the SPAD pixel circuit. Adapted from [6].

The pixel presents both passive and active recharge structures that can be used independently or together. The passive quenching branch is represented by M_4 and can be disabled by the transistors M_5 and M_6. The latter is connected to a 1-bit memory, storing the mask value, used to disable the pixel if noisy, thus improving the SNR. This bit can logically mask the output of the pixel through an AND-gate. The tunable active recharge system, instead, is formed by M_2 and M_3. This last one is connected to a feedback loop that activates it. The loop is formed by a NOR gate a Schmitt trigger and a tunable delay element. The latter is implemented with a current starved inverter (CSI) with a voltage controlled transistor for both the pull-up and pull-down branches (respectively a p- and an nMOS). A current mirror is included to set the controlling voltages of the structure [69]. This delay allows us to tune the hold-off time after one output pulse, a very important parameter to control afterpulsing in large SPADs. This system also enables the control the output pulse width and, in large part, the dead time of the pxiel, from 2 µs to

about 10 ns [14]. The stability of the circuit and sharp edges at the output are ensured by the inverting Schmitt trigger.
A detailed characterisation of SPADs similar to Blueberry's ones and designed by the same group, is provided in [6]. Their circular shape is the one shown in Fig. 2.1a, so also the pixel area will be different. On the other hand, all the other considerations and comments remain valid, and the paper represents the best available reference for results on the implemented SPAD sensor.

2.2 Cluster architecture

We shall now focus on the bottom tier. It is based on an *event-driven* architecture, and it includes a SPAD address tree, photon counters, time-to-digital converters (TDCs), signals distribution, readout scheduler and TDC calibration system [14, 18]. The MD-SiPM is composed of two identical arrays. Each of them is further divided in a 8 x 8 cluster matrix, where each cell is composed by 64 SPADs. Therefore, each cluster (or macro-cell) occupies a square area of 400 x 400 µm^2. The top-level structure of the bottom tier is shown in Fig. 2.3 [14].

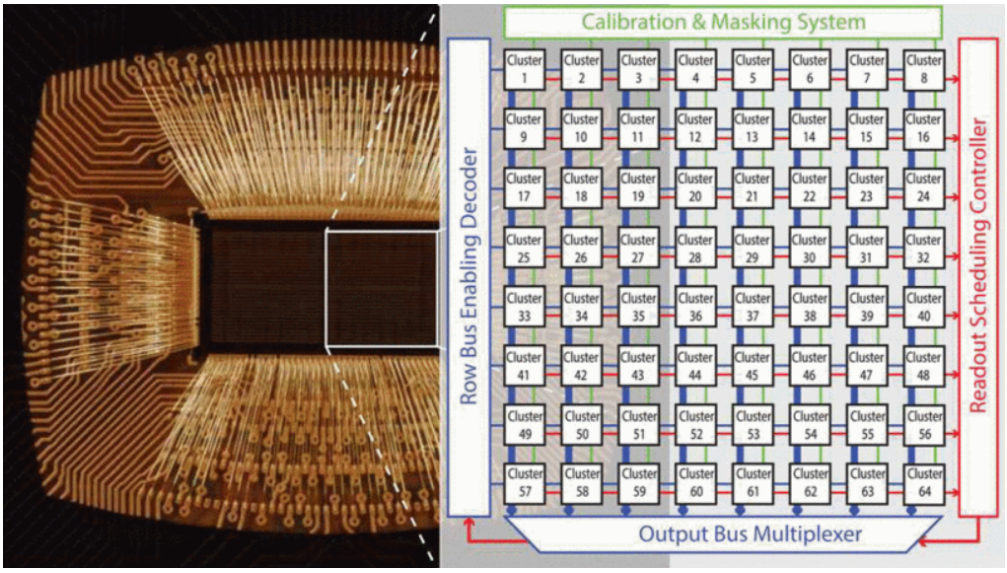


Figure 2.3: Micrograph of the bottom tier, wire-bonded on a custom PCB, used for tests (left) and a simplified sketch of the architecture of a single array (right) [14].

The random-access architecture, given by a row encoder and a column multiplexer, enables the independent readout of each array of the chip (2.3). The cluster

on the same column share a common output bus, whose access is controlled by an high impedance buffer. This buffering stage is enabled, in turn, by the row encoder. Data are output in parallel and each word includes the address of the read cluster, the address of the first SPAD that fired in that frame (the one that triggered the TDC), the TDC digital output code and the final photon count.
When a pulse is generated by the SPAD, it is propagated through an OR-tree for spatial compression within the cluster. The output of this tree is connected to the input of the TDC trigger. Before every OR-gate, a winner-take-all tree is connected (2.3). This circuit allows us to determine the address of the first firing SPAD of the cluster. The data is stored in a register and sent in output with the cluster address (in a ROM) through a dedicated bus, when required. This system increases the spatial resolution of the detector at the level of a single SPAD, improving the flexibility of the sensor for other applications [14, 18].
Each cluster also includes a photon counting system on the propagation tree, to estimate the number of detected photons on the surface. To this purpose, TSPC (True Single-Phase Clock) counters are connected at the output of the fourth OR-gate. Thus, four counter in total are present in each cluster. Their results are later summed by a 6-bits adder. Its current value is sampled with a frame signal (STOP) to be saved in a memory buffer.

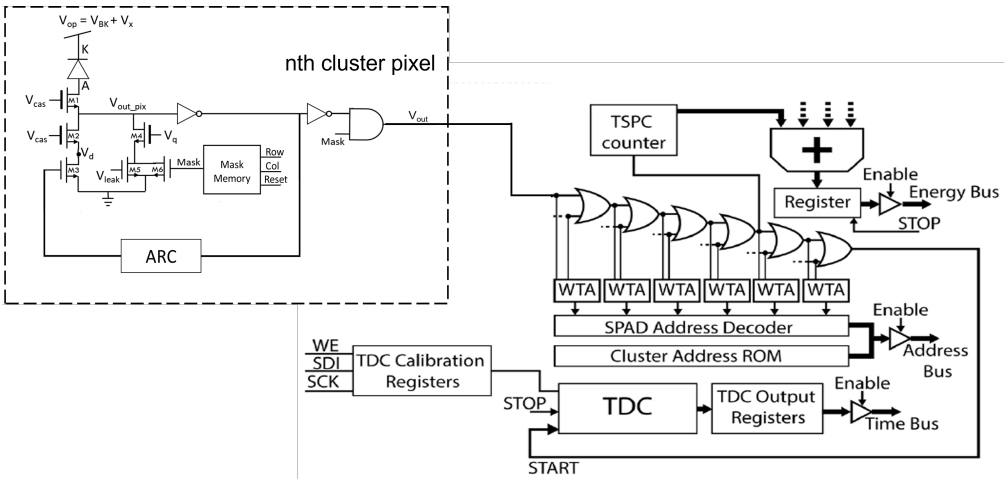


Figure 2.4: Schematic of the cluster architecture. Adapted from [6, 18]

The TDC in each cluster is based on a 9-stages multi-path gated ring oscillator (MGRO), as in [48, 70]. It presents an LSB of around 15 ps and a power consumption below 1.8 mW, as determined from measurements on an isolated structure [14, 18].
A calibration register is directly connected to the TDC (Fig.2.4). It is used to store

a 20-bits string with the calibration data for the TDC. Calibration addresses the delay elements of the TDC. Four banks of TSPC-FF registers sample the phases of the ring oscillator 2.5. Every one of these latches should be delayed 1/4, 1/2, 3/4 and 1 of the LSB value of the oscillator. This allows to divide the LSB by four and perform a linear interpolation within it, increasing the time resolution of the TDC. These delay elements consist of 4 digitally controlled MOS capacitors. Depending on four control bits connected to the gate of as many transistors, a given capacitive load can be added to the line, resulting in an additional delay. Calibration consists in setting these delays to their correct values for each bank of latches, modifying the four control bits. This means that a total 4×4 bits must be set. Four additional bits set the delay of the set/reset signals of the MGRO: the final calibration string will be 20-bits long, as anticipated. The four values stored in the phase register are then moved to the Output processing unit (OPU), together with the value of the loop counter, to calculate the final TDC digital code (the TDC phases are initially in thermometric encoding). The result can then be sent in output through a dedicated bus (2.4). Alternatively, a serial output is also available, to send directly the sampled phases in thermometric encoding, bypassing the OPU. Fig. 2.5 shows a sketch of the architecture of the TDC and its components.

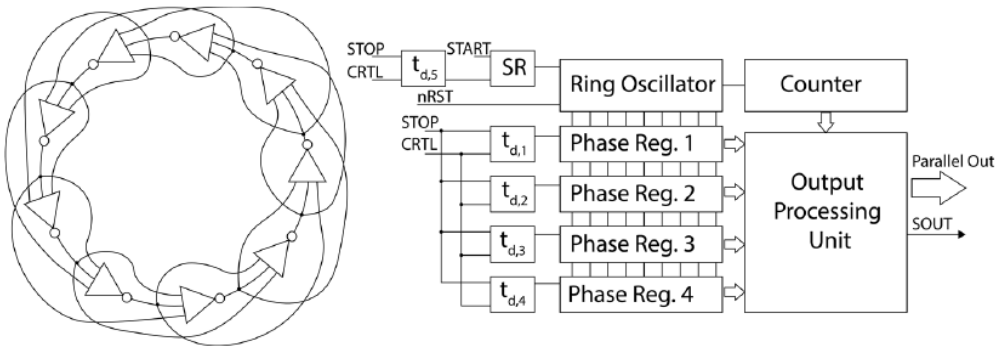


Figure 2.5: Schematic representation of the TDC architecture: 9-stages multi-path ring oscillator simplified sketch (left); component block diagram (rigth). Four banks of registers sample the ring oscillator output. Each of them is triggered by a STOP signal, delayed by a block $t_{d,i}$. The four values are then passed to a processing unit with the loop counter, to compute the final TDC code [18]

When a TDC is triggered, a flag bit is asserted to indicate that a valid data is ready to be sampled. All the flags from the clusters go in input to a control unit, i.e. the readout scheduler. This component implements a priority protocol where every cluster is associated to a given priority. This system is used to significantly reduce the readout time, the data throughput and the resulting power consumption,

thanks to a selective readout of the chip. Especially at low light levels, when not all the clusters have valid data, this component avoid the read out of every one of them. In addition, the system dead time is reduced even more by a buffer mechanism: it is designed to make the sensing and readout phase in parallel. A random-access readout architecture is equally available, enabling the selection of single clusters.

2.3 3D integration and chip bonding

The top sensitive tier and the bottom electronics tier are both fabricated 180 nm CMOS technology, and bound together in a 3D configuration. The connection between the two tiers is ensured by through-silicon vias (TSVs) and bump-bonds (Fig. 2.7). To this purpose, the top tier of SPADs was thinned to facilitate the creation of the TSVs at wafer level, one for each pixel (Fig. 2.6). After 3D-stacking of the two tiers, the correct electrical connection, between the top-tier TSVs and the bottom-tier front-end pixel circuit, is guaranteed by micropillars. These Sn-Ag structures can be seen at the back of the top tier (Fig. 2.8b).

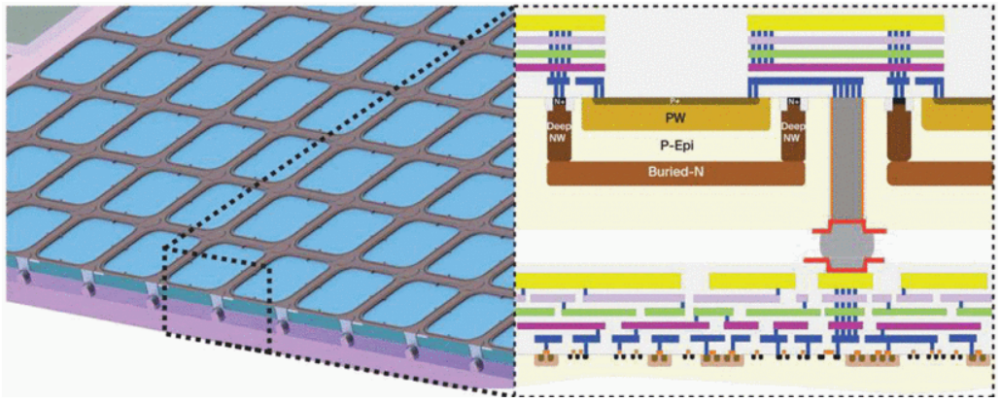
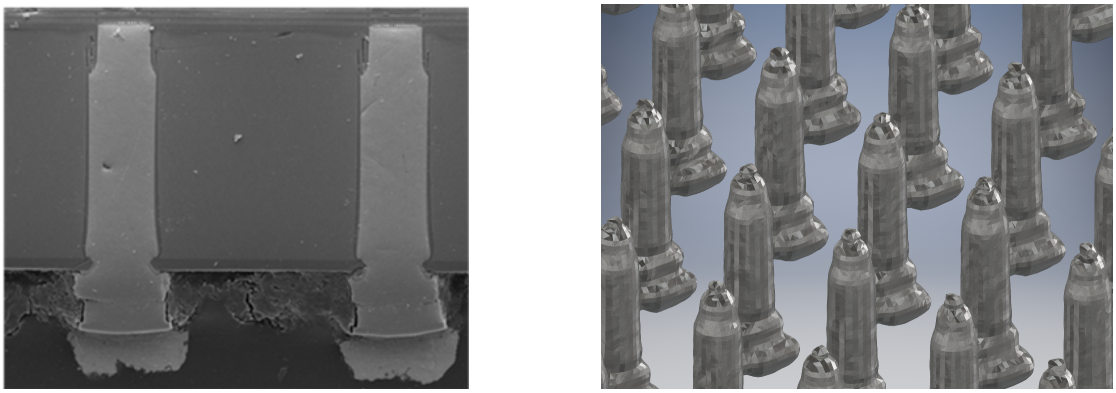


Figure 2.6: The chip top-tier presents square-with-rounded-corners SPADs and is 3D-stacked with the bottom tier (left). The connection between the top-tier's SPADs, their corresponding TSV and the bottom-tier's front-end electronics, is ensured by micro-bumps (right) [14].

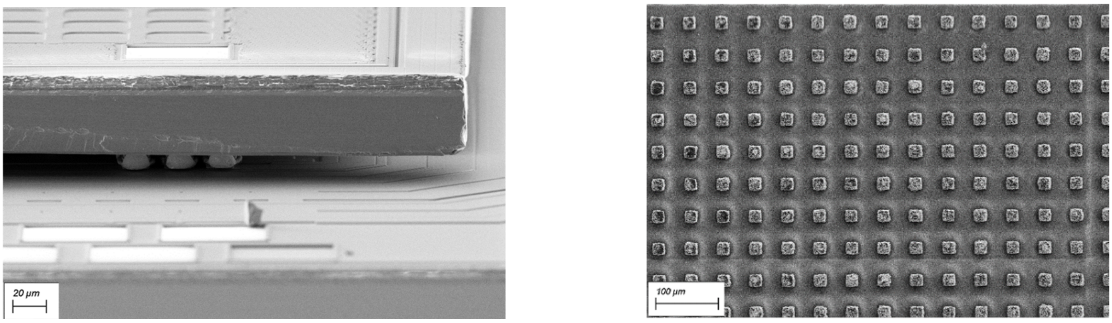
The bonding of the two tiers, besides ensuring the proper electrical connection, creates a mechanically stable structure, also thanks to relatively small TSV pitches [18]. The high voltage for the SPADs is provided directly on the top tier through a set of dedicated bonding pads. This reduces the risk breakdown of the oxide around the TSVs and avoids the need of specific TSV structures, different form the others. In this way, the voltage across the TSV oxide remains always below 5 V.

The wire bonding process might be quite stressful for the structure. In order to prevent possible cracks and damages to the chip during this process, an additional structure is implemented beneath the bonding pad, to provide further mechanical support (Fig. 2.8a).



(a) **(b)**

Figure 2.7: SEM cross-section of the top-tier, with the visible TSVs and the bumps at the bottom (a). X-ray tomography reconstruction of the TSVs and the underlying bumps (b).



(a) **(b)**

Figure 2.8: SEM image of the detail of the bonding pads of the top-tier: they are supported by additional micropillars to improve reliability and mechanical stability, critical especially during wire-bonding (a). SEM image of the Sn-Ag micropillars on the back of the top-tier (b). Taken from [18].

Chapter 3

Testing system

Blueberry testing system includes a PCB module and a software GUI (Graphical User Interface), implemented to test the chip in a semi-automatic way. The motherboard houses board-to-board connectors to connect up to two Opal Kelly boards, hosting one FPGAs each. The board also includes an embedded LVDS oscillator used for synchronization. The FPGA can be controlled by the software. Their task is to manage the communication and all the signals with the chip. On the top side of the board there are two connectors used to attach the unit under test (blue square on Fig. 3.1). Fifteen voltages between 0 V and 3.3 V are available and tunable through the user interface. Two additional voltages from 0 to 5 V are tunable by mean of potentiometers on the PCB.
The motherboard contains a power management system to control the voltages that need to be supplied to the unit under test (UUT) and two slots on the back to attach Opal Kelly boards (XEM 7360). The usage of two FPGAs allows a quite large versatility of the system. In addition, they are necessary to test the whole chip, composed of two identical, 64 by 64 matrices: each FPGA takes care of one half. The power management, in turn is driven only by one of them (FPGA B). Some potentiometer are also present and can be regulated to set some of the voltages, e.g the TDC supply. Fig. 3.1 displays a schematic of the top part of the motherboard, while Fig. 3.2 shows a picture of it.

A set of 20 SMA connectors (red dashed square in Fig. 3.1) is used to sense the generated voltages for monitoring purpose, or to supply the voltages from outside, in case of failure of the power management system. Two of them are used to supply high voltage to the sample (e.g. for SPAD Vop). If external voltages need to be supplied, the correspondent $0\,\Omega$ jumper must be disordered. The jumpers are highlighted by green dashed squares in Fig. 3.1.
In the following sections, we will focus on the firmware designed to perform some tests, and some related sections of the GUI, to control and communicate with

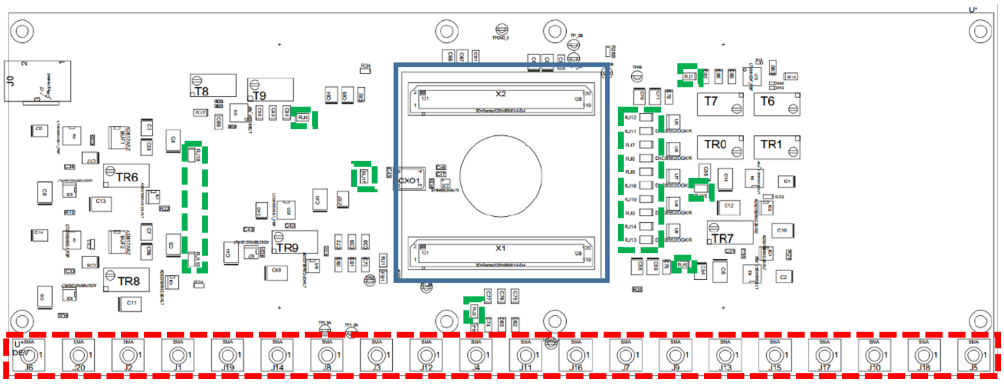


Figure 3.1: PCB schematic, highlighting the UUT connectors (blue), the SMA connectors (red) and the related jumpers (green). Courtesy of Francesco Gramuglia

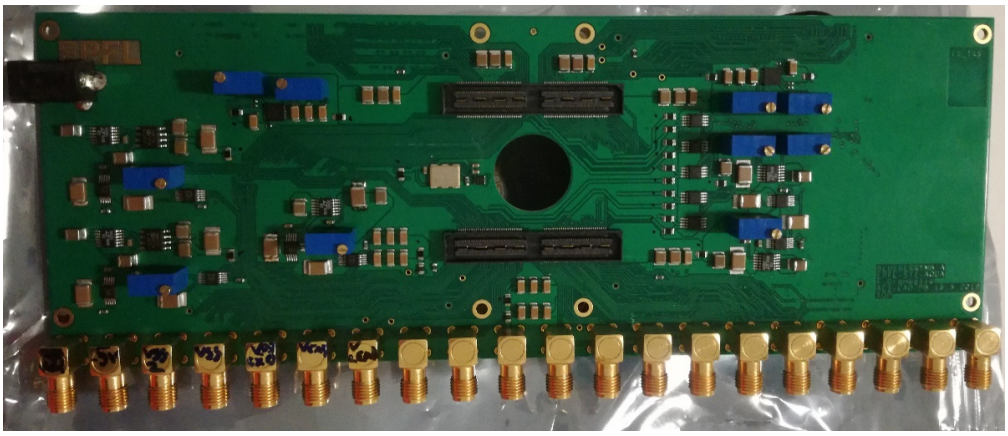


Figure 3.2: Picture of the top side of the PCB. Courtesy of Francesco Gramuglia

FPGA through the Opal Kelly FrontPanel environment. Sec. 3.1 will provide an overview on the main components of the testing structure, implemented on FPGA; Sec. 3.2 will include a brief description of the GUI and its main functions.

3.1 Firmware

The target device is a Xilinx Kintex-7 FPGA, integrated on a XEM 7360 Opal Kelly board with a SuperSpeed USB 3.0 interface for data transfer [71]. The XEM 7360 is fully supported by the FrontPanel Application and its SDK, a C++ class library to interface the software with the board. Its main purpose is to transfer data between PC and FPGA so to provide an effective controllability and observability

of the design.
Besides several components making up the FrontPanel environment, on the FPGA side of the interface, some HDL modules enable the communication with the PC. So-called "Endpoints" are employed to connect FrontPanel components to signals in the design. They basically work just as external pins. To control or observe some signal, it is sufficient to connect them to the endpoint ports. These are all connected to the same shared bus: each one is told when it can assert the data on the bus, otherwise it drives 0. A bit-wise OR operation, performed by the Wire-OR component, passes the requested data in output. A Host Interface module, connected to the same bus, takes care of the communication with the software. Fig.3.3 shows a simplified view of the FrontPanel HDL modules on the FPGA. New endpoints can be added to the design, instantiating them as additional modules. They are designed to consume little FPGA resource so that they should minimally affect the design (precise resources utilisation can be found in [72]).

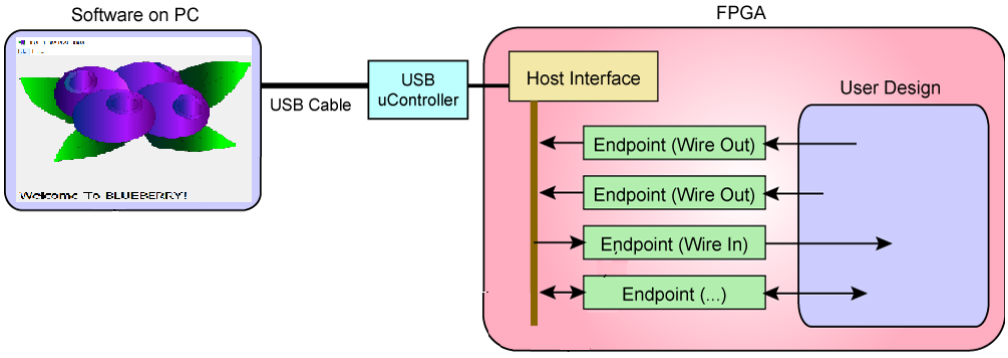


Figure 3.3: Simplified diagram presenting an example of the general structure of the Opal Kelly board and the integrated FrontPanel modules: all the endpoints share the same bus with a Host Interface that controls the communication to the PC through the USB interface on the XEM board. Adapted from [72]

There can be several types of endpoints: Wires, Triggers and Pipes. They are either directed in or out of the design, from the perspective of the FPGA (an "in" endpoint for example, transfer data in the design). Each endpoint can be accessed independently thanks to its associated address. Below, each endpoint is described more in detail for a USB 3.0 interface.

Wires

Wires are asynchronous connection between the PC and the design. Each of them adds 32-bit signals, connected from the FrontPanel to the design or viceversa,

depending on the transfer direction.
Although they remain asynchronous with respect to the design, all the wires are captured and updated simultaneously on the Host Interface clock. This means that all the 64 Wire Ins (or the Wire Outs) are transferred together at the same time (synchronous to the Host Interface clock).

Triggers

Triggers are synchronous connections between the software and the HDL design. The synchronisation clock can be defined in the design and the module itself takes care of the proper clock domains crossing.
A Trigger In creates a one-cycle pulse signal used to initiate an event, e.g to start an FSM. A Trigger Out is set when a signal is asserted on the rising edge of the clock and can be used to signal an event, like the end of an FSM. FrontPanel polls the FPGA periodically and the Trigger Out remains high until the next poll.

Pipes

Pipes are synchronous connections designed to transfer multi-bytes data to (or from) the endpoint. For the design a Pipe is always a master: the PC and the HDL module control the transfer in both directions. Moreover, the communication is synchronous to the endpoint clock, i.e. the Host Interface clock. Clearly, the target FPGA must accept/pass data when required by the endpoint. Coupling this module with a FIFO is a common solution to keep up with the data rate and to transfer a whole block of data. This also ensures a safe crossing of the different clock domains.
The length of single words that can be passed through a Pipe in one clock cycle is 32 bits.

3.1.1 Chip testing

This first component controls the readout of data from the chip. It both provides the control signals to the chip, and stores the output data to be analysed. The component includes a control unit, simply implemented with an FSM whose task is to manage the signals and data transfer with the chip.
The readout occurs following a rolling shutter scheme: after the integration window ends, the data of each cluster are stored in a FIFO, starting from cluster 0 to 63. A whole data string is composed of 36 bits read in parallel. For convenience they are saved as 64 bits-wide string in the memory (zeros are added as MSBs). Once the readout has finished, they can be transferred to a PC through a Pipe Out end point (32 bits at a time), when requested by the software. Several frames can be

ENDPOINT	ADDRESS	SYNC./ASYNC.	DATA TYPE
Wire In	0x00 - 0x1F	Asynchronous	Signal state
Wire Out	0x20 - 0x3F	Asynchronous	Signal state
Trigger In	0x40 - 0x5F	Synchronous	One-shot
Trigger Out	0x60 - 0x7F	Synchronous	One shot
Pipe In	0x80 - 0x9F	Synchronous	Multi-byte transfer
Pipe out	0xA0 - 0xBF	Synchronous	Multi-byte transfer

Table 3.1: Summary of the types of endpoints modules as in [72]

captured in one measure: their value can be chosen up to a maximum of 2047. The length of the integration window can be set by the user. The values range from 1 to 100, and represent the time span in term of number of Opal Kelly clock period ($\approx 10\,\text{ns}$). Fig.3.4 presents a schematic view of the testing system implemented to test the chip.

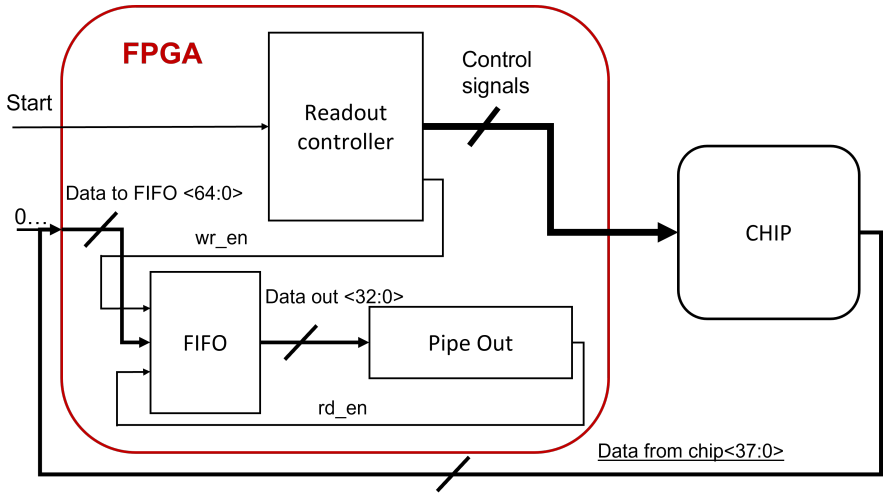


Figure 3.4: Schematic including the readout system main components and their interconnections

The diagram in Fig. 3.5 represents the algorithm implemented for the FSM. A brief explanation of its tasks is provided too. Fig. 3.6 shows the timing diagram

for the main signals.

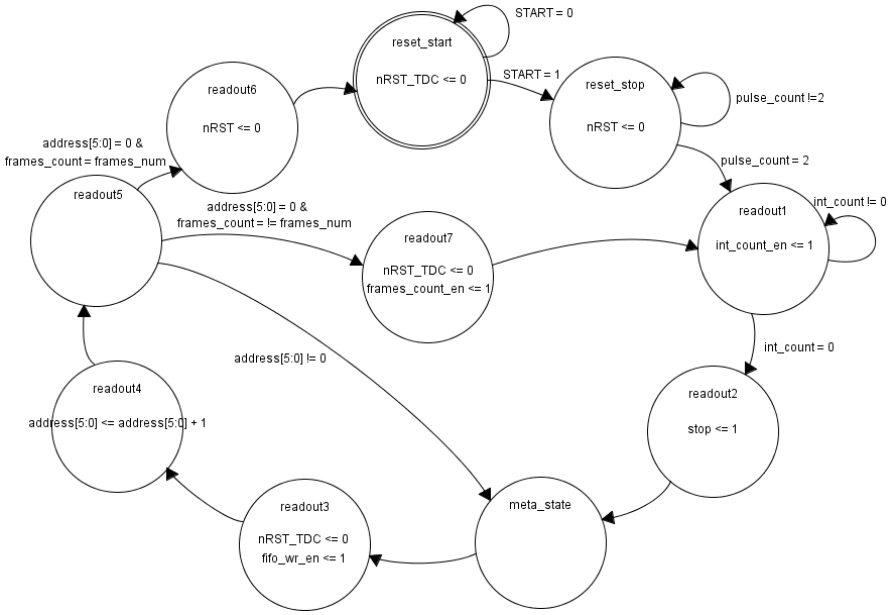


Figure 3.5: Moore-type FSM implemented for the control unit to test the chip. The initial state is "reset_start" (highlighted). All the signals take their default values unless specified otherwise, to avoid latches in the design. The assignation in the states take into account only the transitions to non-default values. The transitions occur when the conditions written on them is verified. Otherwise the other transition, if present, takes place. When only one transition is possible, it occurs after one clock cycle, unless specified.

- **Reset start**: this is the initial state of the FSM where every register is reset and disabled. All the reset signals of the TDCs are asserted.
  If the start signal is received a transition occurs to "reset stop" state.

- **Reset stop**: in this state the counter and the register containing the address of the firing SPAD are reset. A counter is enabled to keep the FSM in this state for two clock cycles, just to be sure that the reset signal is long enough. After two clock cycles the FSM moves to "readout 1" state.

- **Readout 1**: here the integration window begins, connecting the TDCs to the SPADs with the signal *sel_start_readout*, and removing the reset signals. A downward counter is enabled to keep the FSM in this state for number of cycles equal to the chosen integration value. In this state, photons are

collected and data in the chip are updated continuously.
Once the counter reaches 0, it means that the integration window is finished and the state changes to "readout 2".

- **Readout 2**: here the sampling of the data occurs: the stop signal for the TDC is asserted so that the data are stored in the registers on the chip, ready to be read. The assertion of this stop signal corresponds to the end of the integration.
  After one cycle the FSM moves to the next state.
- **Meta state**: this is an additional state to be sure that the data are stable before they are stored on the FPGA.
  The next state is "readout 3".
- **Readout 3**: in this state the rolling-shutter readout begins. The en signal for the fifo, *fifo_wr_en*, is asserted to save the output data from the chip. A 6 bit signal, *address*, represents the address of the cluster to read. It is initially 0 but will be incremented to read the whole chip. The TDC are also reset.
  After one clock cycle the state changes to "readout 4".
- **Readout 4**: here just the address signal to the chip is increased to change the selected cluster.
  The next state is "readout 5".
- **Readout 5**: here we only check if the readout has finished or not. Firstly, we check the address: if it is different from 0, it means that there are still cluster to be read so the FSM goes back to "meta state"; if it is 0 we must check if more frames must be taken or not. In the first case, we go back to state "readout 7", before restarting with a new integration. Otherwise, if all the frames are done the readout has finished and all the data have been saved in the FIFO. In this case we move to state "readout 6".
- **Readout 6**: this is the final state, where the flag *readout_complete* is asserted to inform the software that data are ready to be transferred. The chip is also reset.
  Next state is the initial one, "reset start".
- **Readout 7**: in preparation for a new integration, we assert the chip reset and set the value for the integration window. We also increment the counter keeping trace of the captured frames.
  We move back to "readout 1", ready to start a new integration.
- **Others**: if an error occur in the selection of the state, nothing happens, all the signals are de-asserted.
  A transition towards "start reset" occurs right after.

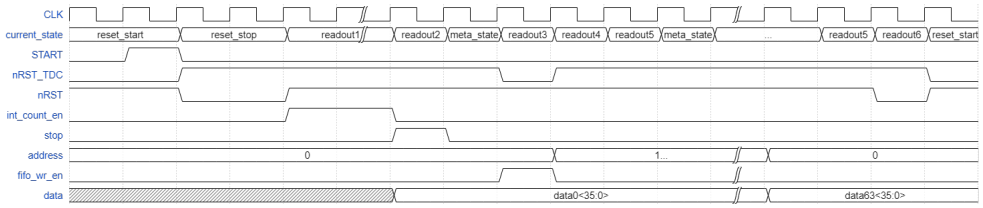


Figure 3.6: Timing diagram of the FSM with its main signals. After the stop signal is used to sample the data in the registers on chip, two clock cycles pass before it it written on the FIFO on the FPGA, to avoid metastability.

Double sampling

During the test of the chip, some issues that could be attributed to an offset on the counting data arose. To check this eventuality, we developed an additional readout scheme. The idea was to sample the data twice: once at the beginning of the integration window, after the reset; once at the end of it, as before. The different data are then saved in two separate FIFOs. In this way, we could check the presence of a different offset in every measurement. Probably the cause was a failed reset of the registers in the chip. Saving both the data, before and after the integration, a relative measure could be performed excluding the effect of the offset.
The number of data is doubled with this scheme and it would be a problem to read the whole chip together. The solution was to select only one cluster to test, without the need of the rolling-shutter readout anymore. The address can be selected arbitrarily as the integration value and number of frames. Here the maximum integration value was increased to 4095 to reach an integration window of more than 40 µs.
The structure of the component and the general structure of the implemented FSM are the same as in section 3.1.1, with some obvious, minor differences. Its timing diagram is shown in Fig.3.7.

3.1.2 TDC testing

This component takes care of the testing of the TDCs on chip, when disconnected from the SPADs. The test has the goal to check the TDC transfer characteristic and verify its linearity. To do so, it is necessary to send the start and stop signal for the TDC at increasing distance in time. For each of these timestamps (the time difference between a start and a stop signal), the output code of the TDC is sampled. The resulting plot should present the typical linear trend of every converter: increasing the timestamp, the TDC output code should grow coherently.

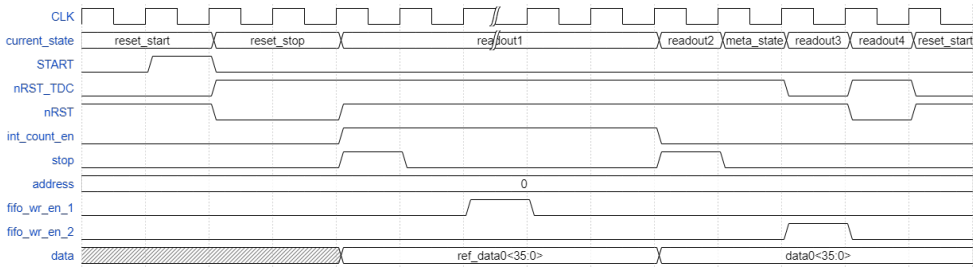


Figure 3.7: Timing diagram related to the double sampling readout. The main differences from the one in Fig. 3.6 are the constant address (only one cluster is tested) and the two enable signals for the FIFO, corresponding to the two different sampling. For both, two cycles pass before the data are stored in the memories.

Moreover, a step-wise trend is expected, where the step width corresponds to the LSB of the TDC output, i.e. its resolution.
The main issue in this case is to provide a significant time-shift between the start and stop signals and between two subsequent timestamps. The TDC resolution is about 20 ps, so a shift of the same amount is necessary to obtain minimum-width steps.
Several solutions were explored to perform the adequate shift on FPGA. The straightforward one is to synthesise a simple counter: in this case the shift corresponds to a clock period; unfortunately, such a small shift would require a clock frequency of several GHz, impossible to achieve on FPGA.
An alternative is to employ the Xilinx primitive IDELAYE2 [73]. The minimum shift depends also in this case on the clock frequency (see [74]). In our case it was impossible to obtain a value below 52 ps, too large for our purpose.
The last and optimal solution is to employ the dynamic phase shift of an MMCM (Mixed Mode Clock Manager) Xilinx IP. This function enables fine shift of the component output clock(s) according to some flags (all synchronous to a clock) and previous shifts. When the enable signal is asserted, a phase-shift increment/decrement is initiated. Every shift increases/decreases the phase of a 1/56th of the VCO (Voltage Controlled Oscillator) period and it takes always 12 clock cycle to complete. With a VCO frequency of 1 GHz, a fine shift of less than 18 ps can be achieved. This scheme of operation is ideal to obtain subsequent shifts one after the other, as required for the test.
A controller component implements an FSM that manages the control signals for both the MMCM and the chip. Once the data are ready, it also takes charge of the sampling, storing them in an appropriate FIFO. The data string has the same length as in Sec. 3.1.1, but in this case only the 18 significant bits of the TDC output are stored.
Fig. 3.8 shows the connections between the different components employed for the

TDC testing. Fig. 3.9 highlights the connections required for the dynamic shift, with all the control signals coming from the Control Unit (CU).

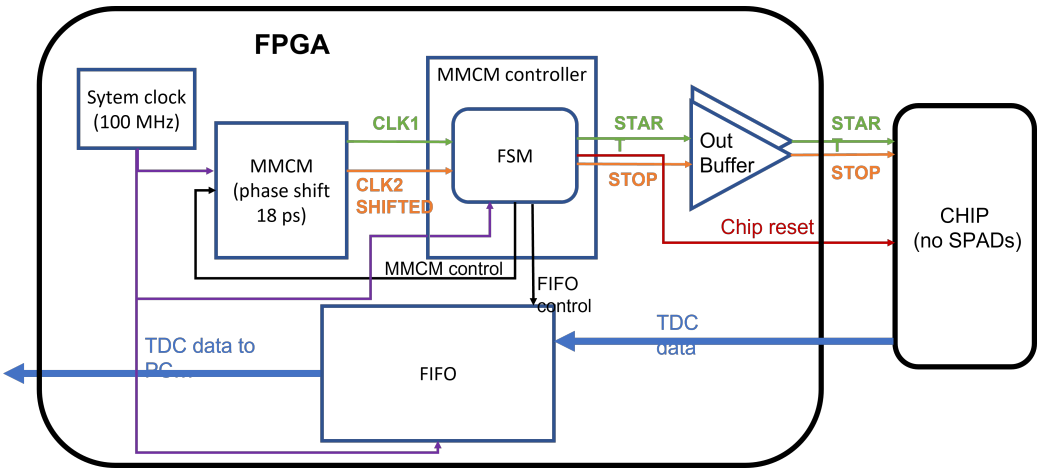


Figure 3.8: Schematic including the TDC testing system implemented on FPGA with its main components and their interconnections.

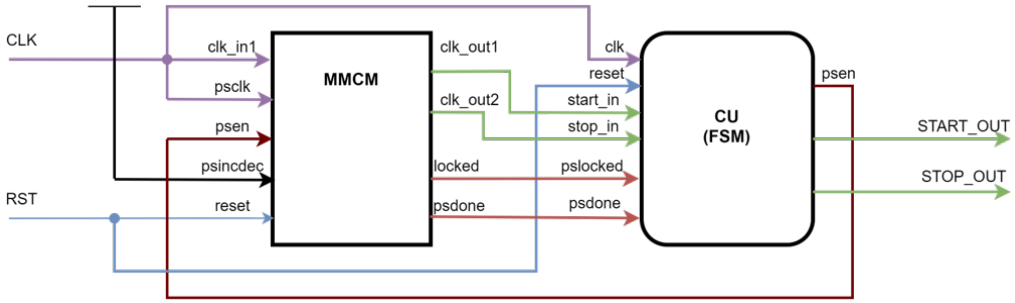


Figure 3.9: Schematic of the connections between the MMCM and the Control Unit. The signal psen is the enable signal for the shift; psincdec fixed at 1 defines the shift as increment; psclk is the reference clock of the component; locked informs when the component is ready, psdone signals when the phase shift is complete.

The number of total shift to be performed can be set arbitrarily up to a maximum of $2^{16} - 1$. The 18-bit TDC output data from the chip are stored, for convenience, in 32-bit-long string, adding zeros as MSB. They are later sent to the PC through a Pipe Out, when requested by software. Clearly, the throughput can be very high, so only one TDC can be tested at a time. The address of the TDC (or the cluster) can be chosen by the user and is directly passed to the chip as 6 bit string.

The implementation of the FSM algorithm is presented in Fig. 3.10 with a brief explanation and a related timing diagram in Fig. 3.11.

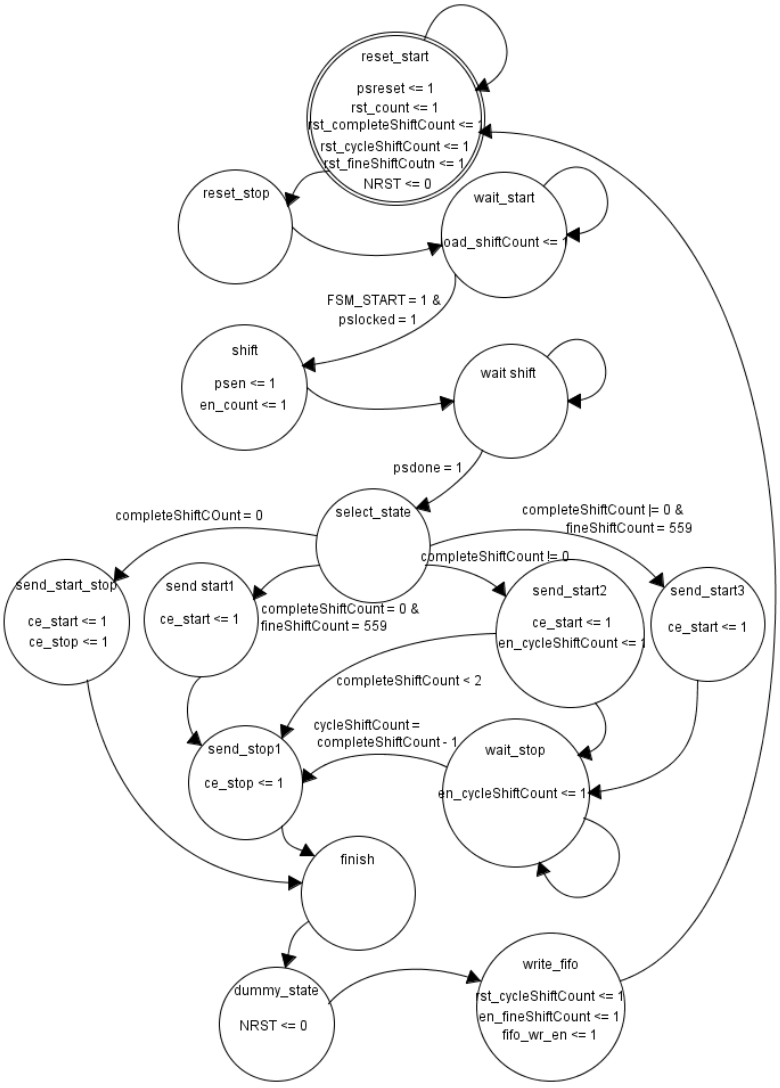


Figure 3.10: Moore-type FSM implemented for the control unit to test the TDC. The initial state is "reset_start" (highlighted). All the signals take their default values, unless specified otherwise, to avoid latches in the design. The signal values in the states take into account only the transitions to non-default values. The transitions from one state occur when the conditions written on them is verified. Otherwise the other transition takes place, if present. When only a transition is possible, it takes places after one clock cycles unless specified.

- **Reset start**: usual initial state where everything is reset. After one clock

cycle a transition to "reset stop" occurs.

- **Reset stop**: resets are de-asserted and all signals take their default values. After one cycle the FSM move to the next state.
- **Wait start**: here we load the down counter that keeps trace of the number of timestamps left. If the start signal is received and the MMCM is locked, the test can begin and we move to "shift" state.
- **Shift**: here the enable signal for the MMCM shift is asserted (*psen*) together with the enable signal for the down counter. After one cycle the state changes to "wait shift".
- **Wait shift**: this is an idle state to wait for the shift to complete. When it does, the *psdone* flag is assserted and we move to the next state, "select state".
- **Select state**: in this state the FSM decides what to do next depending on the point we reached in the test. This information is included in some flags and counters, keeping trace of the shifts. Depending on their values the next state can be "send start stop", "send start 1", "send start 2", "send start 3" (see Fig. 3.10).
- **Send start stop**: here the output buffers for start and stop signals are simply enabled together. After one cycle, the FSM moves to "finish" state.
- **Send start 1**: only the start signal's output buffer is enabled, because here an additional cycle delay must be inserted[1]. Next state here is "send stop 1".
- **Send start 2**: this state works as "send start 1", the only difference being the activation of a counter and the state transition. The latter depends on the value stored in a variable; the next state can be "send stop 1" or "wait stop" (see Fig. 3.10). This state is used after the first 360° shift is reached, so at least one additional cycle delay is required to increase the timestamps.
- **Send start 3**: this works as "send start 2" with the difference that the counter is not activated yet. This will add an additional cycle delay. Next state is "wait stop" where the counter will be activated.

[1]The reason is that after some shifts (559), the next one will be equal to a clock period ($10ns = 560 \times 1/(56 \times 1GHz)$). When this happens, the output clocks from the MMCM (used as start/stop signal as in Fig. 3.9) are synchronous again (phase shift of 360°). This means that additional clock cycles must be manually added to continuously increase the timestamps. In addition, this occurs repeatedly because the clocks are periodic, and every time an additional cycle must be added. This is taken into account with the different "send start" states.

- **Wait stop**: if the FSM reaches this state, it means that more than one additional clock cycles must be added to the MMCM phase shift. A counter (the same activated in "send start 2") keeps trace of them and when their value equals the right one (depending on how many times a complete 360° shift was already reached before), the FSM moves to "send stop 1".

- **Send stop 1**: with one cycle delay with respect to the start signal, the stop buffer is enabled. We can move to "finish" state.

- **Finish**: this state and the next one are idle states inserted to avoid metastability. They introduces an additional clock cycle delay before the data sampling occurs. Next state is "dummy state".

- **Dummy state**: apart introducing a cycle delay, here we also reset the TDCs. The data are already sampled in registers on-chip when the stop signal was asserted, so they are not reset. Next state after one cycle is "write fifo".

- **Write fifo**: the data are finally stored in the FIFO. One counter, keeping trace of the completed measurements, is enabled. Another one is reset. If there are more timestamps to be sent, the FSM goes back to "shift" state, otherwise to "reset start".

- **Others**: if none of the correct states is selected, all the signals take their default values and a prompt transition to "reset start" is imposed.

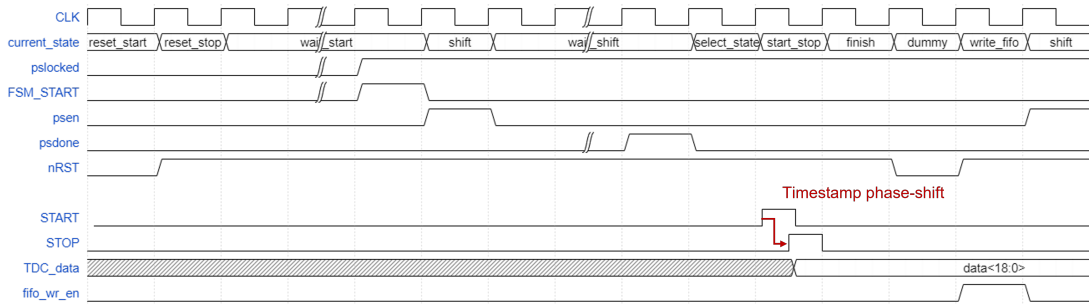


Figure 3.11: Timing diagram for the TDC test. After the dynamic phase shift is complete, the start and stop signal are sent to the chip out of phase. The stop signal samples the TDC data that are sent back to FPGA. After more than two clock cycles to avoid metastability, the data is written in the FIFO. The following shift is initiated right after.

Serial readout

A serial readout is also possible from one pin of the chip. The scheme was implemented for the TDC testing explained so far, but it was not tested yet. The idea behind this readout is that once the data are ready, a load signal sample them in some shift registers on the chip. Then if a sufficiently slow clock is provided (some tens of MHz) these data are sent in output one bit at a time, starting from the MSB (see Fig. 3.13). The whole string in this case is 50-bit long and includes some additional information like a "count valid" flag.
This readout scheme includes a certain dead time given by the data transfer. In addition, the serial clock should not be very fast, to allow a correct communication and avoid metastability. These aspects make it much slower than the parallel readout explained before in this section. To reduce the required time and also limit the memory size, this time maximum total number of timestamps is $2^{14} - 1$. The address is set as 6-bit string and two decoders (row and column) enable the single serializer. As before, for the moment only one TDC at a time can be tested.

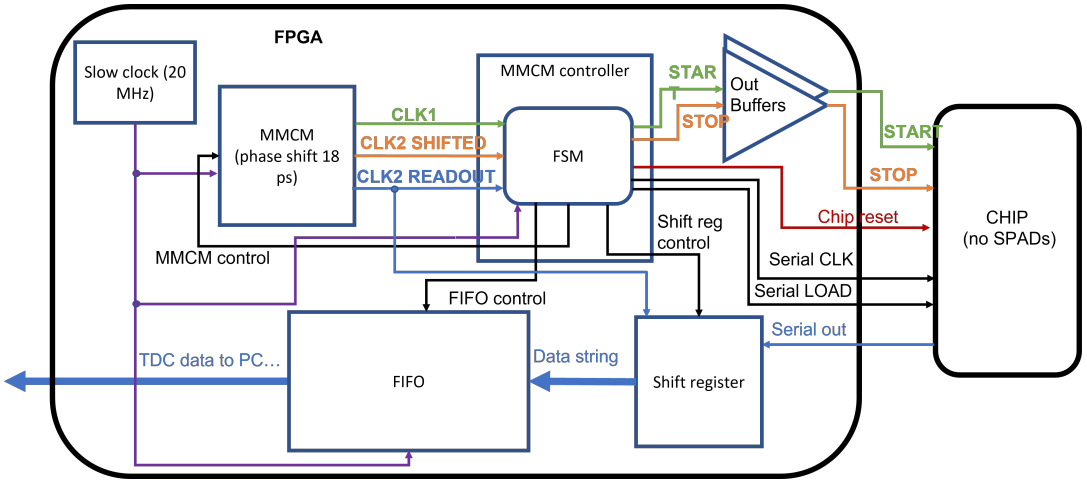


Figure 3.12: Schematic of the implemented TDC data serial readout system with its main components and their interconnections. The main differences from Fig. 3.8 are the serial control signals, the readout clock and the additional shift register.

The structure of the component, sketched in Fig. 3.12, is almost the same of the one with the exception of the serial connections and clocks. The MMCM outputs also a third clock, "clock readout", shifted of 180° and used for the data sampling to avoid metastability. An additional shift register stores the serial data from the chip and then passes the whole 50-bit string to a FIFO. This shift register samples the data with the shifted "clock readout".
Also the FSM follows the same implementation of the one in Fig. 3.10, with some

minor differences. Two states are added after the "dummy state": one is used to send the serial load; the next one to send the serial clock for the data transfer. The output bit is saved as LSB in the enabled shift register. The FSM remains in this state until the whole 50-bit word is sent, then moves to "check finish" state. This state works as "write fifo" state before: it stores the data in the FIFO and checks if the readout is done. If this is the case, we move to an additional state, "finish", where a *done* flag is asserted, otherwise we go back to "shift". Once in "finish" the FSM goes back to "reset start".

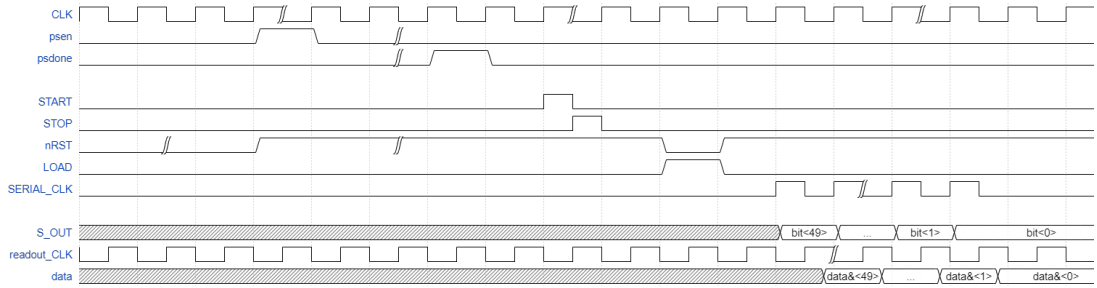


Figure 3.13: Timing diagram of the serial TDC test. The purpose of this diagram is to show hoe the serial communication with the chip works, so many other signals are omitted. After the start and stop signals are sent, the load is activated. Once the data are stored in the registers on chip, the serial clock is sent to get the data in output. These data, are then sampled one bit at a time with a readout clock in anti-phase to prevent metastability, and stored in a shift register, here represented by the variable "data". Later the whole 50-bit word is saved in a FIFO.

3.1.3 TDC calibration

Another component included in the design takes charge of the calibration of the TDCs. Its role is complementary to the one for the TDC test of Sec. 3.1.2 and it will be necessary in the future to obtain correct results. In particular, it is used to improve the temporal resolution of the TDC thanks to the interpolation of timestamps smaller than the nominal LSB of the TDC (see Sec. 2.2 for more details). Calibration addresses the delay elements of the TDC. Depending on four control bits, a given capacitive load can be added to the line, resulting in an additional delay. Calibration consists in setting these delays to their correct values for each bank of latches, modifying the four control bits (Sec. 2.2). This means that a total 4×4 bits must be set. Four additional bits set the delay of the loop counter so that the final calibration string will be 20-bits long.
The calibration circuit is sketched in Fig. 3.14a. The two shift registers, one for the rows and one for the columns, enable the input of the TDC registers. The data in

these shift registers must be loaded serially by the FPGA, also providing a suitable clock. Once the operation is done, the two output enable signals can be asserted: the data stored in the shift registers are presented to the output, so that the chosen clusters are enabled. As shown in the schematic, two binary trees are connected to all the TDC registers. The *SDIconfig* one sends the configuration data serially. The *SCLK* is a common serial clock to sample these data. In summary, two serial communications occur: one to load the enable shift registers; the second to pass the calibration data to the TDC registers.
Every TDC might require a different calibration string. A suitable scheme is to select them one by one and send the appropriate data. The operation can be done sequentially until every TDC is calibrated, or random access, selecting only one of them.

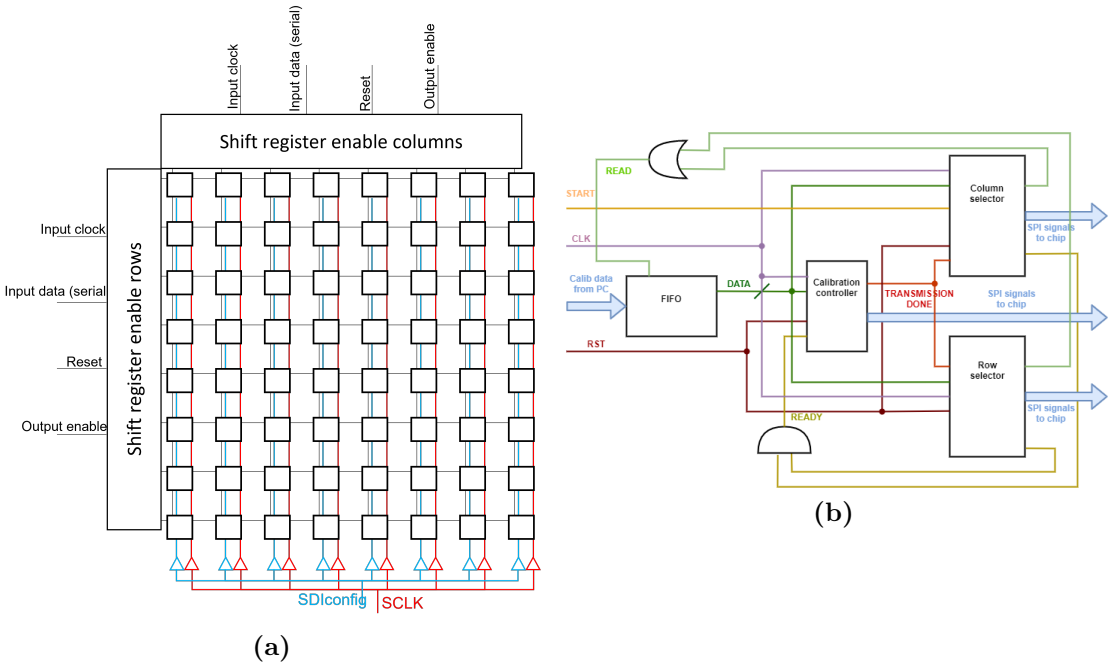


Figure 3.14: (a): Calibration circuit on chip (courtesy of Francesco Gramuglia); (b): sketch of the component structure and the connections between the two control units. The clock CLK, reset RST and start signal come from the design. The SPI signals to the chip include all the required data and clocks, but were merged for clarity.

The component implemented on FPGA includes three different control units. Two of them are employed for the control of the two enable shift registers (column/row selector), the third for the serial load of the calibration data in the registers

(calibration controller).
Data are passed through a Pipe In and stored in a 32-bit wide FIFO. 20 bits are the calibration data, 6 define the total number of TDC left to calibrate, and other 6 correspond to the cluster address. These last 6 bits are sent to the column/row selectors where a decoder defines the sequence to be loaded in the shift registers. The three control units include an FSM each, to manage the serial communication with the chip. All of them are similar because the transfer occurs more or less in the same way. The diagrams in Fig. 3.15 show the implemented algorithms for the shift register controller and the calibration controller. In Fig. **??** a simplified timing diagram of the component is displayed.

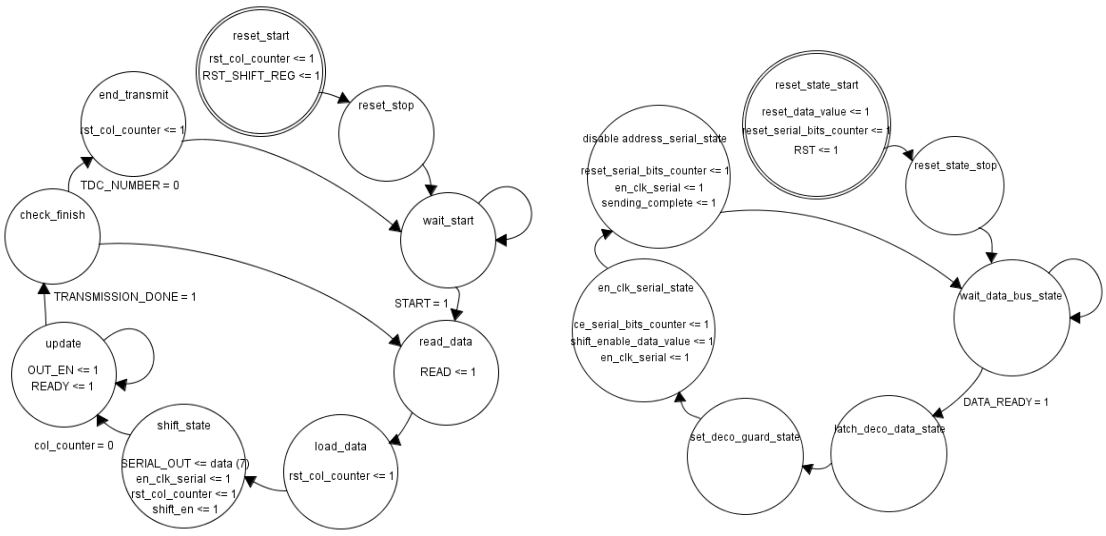


Figure 3.15: (a): Moore-type FSM implemented for the Row/Column selector; (b): Moore-type FSM implemented for the calibration controller. In both diagrams the initial state is highlighted. All the signals take their default values, unless specified otherwise, to avoid latches in the design. For clarity, the signal values in the states take into account only the transitions to non-default values. The transitions from one state occur when the conditions written on them is verified. Otherwise the other transition takes place, if present. When only a transition is possible, it takes places after one clock cycles if no condition is specified.

The two shift register are controlled by the same FSM described below (rox/column selector).

- **Reset start**: initial state where all the signals are reset. After one cycle the FSM moves to "reset stop".

- **Reset stop**: state to de-assert all the reset signals. All the other signal take their default values. Next state after one cycle is "wait start".

- **Wait start**: idle state where no signal is updated. The FSM remains in this state until the *START* flag is received and it can move to "read data" state.

- **Read data**: here the *READ* flag connected to the FIFO is asserted: the first data string is read but only the bits including the TDC address and number are stored. Next state is "load data".

- **Load data**: a state to make sure the data are stable before sending them to the shift registers. The counter (see "shift state") is reset. After one cycle the FSM moves to "shift state".

- **Shift state**: data are sent to the registers. The shift register storing the data on FPGA is enabled to output one bit at a time. The serial clock is enabled and sent through a buffer, phase-delayed 180° with respect to the data. A counter is also enabled to keep count of the sent bits (8 in total). When this down counter reaches 0, the FSM changes state to "update".

- **Update**: in this state, all the shift registers are loaded with the required data. The output enable is asserted to activate the TDC registers. A *READY* flag is passed to the other CU (calibration controller FSM) to start the transfer of the calibration data. Once the transfer is complete, a *TRANSMISSION_DONE* flag is received and we move to "check finish" state.

- **Check finish**: here the FSM checks if more TDCs must be calibrated. In this case it goes back to "read data" state; otherwise it moves to "end transmit".

- **End transmission**: The calibration is done. The counter is reset before going back to "reset start" state.

- **Others**: if some error occurs, all the signals take their default values and the FSM moves to "reset start" after one clock cycle.

On the other hand, the calibration controller FSM manages the serial data transfer to the TDC registers.

- **Reset state start**: initial state where all the signals are reset. After one cycle the FSM moves to "reset state stop".

- **Reset state stop**: state to de-assert all the reset signals. All the other signals take their default values. Next state after one cycle is "wait data bus state".

- **Wait data bus state**: no signal is updated. When *DATA_READY* flag is received from the other two controllers the FSM moves to "latch deco data state".
- **Latch deco data state**: data are sampled from the FIFO and stored in a shift register. Next state is "set deco guard state".
- **Set deco guard state**: a state to make sure the data is stable before sending it to the TDC registers. After one cycle the FSM moves to "en clk serial state".
- **En clk serial state**: data are sent to the registers. The shift register storing the data on FPGA is enabled to output one bit at a time. The serial clock is enabled and sent through a buffer, phase-delayed 180° with respect to the data. A counter is also enabled to keep count of the sent bits (8 in total). When this down counter reaches 0, the FSM changes state to "disable address serial state".
- **Disable address serial state**: the serial bits counter is reset. The clock output buffer is still enabled for one cycle to send the last bit. A *sending_complete* flag is sent to the other control units (see "update" state before).
- **Others**:if some error occurs, all the signals take their default values and the FSM moves to "reset start" after one clock cycle.

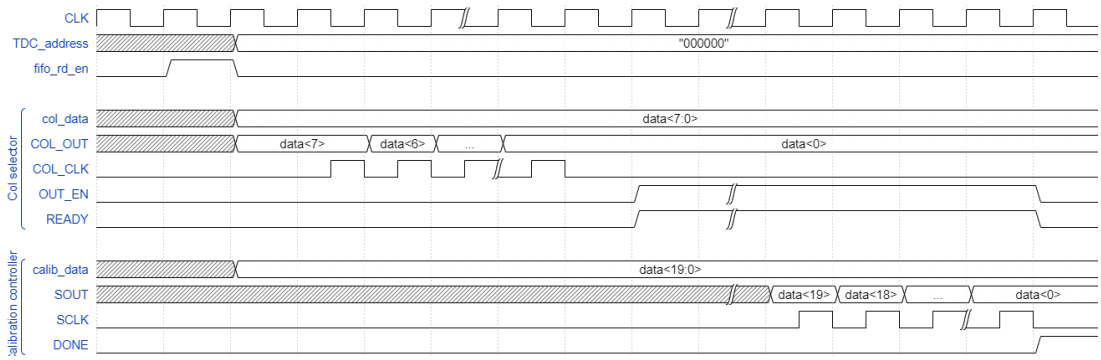


Figure 3.16: Simplified timing diagram of the TDC calibration component, focusing on the serial communication and the combined work of the FSMs. The first block of signals refers to common ones; the second to the row/column selectors ones (both components work in the same way); the third to the calibration controller. Once the data are loaded on the row/column shift register, the output enable is activated. The calibration data are then passed serially by the calibration controller. When the transmission is done, the corresponding flag activates. If another TDC must be calibrated the procedure restarts, otherwise it stops.

3.2 Graphical User Interface

A simple Graphical User Interface (GUI) was developed to perform in a semi-automatic way the tests described so far. It was written in C++ with the help of *Qt*, an open-source software embedding an easy widget toolkit, to create multi-platform applications.
The structure of the application is composed of a main window, shown in Fig. 3.17. It includes several push buttons that open as many different panels, each providing additional controls on the FPGA. The test structures discussed in Sec. 3.1 can all be controlled and observed with some of these dialog windows. Their functions and implementation will be briefly discussed later.
Some of the buttons, however, will not be commented, either because not completed yet, or because their original functions are included in other windows.

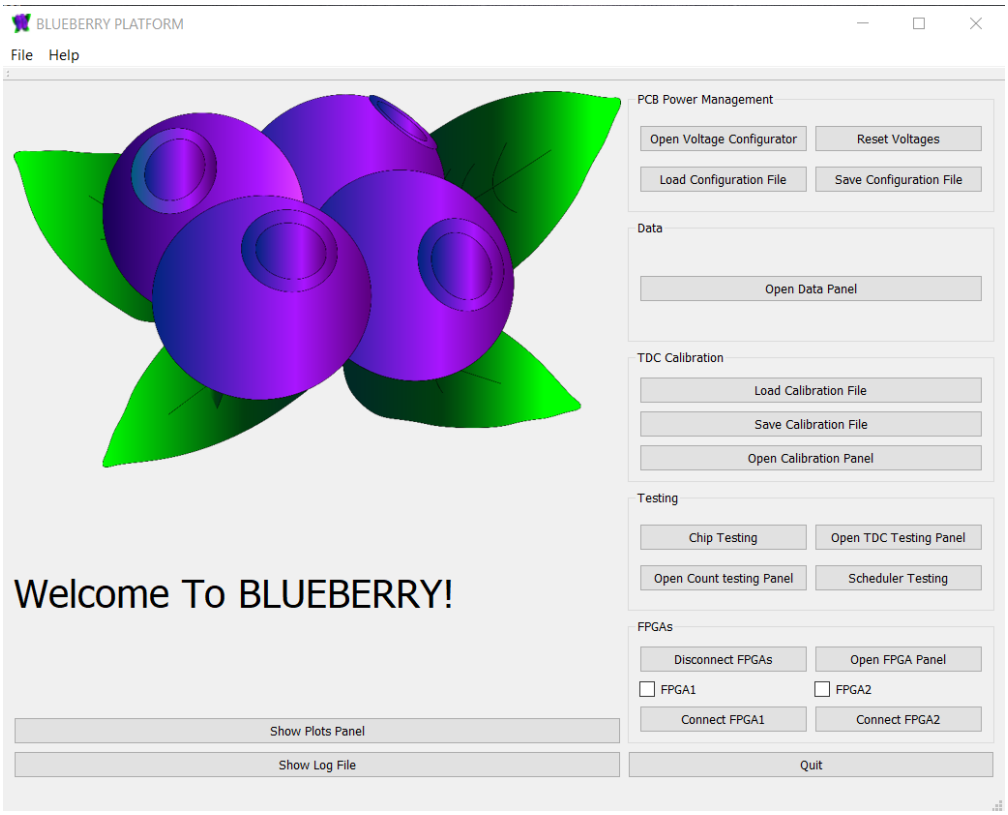


Figure 3.17: Blueberry GUI developed for the chip testing. The different functions will be explained later, with dedicated sections for the most significant ones.

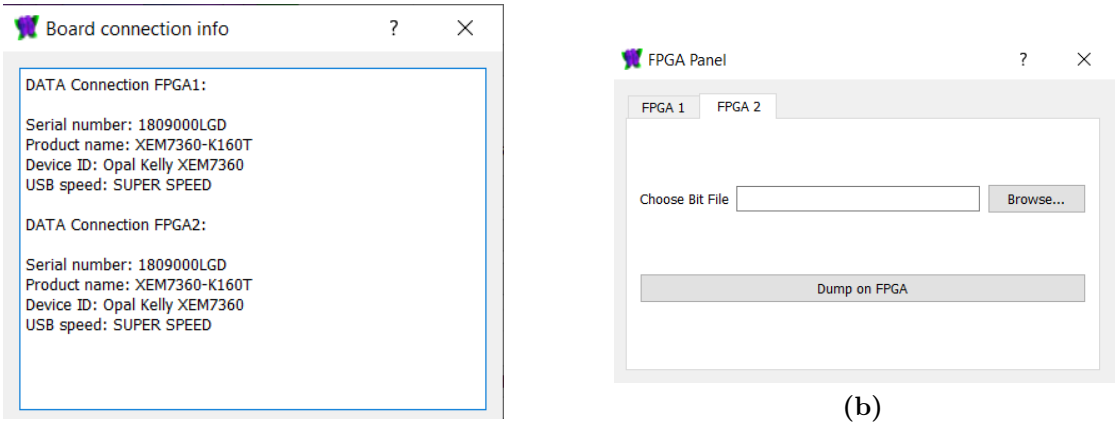
FrontPanel Software Development Kit (SDK) provides the basic functionalities to configure, control and interface with the FPGA hardware and peripherals from

the application. The API was employed to include FrontPanel benefits in our custom applications like device enumeration, FPGA configuration and communication. FrontPanel API contains several methods to interface with the device via USB, especially to directly communicate with the HDL modules in the design like wires, triggers and pipes (see Sec. 3.1). Virtual buttons can be added and connected to points in the design to provide the proper level of control, without increasing the number of physical resources. Also observability is enhanced, because we can have real-time information on signals in the design. FrontPanel allows us to have additional (virtual) I/O to the design at will, without using pins of the FPGA.
The API is provided as Dynamically-Linked Library (DLL) that can easily be included in our application. The library has few necessary classes already implemented that are instantiated in the code to interact with the FPGA. Additional information on the methods employed can be found in the Front Panel API class reference [75].

3.2.1 FPGA connection and configuration

The first basic function to implement is the connection and configuration of the FPGAs. The connection occur directly on the application main window through the "connect FPGA" buttons in the lower part.
Once the devices are connected to the PC via USB cables, we can enumerate their number with the method *GetCount()* of the FrontPanelDevices class. Then the serial number is retrieved with the method *GetSerial()*. An iteration over the serial number is done to find the specific serial number of the two FPGA: this step is required because every FPGA is connected to different pins and has different functions. It is then important to load the correct bit files in the proper FPGA identified by its serial number. While iterating, if the correct serial number is identified, the corresponding device is opened with the method *OpenBySerial()* of okCFrontPanel class. One object of this class will be created for each FPGA connected, to be able to interact with them separately. The connection is now complete.
A visual information on the connection status for both board is given in the main window through two check-boxes. A timer checks periodically if every device is still connected and update the check-box automatically.
Once the available FrontPanel devices are open, their information are retrieved using the *GetDeviceInfo()* method and saved in a object of the okTDeviceInfo structure. The information includes the device serial number, its product name, its ID and the speed of the USB connection (it should always be super-speed in our case).
All these data are written on a log file that can be visualised and read clicking on the "show log file" button, in the main window. An small dialog window appears,

as the one shown in Fig. 3.18a.



(a) **(b)**

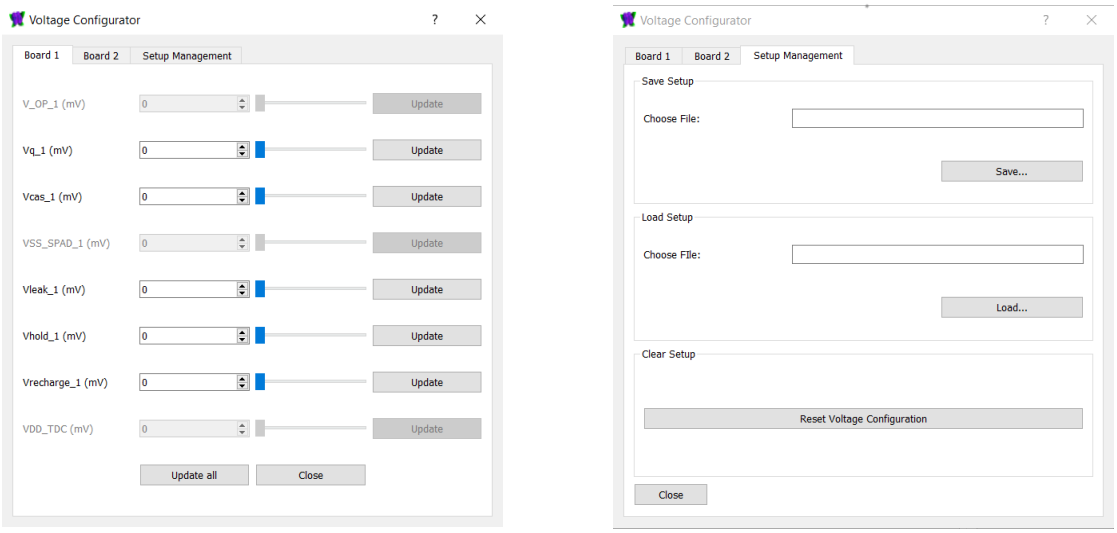
Figure 3.18: (a): Log file panel with the information on the two available devices. (b):FPGA panel for the devices configuration. Two tabs allow the selection of the FPGA. A bit file can be selected with the browse button and then dumped on the device with the corresponding button.

The configuration of the devices implies the opening of a simple secondary panel,in response to the pressing of the "Open FPGA panel" button. The dialog window shown in Fig. 3.18b appears. A tab is used to separate the configuration of the first and the second FPGA. In each tab a "Browse..." button can be used to search for the proper bit file to load. Once the file is selected, it is possible to configure the device. A single method from the okCFrontPanel is required, *ConfigureFPGA()*, specifying the bit file as argument. It returns some standard error code if something fails, so a check can be made: some messages boxes will inform on the task correct execution or failure.

3.2.2 Voltage control

One of the FPGA controls several voltages of the pixels in the chip. These voltages are used to bias several transistors that greatly impact the functioning of the devices. In particular some of them take charge of the SPAD discharge (quenching) whereas some others have effects on the active recharge of the detector. The effect is to modify the SPAD output pulses and, as consequence the dead time. That is why all these voltages must be set at a proper value before beginning any type of measurements.
The corresponding panel opens upon pressure on the "open voltage configurator" button, in the PCB power management section of the main window. in this panel,

two tabs are present because they are used to set the voltages of the two halves of the whole chip (see Sec. 2). However, only one of the two FPGAs is employed for the voltage management; on the contrary, for the control signals and data transfer, each device takes charge of one half of the chip. The look of this panel is shown in Fig. 3.19. The second tab, "board 2" is simply the copy of the first one.



(a) **(b)**

Figure 3.19: (a): voltage control tab. Every voltage can be set either with the slider or the spin box. The values can be updated singularly or all together. Some buttons are disabled because the corresponding voltage is set externally. (b): setup management panel to save and/or load a given voltage configuration. From here is also possible to reset all the voltages to zero.

The values set in the tab are stored in a object employed to handle these parameters. The values in mV can be set either with the spin boxes or with the sliders, and each one has a fixed maximum (1.8 V or 3.3 V, depending on the type of transistor it is controlling). The single "update buttons" can be used to store and send the values independently; the "update all" button works for every voltage instead.
Before sending the data to the FPGA, they are properly converted from mV to binary 16-bit strings. All the data are sent through Wire Ins in the design where they are processed by a component. Each string is included in a 32-bit word with some additional control bits and then transferred to the respective endpoint. The method employed is *SetWireInValue()* that requires the data as unsigned integer and the endpoint address as second argument. The specification of the address

allows us to update the values independently, if necessary. To transfer the data to the FPGA *UpdateWireIns()* method is called: all Wire Ins are updated at the same time by the PC.
From Fig. 3.19a it is clear that some buttons are disabled: V_{OP}, V_{SS_SPAD} and V_{DD_TDC}. The first two are simply set externally with a power supply, while the third one can be tuned with a potentiometer on the hosting PCB.

Some additional useful functions are implemented in the last tab: "setup management". Here we have the possibility to save the voltage configuration or load a previously saved one. The "save" button allows us to create a new .csv file to store the voltage values displayed on the tab as integer. On the other hand the "load" button read one .csv file and set all the values for in the two tabs accordingly.
The last button in this tab forces the reset of all the spin boxes and sliders to zero. The same effect can be obtained pressing the "reset voltages" button in the application main window.

3.2.3 Chip testing

To perform the test of the chip, an independent dialog window is used. Every control signal and data discussed in Sec. 3.1.1 can be set from it. Clicking on the button "chip testing" in the testing section of the main window, the panel appears as in Fig. 3.20.

The first browse button allows us to select a .txt file containing the masking data for the whole chip. These are composed of 64 strings (one for each cluster) of 64 bits (one for each pixel) and are used to locally disable some selected pixels, e.g. the most noisy ones. In the present chip, tested in this work, a bug prevents the correct functioning of this mechanism so that all the pixels must always be active. A fix has already been implemented for future versions. Enabling all the pixels corresponds to sending logical ones to every pixel.
Once the file has been selected, the masking data can be sent to the FPGA. At this point, the data are already stored as char arrays in a local variable. The reason is that the transfer to the device occurs through a Pipe In. For USB 3.0 connections, pipe data is transferred over the USB as 8-bit words. Even if on the HDL side, the Pipe In has a 32-bit word width, on the PC side (API) the width is smaller.
Pressing the "send mask" button, a function is called where the *WriteToPipeIn(...)* method is used. In this case this method is sufficient to transfer the data to the design. Once the communication begins, it will go on until completion, so the FPGA must be able to accept all the data. A coupled FIFO, in this case, is the best solution, as explained before.
The called method returns the number of written bytes or a negative error code

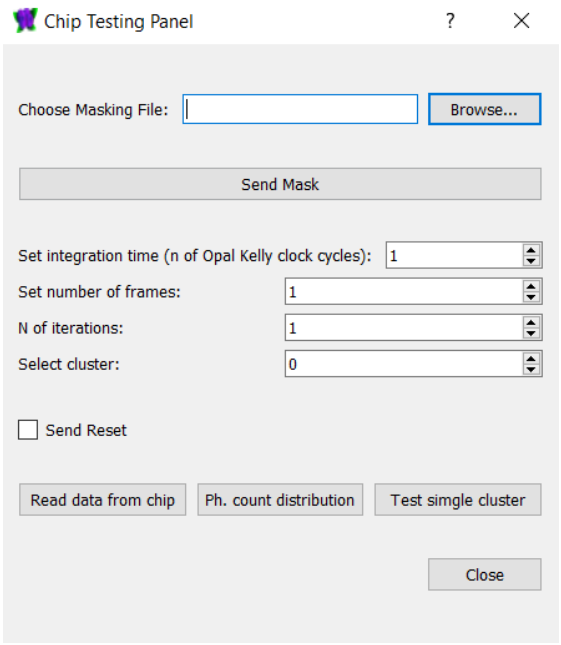


Figure 3.20: Chip testing panel. An explanation of the different functions is provided in this section.

if unsuccessful, so that a check for errors is easily done. If no failure is detected, the start signal for the FSM in the dedicated component is asserted. ActivateTriggerIn(...) method is called to this purpose, passing its address and the specific bit to activate.

After the mask is set, we can deal with the proper chip testing. Three bottons are present at the bottom of the window, depending on which some of the spin boxes must be set.
The first one is "read data from chip" button that is employed to test the chip following the procedure explained in Sec. 3.1.1. The component on the FPGA requires the specification of the integration window width and number of frames to capture. The first two spin boxes serve this purpose. The third one is useful to repeat several times the measurements, allowing us to increase dramatically the statistics. When the button is pressed the values of the first two spin boxes are sampled in two variable and converted to standard strings. They are all later merged in one single 32-bit word with some other control bits, e.g. a reset bit, asserted if the "send reset" check box is checked. The 32- bit word is sent through a Wire in to the design. Again the two required methods are *SetWireInValue(...)* and *UpdateWireIns()*.
This happens automatically in few seconds, after a dialog window appears, to select

or create a folder for the several output files generated. Once the folder is selected the test can start: another trigger is activated to start the FSM. After a brief break in the program execution to let time at the test to finish (few ms), the Wire Out are updated with *UpdateWireOuts()*. The value of the specific endpoint is read after with *GetWireOutValue(...)* method. This is done to check if two bits in this Wire, informing whether the mask and the readout are complete, are asserted or not. If they are, all the data are read from the output FIFO.
The output string from the chip is composed of 36 bits: 18 bits represent the TDC code; 6 bits are the photon count; 6 are the cluster address; the last 6 bits are the address of the first SPAD that fired. These data are transferred in a 64-bit string for convenience (Piep Out data width must be multiple of 32).
At this point, the data are processed on the software. Firstly, one check is performed on the bits corresponding to the TDC data: if they are zeros it means that no photon was detected and the corresponding photon data is forced to zero. Otherwise the photon count data is stored in a variable. Moreover a 64×64 matrix (representing the pixels) keeps count of the occurrence of the SPAD addresses.
Another check performed regards the cluster address. Due to rolling shutter readout, it should increase by one in every data string. If it doesn't happen probably an error occurred and a variable is increased.
Finally, the data are saved in the output files. The photon counts are saved as .csv files where every row represents one acquisition frame. Every row includes the counts for every cluster so it will have 64 values. At the same time, these data are saved in a shared object representing a 3D matrix. Two dimensions represent the square matrix of cluster (8 by 8) on the chip while the third dimension is used to store following frames. This matrix will then be used to directly plot the map of the photon count in another panel of the GUI, as explained in the following section Sec 3.2.4. The pixel occurrence matrix, instead, is automatically updated after every frame, so it is saved as a single 64 by 64 matrix in a file at the end of the measurement.
This happens for every iteration, so the number of output files can be quite large. If necessary the raw output data can also be saved in file just by changing few lines of code. In this case, one additional file for every frame will be generated.

The second button, "ph. count distribution" works differently. The purpose of this test is to retrieve the distribution of the total number of counts in one frame. This is an information that can be used to retrieve the energy spectrum of a radiation source. This makes it a measurement specific to the target application (ToF-PET). In PET The gamma rays hitting the scintillating crystal has a fixed energy so it should generate an almost fixed number of visible photons. When the crystal is placed above the chip, we should observe a peak in the photon count distribution, showing the number of photons corresponding to the gamma ray

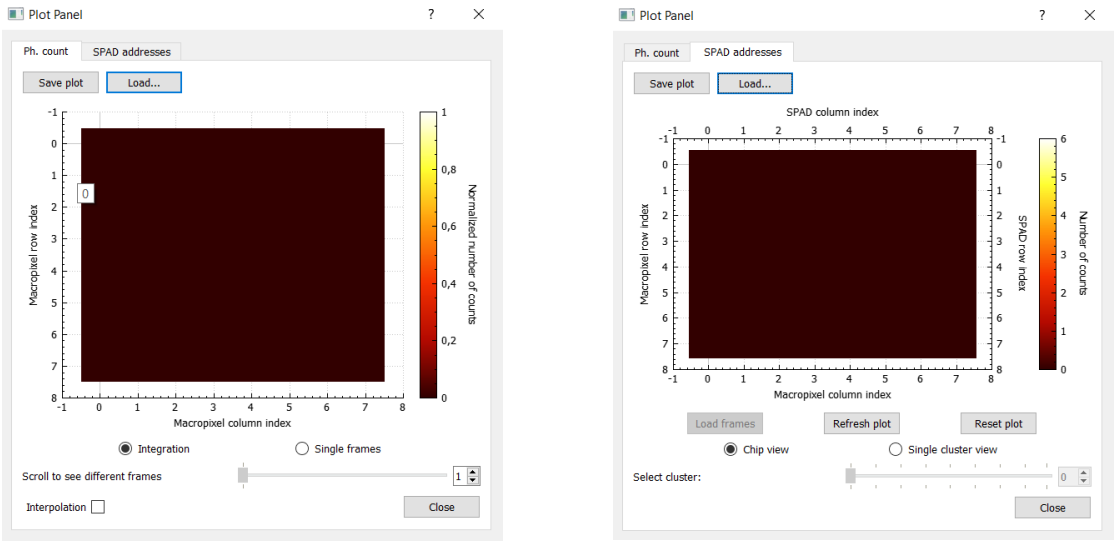
absorption by the crystal.
To retrieve this information test proceeds as exactly as explained above, but the data are not saved anymore to speed up the measurement. For every acquisition frame, the total number of photons is computed, then the count histogram is updated. Once the measurement has finished, the histogram data are saved in one .csv file. The problem with this measurement is that millions of data are required so many iterations must be done. For this reason, the code was optimized to make it as fast as possible, saving the final data only at the end of the measurement.

The last button, "test single cluster", works exactly as the first one. The test procedure is the same apart from the double sampling scheme explained in Sec. 3.1.1 and some minor differences. The first one is the maximum integration value, much higher than before. To represent it a dedicated Wire In must be used to send it to the design. In addition, also the chosen TDC address must be specified in this case. The last spin box is used for this purpose. Its value is transferred to the device with another Wire In, together with the number of frames and the same control signals as before.
Another difference is the direct consequence of the double sampling scheme: twice the data must be read from the FPGA, so two FIFOs and the corresponding Pipe Out are employed. All the data processing occurs in the same way but now the output files will also be doubled. The increased number of data and file makes understandable the choice of focusing on one single TDC at a time.

3.2.4 Plot panel

After a test on the chip is performed, it is useful to directly verify the validity of the results. The best choice is to implement a dedicated panel on which few basic plots are automatically drawn. This panel is implemented in our GUI to show maps of the chip with different information. To include it directly in the Qt-developed application, the dedicated QCustomPlot library is employed. It can be opened with the button "show plot panel".
The panel is formed of two different tabs shown in Fig. 3.21.

The first panel displays the total photon count value for every cluster of the chip. When several frames are acquired in one measure, all the counts are added together. The plot represents the 8 by 8 matrix of macropixels, each coloured accordingly to the scale on the right. All the values are normalized with respect to maximum value of counts possible, i.e 63 times the number of frames. The data for this plot are taken directly from the 3D matrix object created during the readout of the output data, as explained in Sec. 3.2.3. To include very frame in the plot,



(a) **(b)**

Figure 3.21: (a): First tab of the plot window. It shows a 2D representation of the cluster matrix of the chip. For each cluster the photon count information is represented as color scale. The data are normalized with respect to the maximum count possible. (b): second tab of the plot panel. It displays the SPAD address occurrence over the whole chip, represented as 64 by 64 square matrix. The color scale here is not normalized and the values in this picture are default ones. When data are loaded, the colors are re-scaled automatically.

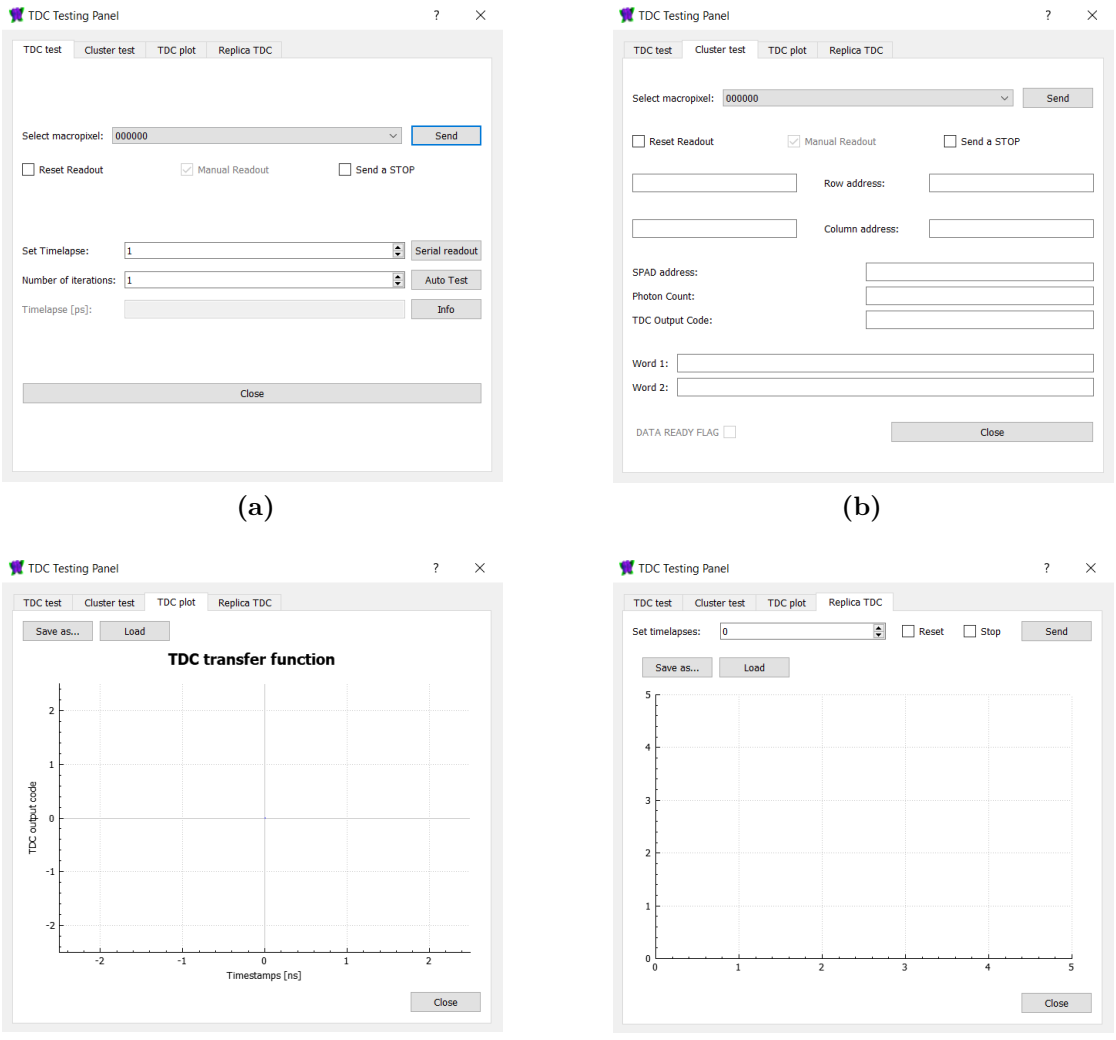
the data of each cluster are summed along the third direction of the matrix (it stores different frames acquisition). The plot in Fig. 3.21a does not display any data because no measurement was performed before, so the matrix is populated with zeros. Different maps will be shown in the next chapter, when discussing the results.
Several buttons are included in this tab. The "save plot" button allows us to save the graph as an image, with three possible format: .png, .jpeg or .bmp. This is easily done with dedicated method in the QCustomPlot library.
On the other hand the "load" button serves the purpose of loading the photon count data from some output files previously stored. When the dialog window appears, it is sufficient to select the appropriate directory containing all the output files of a previous measurement. Then the program fetches the correct file containing the photon counts matrices (see. Sec. 3.2.3) and stores them in the 3D matrix. Once the transfer is complete, the plot is updated to show the new data.
The checkbox on the bottom-left side specifies if the plot should use an interpolation when displaying the color map, instead of using a 1:1 data-to-pixel scale. In our

case we are interested in recognizing the single cluster so this function will rarely be used.
In the lower part, the "single frames" radio button allows us to display the same map of the chip but selecting just one of the acquired frames. This is possible thanks to the 3D matrix that stores the data separately for each frame. The slider and the spin box are enabled when the button is pressed and help in the selection of the desired frame.

The second tab is very similar to the first, but it is used to display the whole pixel matrix (64 by 64). This plot is used to show the occurrence of every SPAD address during the measurement. Every time a count is registered the address of the first SPAD of the cluster to fire is also stored. A explained in Sec. 3.2.3 the occurrence of these address is kept in a matrix that is later saved on a file at the end of the measure. This means that before drawing the plot, the data must be first read from the file. This type of information can be very useful to identify particularly noisy pixels on the chip: if a SPAD is always active, its address will always appear when reading the data coming from its cluster. This means that in the 2D histogram plotted, it will appear particularly bright.
The type of plot is again a QColorMap from the QCustomPlot library as in the other tab. The colour scale in this case is not normalized but directly displays the values. A singe cluster is read once every frame. Even if always the same SPAD fires, its maximum value of occurrence is given by the number of frames (max 2047).
Also in this window two buttons to save the plot as an image and load the data are present. The first button on the at the bottom is "load frames". In case several iterations are performed, additional data can be loaded with this button. The data are accumulated in a variable, but the plot is not updated yet. This means that several iterations can be included to increase the statistics, before displaying the new values.
To update the plot, the second button, "refresh plot", must be clicked. This re-draws the color map taking the updated data in input. Every time also the colors are re-scaled. The third button is used to clear the plot as well as the variable storing the data.
Also in this case a second view is possible: clicking on the "single cluster view" button the plot modifies. The displayed matrix becomes 8 by 8 because it represent the pixels in a single cluster. Basically the effect is a zoom over the matrix. The slider and the checkbox become enabled and help with the selection of the desired cluster.

3.2.5 TDC testing

Another dialog windows to describe is the one employed for the TDC test. The former is shown in Fig. 3.22: the picture displays all the 4 tabs of the window.



(a) **(b)**

(c) **(d)**

Figure 3.22: Picture of the four tabs in the TDC testing panel.(a): First tab for the TDC test. (b): second tab for displaying the information on the chosen cluster. (c): the TDC transfer function can be directly seen on this plot. (d): tab for the test of the replica TDC, currently unused.

The first one is used to send the TDC testing component the necessary data and

control signals. As explained before, the TDC test proceeds one cluster at a time so the first thing is to select the address of the macropixel. The topmost box serves this purpose. Its value is stored as a string and then passed to a Wire In, ready to be transferred to the device. Through the same Wire, also a reset bit and a manual STOP signal can be transferred, when the respective boxes are checked. The usual method *UpdateWireIns()*, corresponds to the beginning of of the transfer.
Once the TDC is selected, we can select the number of timestamps to send to the component for the test. The maximum number is in this case $2^{16} - 1$ as specified in the previous chapter. Clicking on the "Auto test" button, the value is sent to a dedicated Wire In. Right after, a dialog requests the choice of a folder for the output files. When this is done, the Wire Ins are updated, and through *ActivateTriggerIn()*, the FSM is started. A execution break of one second ensures that the component has enough time to complete the test (quite long, especially for many timestamps).
The data are finally read from the proper Pipe Out endpoint, and are automatically saved in a file, where every TDC output string is written on a single row.
The second possibility to test the TDC is to employ the serial output By clicking on the "serial readout" button. The main differences for this readout scheme, regards the implementation of the component on the FPGA, as explained in Sec. 3.1.2. From the software point of view, the maximum value of timestamps is here limited to $2^{14} - 1$. If an higher value is selected, when pressing the button, a message box will inform us about the error and the program will not start. Also the execution pause lasts 1 second more, because this method is obviously slower. The output data will be different, since they will not be processed by the "Output Processing Unit (OPU)" on chip that converts the thermal coding in binary strings. Everything else works as for the parallel readout.
The last button, "info", will open a small message box giving some information on the test. It defines the maximum number of timestamps and the clock frequency, depending on the testing procedure, and explains the value of a single timestamp.

The second tab is simply used to display the data coming from the cluster. In the current configuration is not enabled anymore, but was used in the first tests to verify that the cluster was responding correctly to the signals. The procedure was very similar to the one described above, with the only difference that a single timestamp was selected. For example, in correspondence with the label "row/column address", the sent address of the cluster and the one in output from the chip were compared. If they were different, some error occurred.

The third tab includes a QCustomPlot instance to directly display the TDC transfer function, as result from the performed test. Clicking on the "load" button, the the folder, where the output data are, can be selected. Reading the file, the

code plots the output codes of the TDC as function of the given timestamp. This is possible because every line of the output file represents a code for a given time interval. Moreover, every subsequent timestamp is 18 ps longer than the one before. Some of the resulting plots will be commented in the following chapter.

The last tab is currently unused. Its purpose is to include the function described in the first and the third tab in a single one, to test the replica TDC on the chip. In this case no address must be provided, and only the parallel readout in taken into account. The transfer function can then be immediately displayed in the plot below.

3.2.6 TDC calibration

The last panel left is the one use for the TDC calibration. The panel is shown in Fig. 3.23.

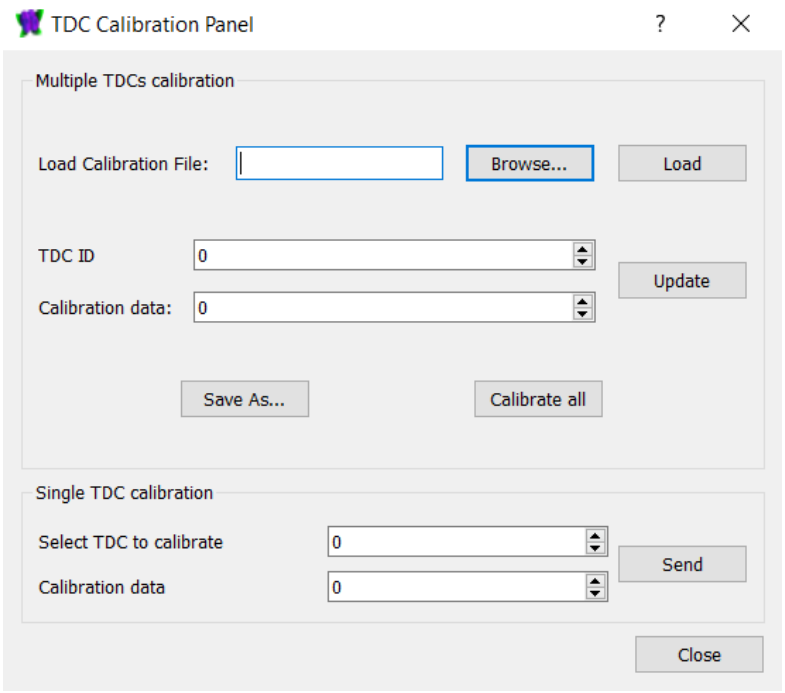


Figure 3.23: TDC calibration panel.

The easiest way to calibrate directly all the TDC is to load an appropriate file with the calibration string. This can be done with the usual procedure, clicking on the browse button and then the"load" one, once the file is selected. The format

of the file must be as follows: on each line, the calibration string is followed by a comma and then by the TDC address. To handle all the data, an object (a pointer called **TDCCalib*) of a dedicated class is used, where we can store the calibration strings and the relative TDC address in arrays. In particular, the load button stores the string and the TDC ID in the object. The same effect can be obtained using the the two spin boxes below. Here we can change the data independently and update their value in TDCCalib, clicking on the "update" button.
"Save as" button saves the current calibration configuration in a file, following the format explained above.
To finally send the data to the FPGA, the "calibrate all" button must be clicked. First thing resets are sent through a Trigger In, for both the input FIFO and the FSM, with *ActivateTriggerIn()*. The data are then written in a single 32-bit string. The 6 LSB are populated with a decreasing binary number starting from 63. This number will inform the FSM on the remaining number of TDC left to calibrate. The last data passed to the component will have 0 stored in these bits. This will automatically inform the FSM that the calibration has finished.
The data transfer occurs through another Pipe In endpoint. It is important to consider that the transfer over the USB occurs again in 8-bit words. The first transferred byte will be written on the LSB of the data bus on the FPGA. This means that data must be passed to the Pipe in starting from the LSB. After some microseconds, the trigger for the FSM start is activated.

The second part of the panel is occupied by the "single TDC calibration" section. The working principle is the same as before but here the single TDCs are calibrated with random access approach. Since only one TDC is considered, the 6 LSBs will be all zeros. Once the TDC address and the value are selected, the "send" button can be clicked. From this point the code is the same as before.
One important difference is that with this method the TDCCalib object is not updated anymore, but data can only be directly sent to the FPGA.

Chapter 4

Results

4.1 TDC characterization and test

Obtaining the TDC transfer function is necessary to understand, first of all, if it is active and working. From the slope and the step-wise trend it is also possible to identify the LSB, i.e. the resolution of the TDC and confront it with the expected one.
This was one of the first test performed, also because the bottom tier of the chip is sufficient, so the setup does not require many instruments. The testing system, developed on FPGA, independently provides the start and stop signal to the TDC, as discussed in Sec 3.1.2. This means that the circuit does not have to be connected to the SPADs, placed on the top tier, to work. Only the bottom tier of the chip is then sufficient to test the TDC. Such samples were available already wire bonded on a small PCB, ready to be connected to the large one, including connections with the two FPGAs. Only one half of the chip is tested at at time so only one FPGA is necessary. The power to the chip come directly from the power cable connection of the XEM7630 so no additional instrument was needed.
As explained in the section dedicated to the TDC testing component, different design solutions were explored. The first and simplest one includes the implementation of a counter. The single timestamp corresponds to a clock period, in this case of 10 ns. The resulting plot is shown in Fig. 4.1.

As expected the typical, linear trend of a converter is found. Beside this promising result, another important feature shown in this plot is the limit of the TDC dynamic range. After 4 µs the trend is clearly interrupted and re-start right after. The reason is the overflow of the TDC loop counter for such an high timestamp value. This is supported by the fact that for the next timestamp, it restarts from zero.

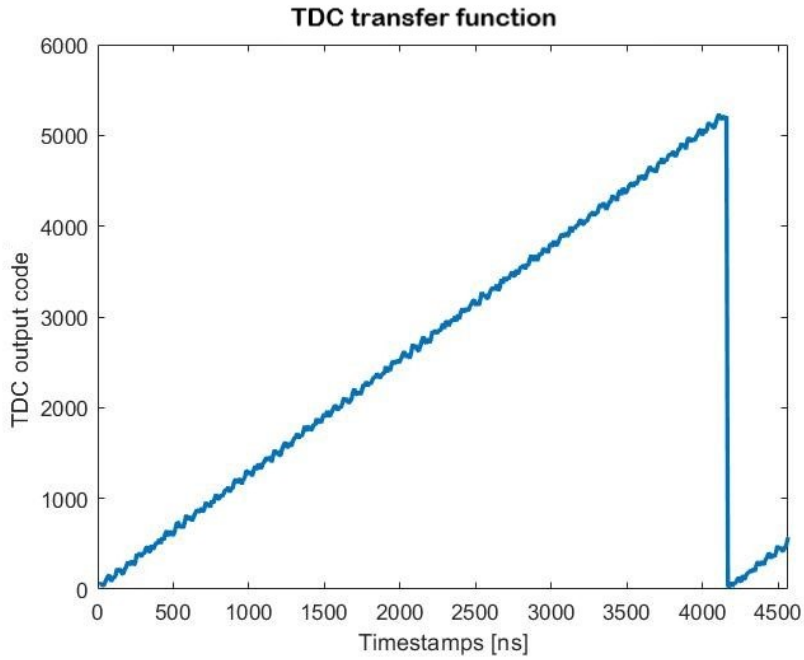
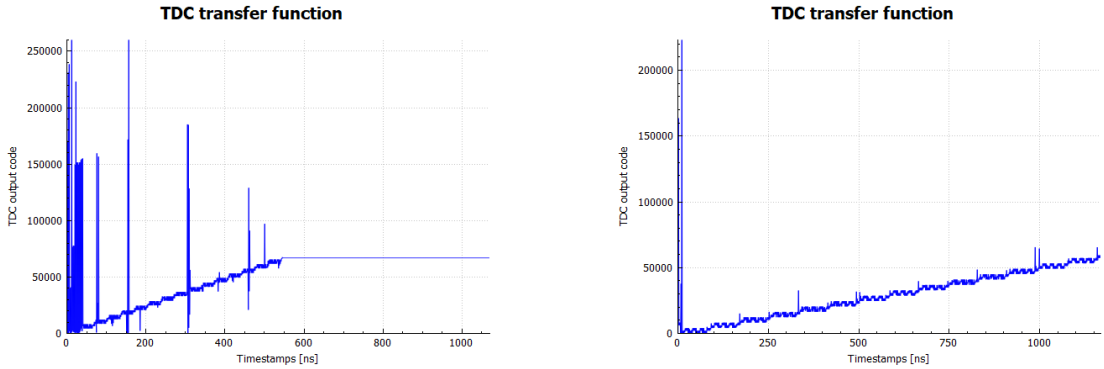


Figure 4.1: TDC transfer function. The used implementation for the definition of the timestamps is a simple counter on the FPGA. The pulses are coarse but all the dynamic range of the TDC can be observed.

On the other hand, focusing on the linear section of the plot, the stepwise trend is not clear yet. All over the range, many fluctuations are present instead. These can not represent the expected steps because their width is at least of tens of ns, whereas the LSB is expected around 20 ps. The fact that the TDC was not calibrated probably generated this weird trend. In addition, such a coarse test, with such large timestamps, can not pretend to visualize the correct LSB of the TDC. The main reason to perform it, was to verify the activity of the TDC, the correct trend of the transfer function and to define the device's dynamic range.

More precise measurements required the implementation of another solution, able to provide sufficiently small shifts. The dynamic phase shift feature of an MMCM resulted the most promising one. The next tests were performed with a unit timestamp of 18 ps, plotting the data on the dedicated panel of GUI (see Sec. 3.2.4). At the beginning, we obtained the plot in Fig. 4.2a, selecting a total number of 65000 shifts, to reach a final timestamp larger than 1 µs ($18\,\text{ps} \times 6500 = 1.17\,\mu\text{s}$). The plot presents several issues. First of all half of the function becomes constant as if the read output data is always the same. This was very likely related to an error in the data readout circuit. The second problem is the obvious presence of

unexpected, very high spikes along the function. The first solution was to review the readout component to check for some errors in the communication between chip and FPGA, or between FPGA and computer.
The right solution appeared when the readout clock frequency of the component was halved. The resulting plot is shown in Fig. 4.2b. Two idle states were already included in the FSM from the beginning, precisely to avoid metastability of the output data (for more details, see Sec 3.1.2). Nonetheless, the time interval before sampling the data was apparently not enough in the first configuration. From this analysis it was roughly estimated that at least 20 ns are required to the data to become stable, avoiding any related issue. With the new clock frequency, 30 ns pass before the output data are registered and stored.

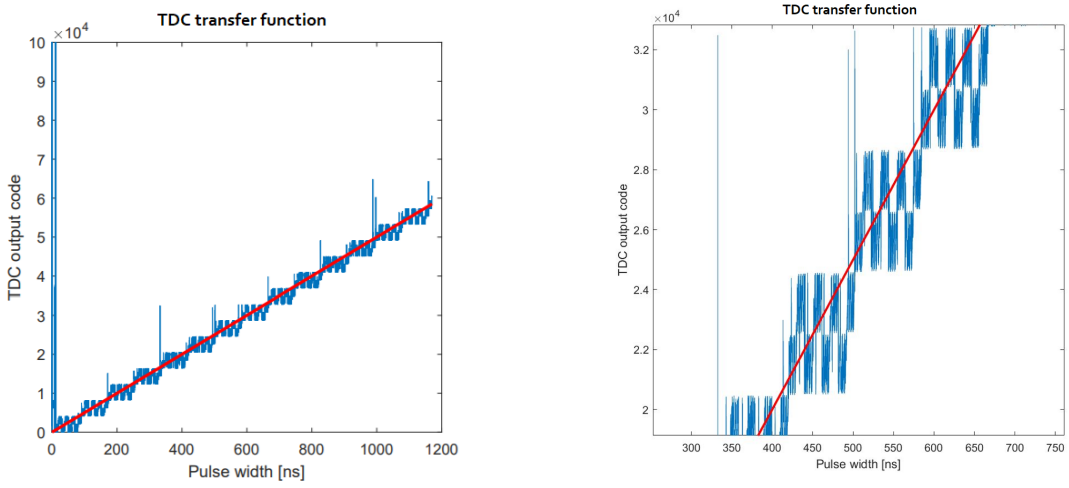


(a) **(b)**

Figure 4.2: TDC transfer functions. With a total number of timestamps of 65000, and a readout clock of 200 MHz, the plot presents unexpected spikes, and a constant trend for half of the range (a). With the same input data, but slowing down the clock to 100 MHz, the resulting transfer function appears much neater, even if some spikes are still present. The problems were most likely related to data metastability (b).

In this new plot we can notice how the reduction of the clock frequency resulted in a much neater trend. Some spikes are still present, but they are much smaller than before, apart from the initial ones for small timestamps. Some of them can be considered acceptable random errors in the transfer function; others might disappear after a correct calibration. Unfortunately the calibration requires a precise study of the device to find the right combination of bits (see Sec. 3.1.3 for further explanations). This was point has not been faced yet, but it certainly will be for future tests and further improvement of the TDC characterisation.
Besides the spikes, the correct linear trend is clear and can be seen in Fig. 4.3a.

This is a further verification that the TDC is working. On the other hand, another displayed issue is the width of the steps and their internal periodicity. This feature is displayed more clearly in the plot in Fig. 4.3b, representing a zoomed section of the one on its left.



(a) **(b)**

Figure 4.3: Linear fit of the TDC transfer function of Fig. 4.2b (a). Zoom on the TDC transfer function highlighting the periodicity on the steps (b).

The steps in this case are obviously not the ones we were expecting, defining the LSB of the oscillator. However, the linearity and the fact that a periodic pattern is also clear in each step, might suggest that there is a problem in the ordering of the bits in the output string from the chip. They are manually reordered on the FPGA, but only a careful check of the circuit design and the bit's routing can confirm this hypothesis.
With the first implementation of the counter (see Fig. 4.1) this issue might also have generated the observed fluctuations on the linear trend. Furthermore, because the timestamps were coarser, only few, different bits in the output string should change from one shift to the other (likely, the most significant ones). This could result in a neater linear trend, with no periodic pattern or steps.

In summary, it looks like the remaining issues in the test might result from a corruption of the output bits, rather than from the TDC itself. If this is the case, two possible error sources can be identified. The first is the OPU that process the raw data from the TDC directly on chip, and outputs them as 18-bit binary strings. The second is the readout scheme. In both cases a viable solution to solve the problem, is to read directly the TDC raw data from the serial output,

excluding the OPU. Although the readout becomes slower and the data have to be post-processed, it would help to find the main issue. This alternative is currently being developed starting from the dedicated component already implemented, as described in Sec. 3.1.2.
Another planned verification is a "single-shot" test. A defined timestamp is sent many times to the TDC and the output data are analyzed. If the circuit is working properly, all the data should follow a thin Gaussian distribution, centered on the correct output.

4.2 Photon count measurements

The activity and the performance of the chip as light detector can be assessed analyzing the photon count under different conditions. One of the first parameter to consider is the DCR, i.e. the rate of detection under no illumination. The results in this case are worse than expected, but show a promising coherence with the theory. For example the temperature dependency shows a correct trend. On the other hand, the results under illumination with a laser or a LED, change visibly. Although some issues are present, preliminary PDE measurements were also performed, again with encouraging results. After all, the planned tests aimed at a preliminary, qualitative characterisation of the device.
In the following sections, the most significant results are presented. They are all coming from one of the available chip. The choice was done to avoid the most noisy ones as explained in the following. In addition, several devices were visibly damaged and broken, and were therefore discarded.
One of the main issue for these kind of test was the malfunctioning of the counters. We discovered that a variable offset was always present at the beginning of the measurement. No reset seemed to work to bring that value back to zero. For this reason we were forced to adopt a double sampling scheme, to get rid of the offset. Many of the data in the following sections are taken with this readout technique so are limited to a single cluster (see Sec. 3.1.1, "Double sampling scheme" for more details). By contrast, when representing the data on the whole matrix, the standard readout was obviously used, to consider all the cluster. As consequence, in this case, the values are generally over-estimated.
As explained in previous sections, the measurements could include the acquisition of several frames. This can be useful to increase the statistic of the data and reduce their fluctuations. A first check can be to verify how many frames are necessary for the data to converge. Twelve clusters were selected and tested with two different integration window. The resulting gradient of the mean value is plotted in Fig. 4.4.

The plot represents, in fact, an average value: the gradient of the movable mean

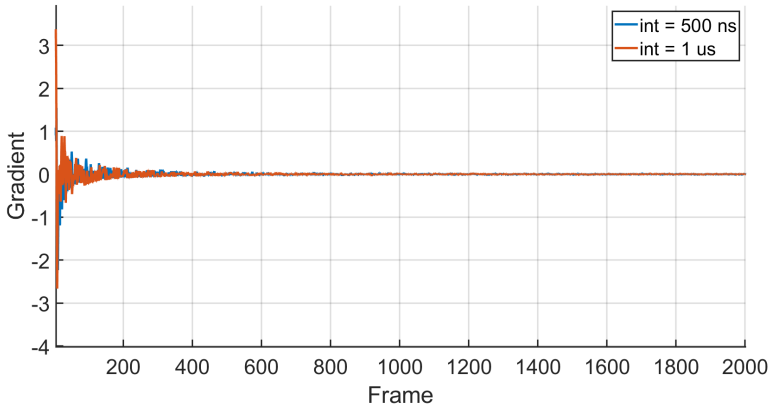


Figure 4.4: Gradient of the mean value as function of the number of frames with two different integration window. The mean value shows a good convergence already after few hundreds of acquired frames.

over the acquired frames of twelve different cluster were extracted to compute it. The graph shows a significant convergence already for few hundreds of frames for both the integration values. After one thousand frames the mean value can be considered completely stable. For shorter integration window the mean is expected to fluctuate more, so in general more frames are acquired, For the following tests, 2047 frames are always taken anyway, to avoid any possible convergence problem.

4.2.1 Dark count and DCR

Initial dark count evaluations were performed to verify the activity of the detector. More importantly, they helped with the fast recognition of failing chips and noisy ones. Qualitative plots could be directly displayed on the GUI, making the procedure much simpler and faster. Even broken chips were quickly tested on both halves, to verify that the damage led irreversibly to a failure.
Fig. 4.5 represents the cluster map of six different devices, tested under the same condition to confront their activity. They are selected between both halves of the three most promising chips. These first results drove the choice of the designated chip for the next tests.

The test is performed setting the integration window to 100 ns and the excess bias to 2 V. 2047 frames are acquired in total, and the counts for each of them are added together. The colour scale is normalized with respect to the maximum count possible for each cluster, corresponding to $63 \times$ [number of frames].
The first evidence is that the chip represented in Fig. 4.5b is not working: no

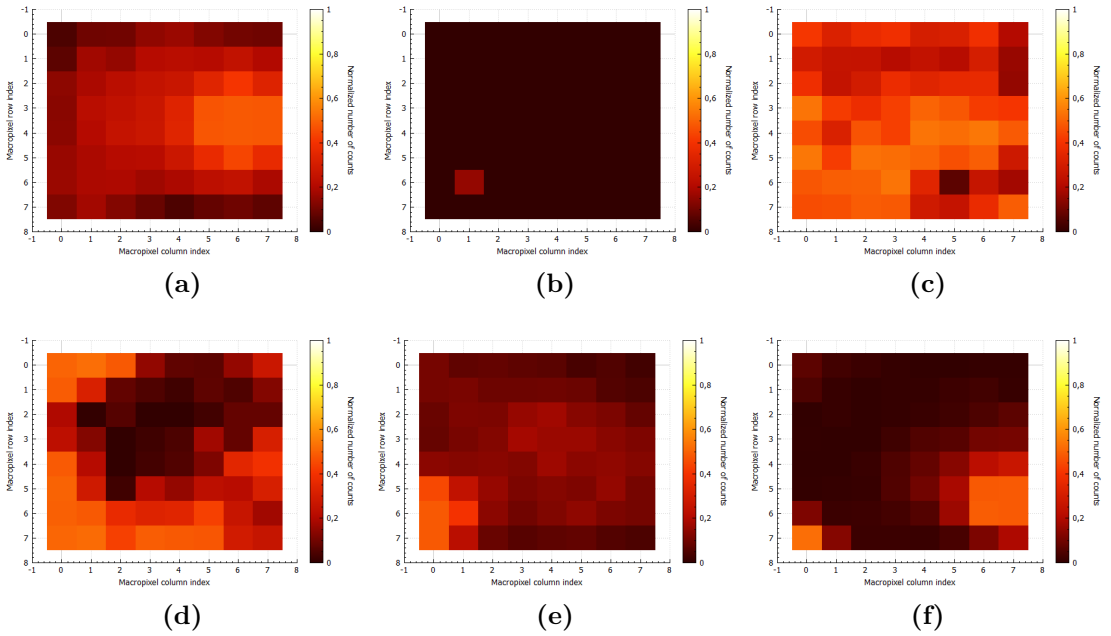
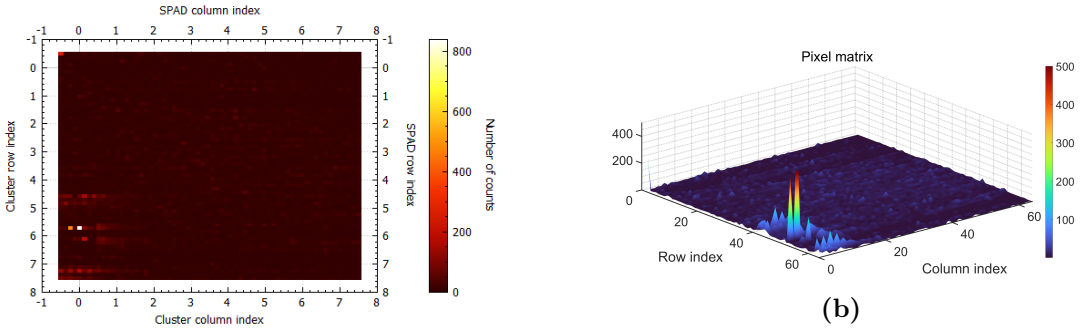


Figure 4.5: Cluster maps representing different chip's activity under no illumination, i.e. the dark count. The integration window is 100 ns long, the excess bias is 2 V and 2047 frames were acquired in total and the counts are added to form these maps. The data are normalized with respect to the maximum count possible. Chip 5, pad 1 (a). Chip 5, pad 2 (b). Chip 4, pad 1(c). Chip 4, pad 2 (d). Chip 2, pad 1 (e). Chip 2, pad 2 (f). The Plot (b) shows an inactive chip. The others are generally very noisy, but working. The ones (e) and (f) are the most promising ones.

photon counts are recorded apart from one single cluster that is evidently always active. By contrast plots (a), (c), (d) result very noisy even at such a moderate excess bias and integration window. The most promising ones are the two last ones, representing both halves of chip number 2. The dark count in the first half, or "pad 1", in Fig. 4.5e, is generally more uniformly distributed on the map, apart from four visibly noisy cluster. On the other hand, Fig. 4.5f shows a lower dark count all over the chip, but presents two very active regions. In total, five clusters look problematic, whereas, in the previous one, only four.
In the end, the first pad of the chip number 2, whose cluster map is displayed in Fig. 4.5e, was the designated one for the upcoming tests.

A further insight of the chip activity can be taken analysing the SPAD's address matrix. Every time a cluster is read, the address of first SPAD to fire is sent in

output. Collecting several frames, a 2D histogram representing the 64x64 pixel matrix is drawn. In each point, the occurrence of the corresponding SPAD is stored, as previously explained in Sec. 3.1.1 and 3.2.4. This allows us to appreciate the activity of the single pixels and identify the noisy ones. The noisy clusters from the previous plots, in fact, can originate either from a high activity of all its pixels or from few very noisy SPADs (even a single one). The identification of noisy pixels is important because they can individually be shut down with the masking system, once its bug is corrected.
The corresponding plots for the chosen chip are displayed in Fig. 4.6 where the integration window and excess bias are the same as before (100 ns and 2 V respectively).



(a) (b)

Figure 4.6: Pixel matrix corresponding to the chip 2, pad1, taken in the same conditions of Fig. 4.5e (a). The noisy pixels are clearly visible and are located in correspondence of the noisy cluster previously identified. Same histogram represented in 3D to better appreciate the relative dimensions of the peaks(b). Apart from the noisy pixels, the remaining matrix shows a significant uniformity. The z-axis scale is in this case limited, to highlight also the lower peaks.

The first plot in Fig. 4.6a is the one directly obtained in the GUI panel. The noisy pixels are clearly identified and are located in correspondence of the noisy clusters of Fig. 4.5e. It is clear that, in this case, few noisy pixels contribute to all the registered counts. The same histogram is drawn in 3D in the plot 4.6b to better appreciate the relative dimensions of the peaks. In this case, the z-axis range is reduced to better distinguish the lower peaks. It is interesting to notice that the noisy pixels are generally distributed one close to the other. The reason might be a little cross-talk between nearby SPADs.
Besides the peaks in the noisy clusters, the rest of matrix appear uniform, as expected.

A first simple test to verify the correct functionality of the chip (and the testing

system) is the measure of the dark count as function of the integration window. For larger integration intervals, a larger number of counts is expected until the maximum value is reached. The cluster maps and the pixel matrices of the chip, taken with the same excess bias, are also displayed in Fig. 4.7 to highlight visually the difference between a short and a long integration window.

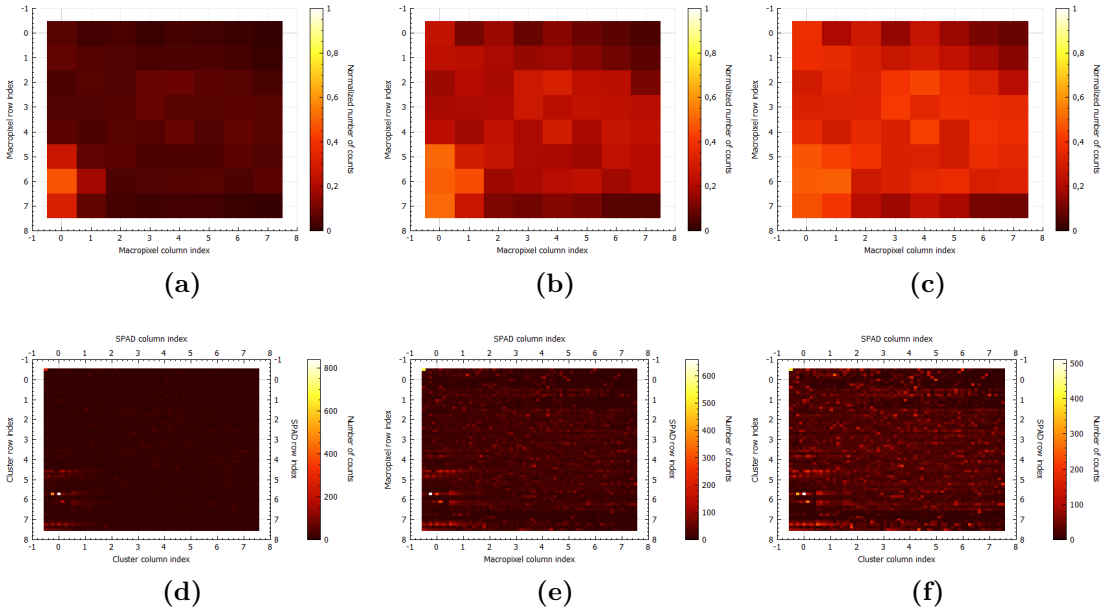
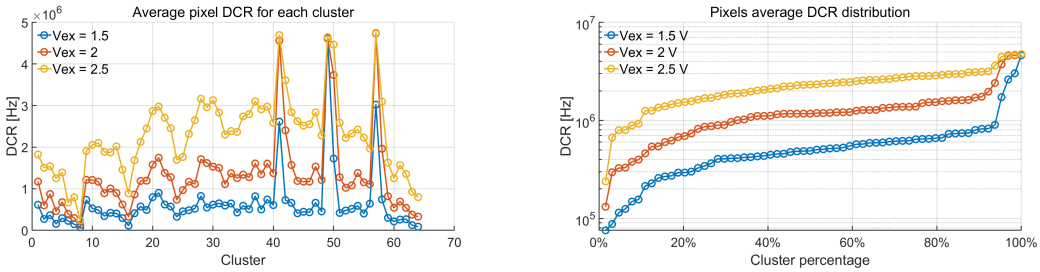


Figure 4.7: These maps are taken at V_{ex} of 1.5 V and increasing the integration window: 100 ns (a), (d); 500 ns (b), (e); 1 µs (c), (f). The second row of plots represents the corresponding pixels map of the plots above.

These plots are taken with a constant excess bias of 1.5 V, and the total number of frames is 2047. The integration window widths are, from left to right, of 100 ns, 500 ns, 1 µs. A nice evidence in the pixels matrices is the increasing number of counts, results in a higher activity of the pixels all over the chip. Another significant feature, that becomes visible for higher integration values, is that the noisy pixels mask all the others of the cluster. The reason behind is that, being so active, they are always the first to fire, with or without a real event detection.
The aim of these analysis is always to assess the DCR, a significant figure of merit of every detector. We can compute its mean value over multiple frames, knowing the photon count values for each cluster. From it we can also estimate the average DCR for each pixel of the cluster. For example, it is interesting to check how it changes from one macropixel to another: we know that in the chip there are four noisy ones, so we expect to find high values for the corresponding DCR. This

is what is shown in Fig. 4.8a. For each cluster the average DCR of its pixels is plotted.



(a) **(b)**

Figure 4.8: Average pixel DCR for each cluster of the chip (a). The noisy ones are clearly visible. Average pixel DCR distribution over the cluster (b). Most of the macropixels have a DCR of the same o.o.m, a fraction has lower or higher values instead. For each curve, four points have clearly larger value: they correspond to the four noisy cluster identified before.

Data taken with three different excess bias are considered. All the three curve in plot 4.8a share a common trend. The four noisy clusters identified below are clearly visible also here. Their DCR values are significantly higher than the others, as expected. The number of the cluster corresponds correctly to the ones displayed in Fig. 4.7 (the numeration starts for the top-left corner of the matrix and proceed row by row).
An additional result from these data is displayed in Fig. 4.8b. It shows the DCR distribution over the cluster of the chip. The DCR here is again computed for the single pixels and averaged on 2047 acquisition frames. The trend is coherent with the expectations: a fraction of the cluster, around the 10% depending on the curve, have a significantly lower value. The largest portion (60%-70%) presents similar values, of the same order of magnitude. Another smaller fraction, instead, has clearly higher DCR and it represents the noisy clusters. In particular the last four points of the plot must represent the four noisy macropixels identified before.
The main issues of these plots is the resulting, incredibly high DCR. The expected values should be not less than one or even two orders of magnitude lower. Even considering an extremely noisy detector, such high values of DCR might prevent the detection of photons. As already mentioned, some issues regarding an offset and a failing reset of the counters on the chip was discovered. This can obviously be one of the causes of these unusual values. In some cases, it might happen than even without any real count, a large number is stored in the registers, inevitably affecting the validity of the output data. Unfortunately, to avoid this problem

the double sampling scheme is required, but it can not be employed for the whole matrix.
This suggests us to move to the analysis of single clusters, also to have a more quantitative estimation of the dark count. Starting from this, longer integration windows and higher excess bias are set, also to appreciate the saturation of the counts. The chosen cluster is the tenth, given its noise level in the average (check plot 4.8a, but consider the 11^{th} cluster, because the numeration starts from 1 there). The obtained plot shown in Fig. 4.9, displays the dark count of the cluster over two different ranges, to better appreciate the trend.

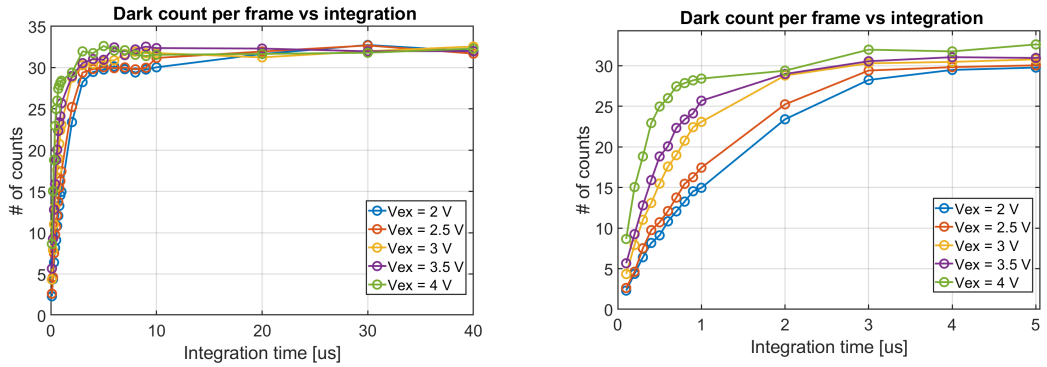


Figure 4.9: Dark count of the single cluster vs integration. The data are an average on 2047 frames, acquired with double sampling scheme on the tenth cluster alone. Different excess bias where used.

At this point we have the evidence that some problems in the counters are, in fact, present. The saturation, in particular, highlights that this component does not work as expected in that regime: instead of saturating around the maximum value, 63, it occurs around 30. This surely has to do with an error in the counting system. The fact that the counter, after reaching 63, overflows and restart from 0 is another issue to be taken into account. The double sampling scheme helps to deal also with it, because we can recognize when an overflow occurs from the relative values of the counts, before and after the integration. If the latter is smaller than the former, it means that the maximum value was reached and the count restarted. On the other hand, we are not able to recognize if multiple overflows occur. For example, in presence of a large number of counts, the same effect can results from two consecutive overflows, but we would not know it.
Nonetheless, the trend follows the expected one, especially the initial linear trend and the saturation regime. In addition, the effect of the increasing bias voltage is

significant too. A higher excess bias generally results in a higher number of counts and this is evident from the plot 4.9b. This also results in an earlier saturation, clear with a closer look at the data.
To have an additional verification and a comparison with the plots in Fig. 4.8a, we can easily compute the DCR of the pixels in this cluster. The result is shown in Fig. 4.10.

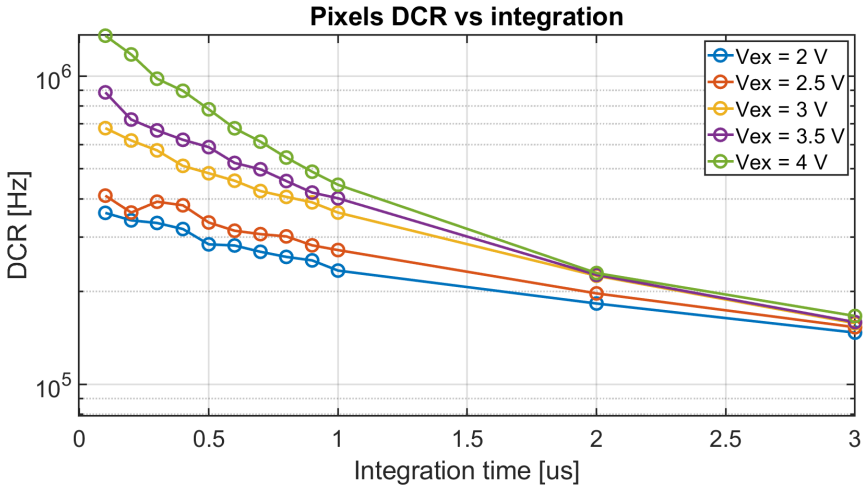


Figure 4.10: Pixels' DCR for different integration windows and excess bias. Only the tenth cluster is analysed here. In general the values are lower than the ones taken with the whole chip (see Fig. 4.8b)

The plot shows the pixels' DCR with different integration window and excess bias. It appears clearly an unexpected dependency of the DCR on the integration time that gets stronger for higher V_{ex}. This is again very likely due to the errors described before. Intuitively, the cause might be the saturation of the counts for longer integration window, whereas the value is expected to grow steadily. The fact that the effect is more significant for higher excess bias, where the saturation occurs earlier (see plot 4.9b) might support the hypothesis.
On the other hand, it is noteworthy that these values are much lower than the ones presented in Fig. 4.8b. In particular, looking at the data for the two lowest excess bias, the values remain always well below the million of cps. This does not seem to to fit with the much higher values of the cited plot. As anticipated, the use of the double sampling surely reduces significantly the level of DCR, mitigating the effect of the offset. Nevertheless, the values remain still very high, suggesting the presence of some bug in the counting system, as also expressed after the analysis of Fig.4.9.
In summary, some issues appear with the analysis of these last data. They will surely affect the performance of the detector, but only following measurements will

be able to assess how much. Besides the qualitative trends presented so far, further analysis will be necessary to understand precisely the problem, before being able to obtain reliable, quantitative data.

DCR temperature dependence

One intuitive solution to lower the noise level of the detector is to decrease the operating temperature. In this condition the DCR should fall steeply, because the thermal generation of carriers, one of the main source of noise, is dramatically reduced. For low temperatures, only field-enhanced generation by tunneling effect dominates, and the values reaches a plateau.
The test is performed in a dark temperature chamber with a fixed integration time of 1 µs. This higher value is set because the active recharge circuit was disabled, preventing possible oscillations in that branch, due to the masking circuit. As consequence, only the passive discharge of the SPAD takes place, which is much slower. A longer integration time is then required. The results with three different bias voltages are shown in Fig 4.11. The values represent the average DCR of the pixels of the tenth cluster, read with the double sampling scheme.

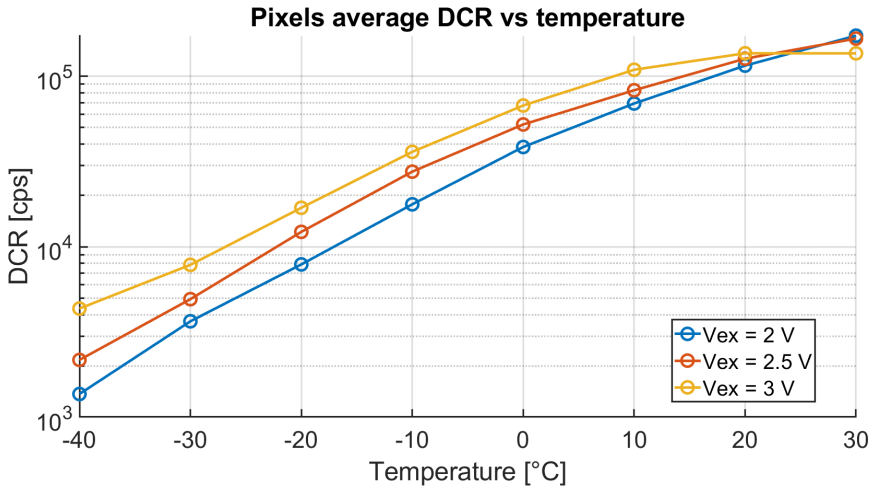


Figure 4.11: Average pixels DCR vs temperature for three excess voltages. The data are averaged on 2047 frames, acquired with double sampling, always on the tenth cluster of the matrix.

The plot shows a good coherence with the expected theoretical results. All the three curves display an exponential dependency on the temperature until −40 °C. The DCR decrease almost of two order of magnitude along the whole range. At higher temperatures, around 20 °C, the saturation of the counts occurs and is clearly visible when the excess bias is larger. The fact that the values are generally

lower than the ones presented before, is to ascribe to the longer integration interval. As we commented before, the DCR unexpectedly decreases with wider integration windows. Another reason is surely the passive discharge of the SPADs. A longer dead time of the pixel can reduce the total count.
Below $-20\,°C$ the values begin to show an acceptable DCR level, although still high. This analysis suggests that the right strategy might be performing the tests at that temperature, also under illumination. Lowering the temperature is, indeed, a good solution to lower the noise. Despite a quantitative analysis is premature because the values are not completely reliable, the trend confirms our expectation. In addition it also provides us with a significant hint for future tests to reduce the high noise level of the detector.

4.2.2 Laser illumination

After the analysis of the DCR, we move to the measurements in presence of a light source. The first idea is to check if the chip responds correctly in presence of external photons and if their are detectable above the noise level. The first tests were performed with a laser with the purpose to verify qualitatively if Blueberry can detect the incoming photons. The advantage of the laser is that it can be focused in a relatively small spot. If the spot is small enough, when it hit the detector only few clusters will activate. In this way there will be a clear difference in the count values between the background and the active clusters, provided that the dark count is not too high. On the cluster map, few bright macropixels will be highlighted on a uniform background. The solution to reduce the DCR was to perform the measurements at $-40\,°C$, as suggested before. This avoid that a too high noise level prevents the detection of photons.
A 517 nm fibre-coupled laser is employed for this test. The size of the spot in output from the fibre optic is further reduced, using a small diaphragm with several circular slits of different dimensions. The alignment was not easy given the small dimension of the spot, the slit and the target. Fortunately, the laser light in the visible range simplified the task a bit.
The first measurements are performed on the whole chip to obtain a cluster map at different laser intensities. The excess bias is fixed at 2.5 V and 2047 frames are acquired each time for 60 iterations. The integration window is fixed at 1 µs because, also in this case, the active recharge circuit is deactivate. The resulting plots are displayed in Fig. 4.12, starting from no illumination and increasing the light intensity from left to right, row by row.

First of all, it is noteworthy the difference of the maps in plot 4.12a at 2.5 V of excess bias and the one in Fig. 4.7c at 1.5 V. The comparison further explains

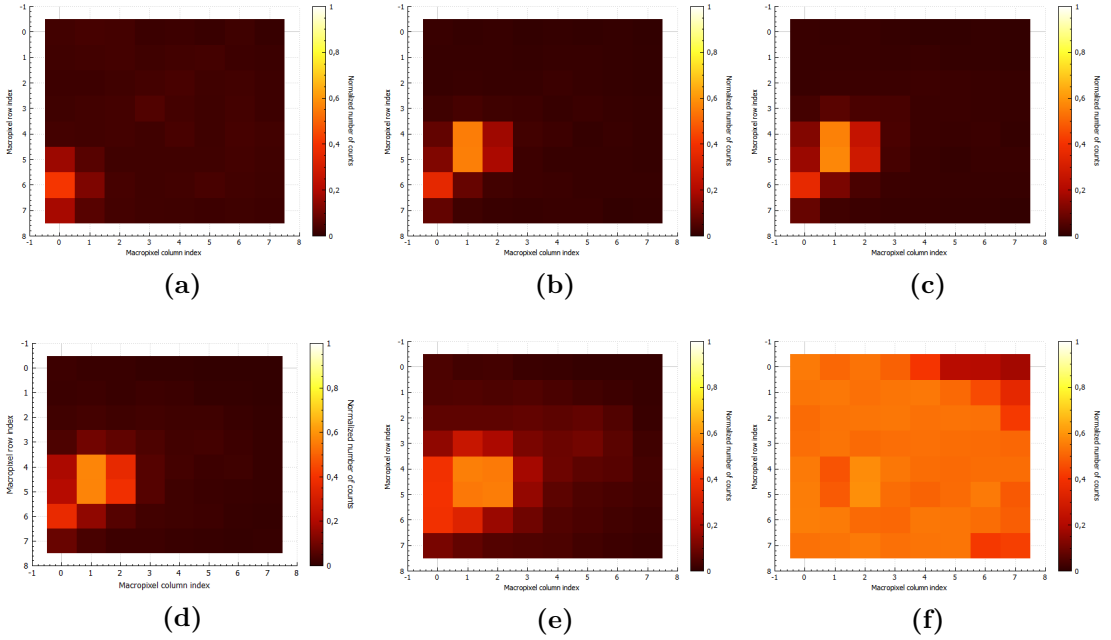


Figure 4.12: These maps are taken at V_{ex} of 2.5 V. (a) is taken with no illumination. (b) to (f) are taken with increasing laser intensity. The scale is not linear and the last plot shows a saturation of all the clusters.

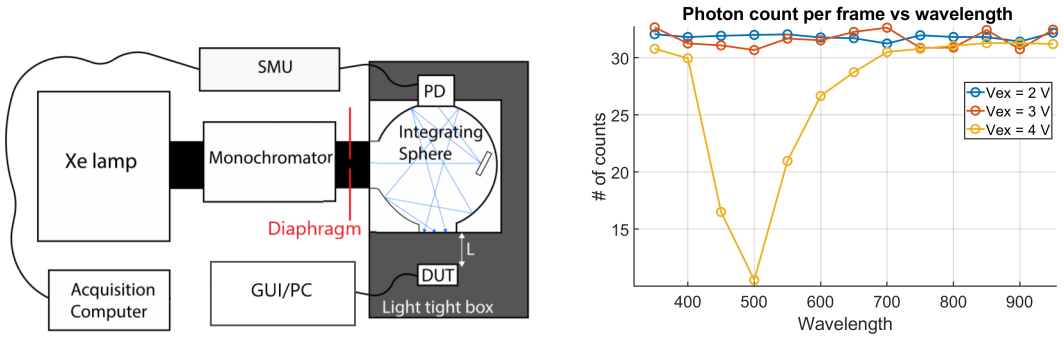
visually the great advantage of decreasing the temperature. Even with a lower excess bias the previous plot had a much higher dark count on every macropixel. The clear difference between the first map, Fig. 4.12a, and the second one, Fig. 4.12b, highlights the spot of the laser. It is located on four cluster of the matrix, close to the noisy ones, even if only two of them look especially active. These four clusters are almost not-triggered in absence of illumination (first plot), so we can be positive enough that their activity is given by the laser spot. Increasing the laser intensity we can notice how also the cluster next to the spot start to activate. This is in accordance with the expected results because the focused laser spot is anyways a Gaussian beam. The light intensity profile then follows the same distribution, meaning that not only the cluster in the centre of the spot will be activated. The effect becomes more evident increasing the laser intensity because also the Gaussian tails of the distribution will become able to trigger more easily the detectors. For example in the plot 4.12e the activation of nearby macropixels is evident, even if the distribution has not a perfect profile.
The last plot, Fig. 4.12f shows a situation where this effect is brought to its limits. In this case the almost the whole matrix is triggered, as it almost reaches saturation. Nonetheless, the photon counts are still decreasing towards the right edges of the

chip, suggesting the Gaussian profile of the intensity. It is important to notice the difference in this last plot concerning the four initial cluster. They are still distinct from the others, but in this case, unlike in the other plot, the two on the right are lighter than the ones on the left. This is understandable reminding that the test of the whole chip prevents the use of the double sampling scheme. Therefore, these results are not a surprise, after the previous analysis: evidently the high number of counts registered in those two cluster, caused the counters to saturate and overflow, restarting from zero. Unfortunately, we are not able to correct this error in this case, without the help of the double sampling readout. Still, this looks as a reasonable explanation of the effect.

4.2.3 Preliminary PDE measurements

Together with DCR, PDE is one of the main figure of merit to characterize a light detector. It is defined as the probability of a photon hitting the device's surface to generate a pulse, i.e. to be detected [9]. It also represents the main parameter to characterize SPADs' sensitivity and it strongly depends on the wavelength and the excess bias voltage. In general, CMOS SPADs have a sharp PDP peak arounf 400-500 nm, due to the shallow junction just below the surface [21].
To perform this test a standard setup is employed. It is formed by a Xe lamp whose light is focused on a monochromator, i.e. a diffraction grating to disperse the different wavelengths. The grating is placed on a motorized stand that can be controlled by the computer to select the desired narrow band of wavelengths, from 340 nm to 960 nm. This range is ideal to assess the device sensitivity peaked in the visible range. An integrating sphere acts as a diffuser to shine the beam on the whole chip. A calibrated photodiode is used to detect the number of arriving photons, before collecting them with the device under test (Fig. 4.13a).
Before considering the exact PDE of the device it is a good idea to simply verify the photon counts dependency on the wavelength. The trend should be more or less the same expected for the SPADs' PDP, and it gives us an idea if the device is working correctly. In addition, these data are then necessary to obtain the PDE values, after the calibration of the photodiode is performed. To get more reliable results, the double sampling scheme is always used in the following, limiting the analysis to a single cluster.
The first run of measurements are carried out relying only on the passive recharge circuit as in previous cases. Given the long dead time, the integration window is further increased to 10 µs. Different excess bias voltages are also used and only few wavelengths are selected. The resulting plot is shown in Fig. 4.13b

The plot results are very different from what we expected. For lower voltages, the counts remain constant, whereas for the highest one the trend is obviously



(a) **(b)**

Figure 4.13: Schematic representation of the PDP setup (a), adapted from [6]. Cluster counts vs wavelengths with 10 µs integration window (b). The values are averaged on 2047 frames

inverted. Where the peak of sensitivity is expected, the counts reach the lowest value. On the other hand, besides being upside-down, the yellow curve has a reasonable trend. This can help us to understand the source of the error.
First of all, the fact that the two first curves are practically constant suggest that the counters are saturating. As was previously discussed in Sec.4.2.1, this effect is indeed observed when the counts reach the values around 30. The saturation is caused by the combination of the high number of photons and the very long integration window. With illumination we can expect that the number of counts is generally larger. In addition, despite the long dead time, the integration interval is large enough to register many events and to cause saturation. The reason because it does not happen for the higher excess voltage, can also be explained. We must consider that, increasing the bias voltage, the SPADs are more active so the situation is even worse: the number of triggering event gets still higher. If the discharge of the SPAD is not fast enough, it happens that, after an output pulse, the next event triggers an avalanche before the voltage at that node is lowered. This means that an additional pulse is generated before the previous one finished. The result is that the voltage level at the output node remains high. As consequence the following inverter does not trigger because the threshold voltage of the component is not reached (the voltage remains almost constant). Therefore, the edge-triggered counter is not enabled, although an event occurred. Then, the resulting count value will be much lower than expected. This explains why the yellow curve has lower values in correspondence of the sensitivity peak and because this effect is observed only for the highest excess bias. It is also true that, if that is the problem, the trend is promising even if upside-down and we can easily correct it.
The straightforward solution is to reduce the pulse width, acting on the discharge

of the SPAD. The best way is to re-enable the active recharge circuit. To avoid the risk of oscillation in the circuit, a small modification is done to the firmware, to force a control bit always to one. The same reasoning is valid in this case: to avoid the saturation of the counters, the integration window must be greatly reduced. The chosen value is 100 ns.

Before proceeding with a new measurement, it is necessary to adjust the light power arriving on the detector. This can be done acting on a diaphragm, placed before the integrating sphere: the width of the slid can be modified to reduce the incoming light on the detector. The aim in this case is to generate a clear difference in the photon count, when the light is around 500 nm and the other wavelengths. If the light intensity is too high, the counters saturate again for every wavelength, so no trend could be appreciated. Some measurements are performed on the whole matrix with this purpose. A quick check on the cluster maps help us to visually notice the difference and set the proper slid aperture. At the same time a still naiver analysis can be done, controlling the drained current. Also here, the values in case of 500 nm light and other wavelengths must be significantly different, without reaching the maximum limit or saturation. The same tests must be performed with several excess bias voltages.
Once the proper configuration is set, the real test can start. The plot in Fig. 4.14 displays the results.

All the curves in the plot follow the expected trend, with a clear peak between 450 nm-500 nm. This confirms that the previous analysis and the the following precautions were correct. Also the dependency on the excess bias voltage is verified, apart in the region of the peak where all the curves overlap a bit. A significant ratio of three-four is present between the values on the peak and the lowest counts. The two smaller peaks after 800 nm are the only unusual features. The fact that they are exactly repeated in every curve at increasing voltages, suggest that this is not a readout issue but it is rather caused by the setup. A proof of this hypothesis is found looking at the irradiance distribution of the emitting lamp in Fig. 4.15 [76]. The curve to consider for our lamp is the one indicated as 6258. Several peaks between 800 nm and 1,000 nm are clear and they are the source of the corresponding larger number of counts in the previous plot. This also explains the necessity of using the calibrated photodiode for the real PDE evaluation. These lower peaks are an artifact due to the variable lamp emission and can be eliminated. The photodiode is used to practically normalize the number of events with respect to the actual number of arriving photons. For these wavelengths a larger number of photons are hitting the detector and this will be considered with the photodiode output. This takes into account the lamp's profile so that the PDE measurement will not be affected by it. The result is the elimination of the two smaller peak.

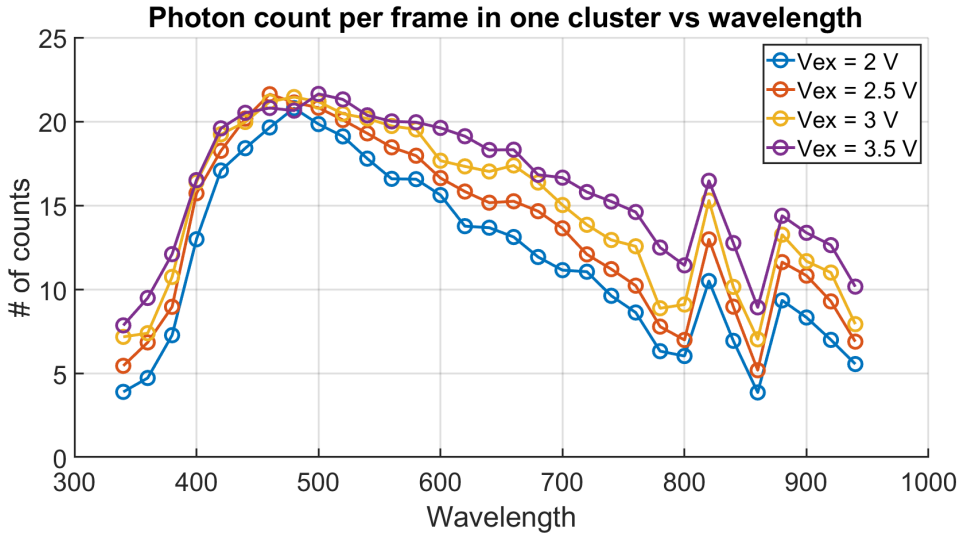


Figure 4.14: Cluster counts vs wavelengths with 100 ns integration window. The values are averaged on 2047 frames. The smaller integration window is justified by the activation of the active recharge circuit that significantly reduces the dead time. The curves follow the expected trend for all the excess bias voltages. The smaller peaks around 800 ns are caused by the light source irradiance distribution (see Fig. 4.15).

The calibration of the photodiode was not performed, so that more quantitative results for the PDE are not yet available. Nonetheless, given the promising trend in Fig. 4.14, we can expect a positive result from the test, already scheduled for the near future.

Additional verification can be undertaken to confirm the assumptions done so far. First of all, we can check if the choice of the smaller integration window was the correct choice. With a fixed bias the same test are performed, this time varying the integration interval: for larger value we expect to observe saturation. The results are shown in Fig. 4.16a.
Similar effects can be observed modifying the recharge time of the SPAD. A smaller dead time, directly results in a larger number of events that can be detected. This is achieved modifying some of the voltages controlling the transistor of the recharge circuit. Depending on it, their conduction varies, resulting in a faster or slower recharge. The effect can be appreciated in the plot of Fig. 4.16b, Here only the conduction of one transistor was modified to reduce the fall time of the output pulse.

These plots verify all the previous assumptions and are coherent with the expected results. Plot 4.16a confirms that the choice of the integration window of

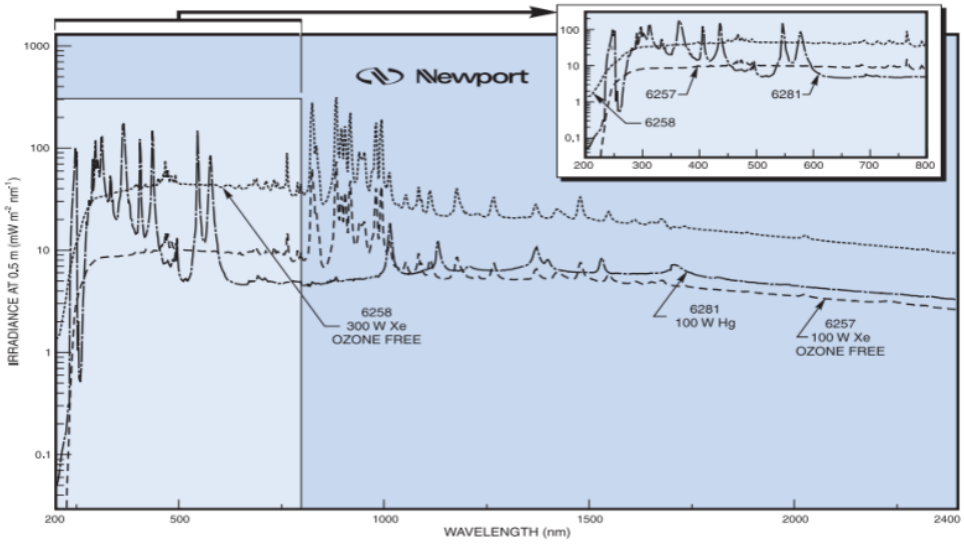


Figure 4.15: Spectral irradiance of various arc lamps from Oriel/Newport [76]. The one employed in our setup is the 6258. Some peaks are evident from 800 nm to 1,000 nm.

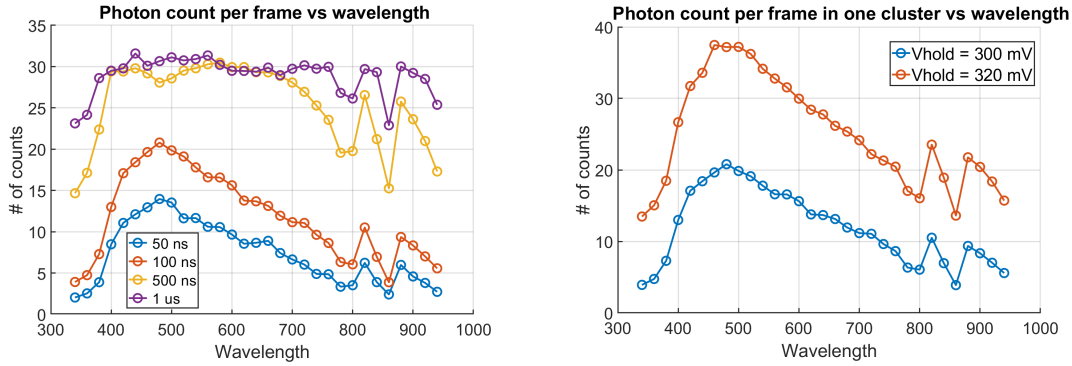


Figure 4.16: Average cluster counts vs lambda with $V_{ex} = 2\,V$, varying: the integration window (a); the voltage of one of the transistor in the active recharge circuit, affecting the dead time of the SPADs (b). In this second case the integration is set again to 100 ns. All the counts in both plots are averaged on 2047 acquired fraes.

100 ns was correct. For lower values, the interval would not be enough to collect a considerable number of events. This is true especially for the peak, that results lower and flatter. If the integration window increases more, the counts, in turn,

grow coherently, but the problem becomes the saturation, as expected. It is especially evident in the last, violet curve, showing almost a constant trend over all the wavelengths. Anyway, also for lower values, e.g. 500 ns, the saturation masks completely the peak of sensitivity of the detector.
The plot on the right, in Fig. 4.16b, shows a similar effect caused by a variation in the voltage V_{hold}. A higher value corresponds to a shorter output pulse and a lower dead time. This allows the detection of more events, under the same conditions. In the plot this is verified by the significantly higher number of counts registered in the second, red curve. A voltage variation of 200 mV is sufficient to almost double the counts. The increased number of events also corresponds, macroscopically, to an increased drained current. At this point, setting this voltage at higher values, seems the right choice, because the peak results sharper. On the other hand, we must consider that these data were taken with an excess voltage of 2 V. If we want to increase it further, it is very likely that the larger number of counts might cause again saturation. As consequence, the best choice in this case, is to keep this value around 300 mV to have a significant sensitivity and a good range of operation, as in Fig. 4.14.

Chapter 5

Conclusion and future work

A complete, stable and user-friendly testing system was presented in this work. It allows direct communication with the analysed FSI 3D-stacked MD-SiPM and helps to perform the necessary tests in a semi-automatic way. Thanks to it, a preliminary characterisation of the innovative photodector was possible. This, in turn, verified the correct functionality of the implemented testing system.
The resulting measurements are presented in Chapter 4, starting from the TDC characterisation. The device is proved to be active and working, even if some issues are present and will need further testing to be resolved. Some photon counting measurements are also reported, to characterise both the DCR of the chip as well as its activity under illumination. The noise level results particularly high compared to usual values [9, 10, 11, 12, 14, 17, 19, 21, 22, 28, 29, 30] for isolated pixels and 3D-stacked structure. On the other hand, this test helped to understand that some artifacts might be present due to some problems in the counting system. However, an additional test showed that the dark count reaches acceptable values when the detector is cooled, verifying the correct trend in temperature and opening a new reliable possibility for future tests. The chip seems to respond correctly under laser illumination too. A preliminary PDE measurement is presented in Sec. 4.2.3, with promising results (as allowed by the aforementioned issues), showing the expected trend.
In general, from these preliminary results, the chip seems to work correctly apart from some minor issues that was possible to identify thanks to the developed testing system. The 3D integration process briefly presented in Sec. 2.3 looks successful and can be exploited for future designs. Nevertheless, a more detailed characterisation of the system is still ongoing and more accurate results will be available in the future. For what concerns the TDC, the best strategy will be to sample the data directly from them, bypassing the output processing unit. This technique was already employed on an independent TDC test structure and demonstrated its validity. Additional measurements will also be performed to obtain a complete PDE

characterisation. A calibration with the photodiode is necessary for this purpose and will surely be performed on the already available setup. A new batch of chips with some minor changes, especially in the masking circuit, will be tested, hopefully leading to a lower, acceptable DCR. These and future tests will be performed with the help of the testing system described in this thesis.
In this context, it is complex to give a comparison of the performances with respect to other developed solutions. First of all, we still have to precisely assess the capability of the detector, as explained above, and only future tests will provide us with quantitative results. It is also true that not many comparable, complete and already published systems are present in the literature. This is, to our knowledge, the first FSI 3D MD-SiPM of these dimensions, successfully integrated and working. A recent review on other 3D-stacked, or similar, solutions can be found in [30]. Still, the comparison of these emerging technologies remains problematic, given the lack of data due to ongoing development.
The development of the testing system, as first goal of the thesis, was completely successful, leading to a comprehensive and user-friendly environment that will be used also in the future. Its contribution to a fast and semi-automatic test of the chip allowed us to verify the detector functionalities and the reliability of the exploited 3D integration process. The following preliminary characterisation represents the second main contribution of this work. Besides testing the different systems on the chip, it also led to discovery of some main issues of the design. This will be further investigated and will undoubtedly form a solid basis for forthcoming evolution of the detector and designs. Even if not complete yet, also the analysis and the performed tests represent a valid starting point for more advanced and quantitative tests.

Bibliography

[1] Peter Seitz and Albert JP Theuwissen. *Single-photon imaging*. Vol. 160. Springer Science & Business Media, 2011 (cit. on p. 1).

[2] G S Buller and R J Collins. «Single-photon generation and detection». In: *Measurement Science and Technology* 21.1 (Nov. 2009), p. 012002. DOI: 10.1088/0957-0233/21/1/012002. URL: https://doi.org/10.1088/0957-0233/21/1/012002 (cit. on pp. 1, 10).

[3] Silvano Donati and Tiziana Tambosso. «Single-Photon Detectors: From Traditional PMT to Solid-State SPAD-Based Technology». In: *IEEE Journal of Selected Topics in Quantum Electronics* 20.6 (2014), pp. 204–211. DOI: 10.1109/JSTQE.2014.2350836 (cit. on pp. 1, 7, 9, 11).

[4] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa. «Avalanche photodiodes and quenching circuits for single-photon detection». In: *Appl. Opt.* 35.12 (Apr. 1996), pp. 1956–1976. DOI: 10.1364/AO.35.001956. URL: http://ao.osa.org/abstract.cfm?URI=ao-35-12-1956 (cit. on p. 1).

[5] Francesco Ceccarelli, Giulia Acconcia, Angelo Gulinatti, Massimo Ghioni, Ivan Rech, and Roberto Osellame. «Recent Advances and Future Perspectives of Single-Photon Avalanche Diodes for Quantum Photonics Applications». In: *Advanced Quantum Technologies* 4 (Dec. 2020), p. 2000102. DOI: 10.1002/qute.202000102 (cit. on pp. 1, 12).

[6] Francesco Gramuglia, Ming-Lo Wu, Claudio Bruschini, Myung-Jae Lee, and Edoardo Charbon. «A Low-noise CMOS SPAD Pixel with 12.1 ps SPTR and 3 ns Dead Time». In: *IEEE Journal of Selected Topics in Quantum Electronics* (2021), pp. 1–1. DOI: 10.1109/JSTQE.2021.3088216 (cit. on pp. 1, 24–28, 81).

[7] S. Cova, A. Lacaita, M. Ghioni, G. Ripamonti, and T A Louis. «20-ps timing resolution with single-photon avalanche diodes». English. In: *Review of Scientific Instruments* 60.6 (1989), pp. 1104–1110. ISSN: 0034-6748. DOI: 10.1063/1.1140324 (cit. on p. 1).

[8] Frédéric Nolet, Samuel Parent, Nicolas Roy, Marc-Olivier Mercier, Serge A. Charlebois, Réjean Fontaine, and Jean-Francois Pratte. «Quenching Circuit and SPAD Integrated in CMOS 65 nm with 7.8 ps FWHM Single Photon Timing Resolution». In: *Instruments* 2.4 (2018). ISSN: 2410-390X. DOI: 10.3390/instruments2040019. URL: https://www.mdpi.com/2410-390X/2/4/19 (cit. on pp. 1, 19).

[9] Cristiano Niclass, Marek Gersbach, Robert Henderson, Lindsay Grant, and E. Charbon. «A 130-nm CMOS single-photon avalanche diode». In: *Proceedings of SPIE - The International Society for Optical Engineering* 6766 (Oct. 2007). DOI: 10.1117/12.728878 (cit. on pp. 1, 12, 13, 80, 86).

[10] E. Charbon, Hyung-June Yoon, and Y. Maruyama. «A Geiger mode APD fabricated in standard 65nm CMOS technology». In: *2013 IEEE International Electron Devices Meeting* (2013), pp. 27.5.1–27.5.4 (cit. on pp. 1, 86).

[11] Chockalingam Veerappan and Edoardo Charbon. «A Low Dark Count p-i-n Diode Based SPAD in CMOS Technology». In: *IEEE Transactions on Electron Devices* 63.1 (2016), pp. 65–71. DOI: 10.1109/TED.2015.2475355 (cit. on pp. 1, 25, 86).

[12] Myung-Jae Lee, Pengfei Sun, and Edoardo Charbon. «A first single-photon avalanche diode fabricated in standard SOI CMOS technology with a full characterization of the device». In: *Opt. Express* 23.10 (May 2015), pp. 13200–13209. DOI: 10.1364/OE.23.013200. URL: http://www.opticsexpress.org/abstract.cfm?URI=oe-23-10-13200 (cit. on pp. 1, 86).

[13] Chockalingam Veerappan and Edoardo Charbon. «A Substrate Isolated CMOS SPAD Enabling Wide Spectral Response and Low Electrical Crosstalk». In: *IEEE Journal of Selected Topics in Quantum Electronics* 20.6 (2014), pp. 299–305. DOI: 10.1109/JSTQE.2014.2318436 (cit. on p. 1).

[14] Francesco Gramuglia, Andrada Muntean, Esteban Venialgo, Myung-Jae Lee, Scott Lindner, Makoto Motoyoshi, Andrei Ardelean, Claudio Bruschini, and Edoardo Charbon. «CMOS 3D-Stacked FSI Multi-Channel Digital SiPM for Time-of-Flight PET Applications». In: *2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. 2020, pp. 1–3. DOI: 10.1109/NSS/MIC42677.2020.9507833 (cit. on pp. 1, 2, 19, 24, 27, 28, 30, 86).

[15] Stefan Gundacker and Arjan Heering. «The silicon photomultiplier: fundamentals and applications of a modern solid-state photon detector». In: *Phys. Med. Biol.* 65.17TR01 (2020) (cit. on pp. 1, 15–18, 20).

[16] S. Gundacker, E. Auffray, P. Jarron, T. Meyer, and P. Lecoq. «On the comparison of analog and digital SiPM readout in terms of expected timing performance». In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 787 (2015). New Developments in Photodetection NDIP14, pp. 6–11. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2014.10.020. URL: https://www.sciencedirect.com/science/article/pii/S016890021401167X (cit. on p. 1).

[17] Shingo Mandai and Edoardo Charbon. «Multi-channel digital SiPMs: Concept, analysis and implementation». In: *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*. 2012, pp. 1840–1844. DOI: 10.1109/NSSMIC.2012.6551429 (cit. on pp. 1, 24, 86).

[18] Francesco Gramuglia, Andrada Muntean, Carlo Alberto Fenoglio, Myung-Jae Lee, Scott Lindner, Makoto Motoyoshi, Andrei Ardelean, Claudio Bruschini, and Charbon Edoardo. «CMOS 3D-Stacked FSI Multi-Channel Digital SiPM for Time-of-Flight Vision Applications (submitted for publication)». In: Aug. 2021 (cit. on pp. 2, 24, 25, 27–31).

[19] Edoardo Charbon, Claudio Bruschini, and Myung-Jae Lee. «3D-Stacked CMOS SPAD Image Sensors: Technology and Applications». In: *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. 2018, pp. 1–4. DOI: 10.1109/ICECS.2018.8617983 (cit. on pp. 2, 19, 20, 86).

[20] F. Nolet et al. «A 2D Proof of Principle Towards a 3D Digital SiPM in HV CMOS With Low Output Capacitance». In: *IEEE TRANSACTIONS ON NUCLEAR SCIENCE* 63.4 (Aug. 2016) (cit. on pp. 2, 19, 24).

[21] Myung-Jae Lee and Edoardo Charbon. «Progress in single-photon avalanche diode image sensors in standard CMOS: From two-dimensional monolithic to three-dimensional-stacked technology». In: *Jpn. J. Appl. Phys.* 57.10 (Sept. 2018) (cit. on pp. 2, 14, 18, 20, 80, 86).

[22] E. Charbon, M. Scandini, J. Mata Pavia, and M. Wolf. «A dual backside-illuminated 800-cell multi-channel digital SiPM with 100 TDCs in 130nm 3D IC technology». In: *2014 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. 2014, pp. 1–4. DOI: 10.1109/NSSMIC.2014.7431246 (cit. on pp. 2, 19, 86).

[23] Frédéric Nolet, William Lemaire, Frédérik Dubois, Nicolas Roy, Simon Carrier, Arnaud Samson, Serge A. Charlebois, Réjean Fontaine, and Jean-Francois Pratte. «A 256 Pixelated SPAD readout ASIC with in-Pixel TDC and embedded digital signal processing for uniformity and skew correction». In: *Nuclear Instruments and Methods in Physics Research A* 949, 162891 (Jan. 2020), p. 162891. DOI: 10.1016/j.nima.2019.162891 (cit. on pp. 2, 19).

[24] Samuel Parent, Maxime Côté, Frédéric Vachon, Robert Groulx, Stéphane Martel, Henri Dautet, Serge A. Charlebois, and Jean-François Pratte. «Single Photon Avalanche Diodes and Vertical Integration Process for a 3D Digital SiPM using Industrial Semiconductor Technologies». In: *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*. 2018, pp. 1–4. DOI: 10.1109/NSSMIC.2018.8824571 (cit. on pp. 2, 19).

[25] Benoit-Louis Bérubé, Vincent-Philippe Rhéaume, Samuel Parent, Luc Maurais, Audrey Corbeil Therrien, Paul G. Charette, Serge A. Charlebois, Réjean Fontaine, and Jean-François Pratte. «Implementation Study of Single Photon Avalanche Diodes (SPAD) in 0.8 μm</tex> </formula> HV CMOS Technology». In: *IEEE Transactions on Nuclear Science* 62.3 (2015), pp. 710–718. DOI: 10.1109/TNS.2015.2424852 (cit. on pp. 2, 19).

[26] Augusto Ronchini Ximenes, Preethi Padmanabhan, Myung-Jae Lee, Yuichiro Yamashita, D. N. Yaung, and Edoardo Charbon. «A 256×256 45/65nm 3D-stacked SPAD-based direct TOF image sensor for LiDAR applications with optical polar modulation for up to 18.6dB interference suppression». In: *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*. 2018, pp. 96–98. DOI: 10.1109/ISSCC.2018.8310201 (cit. on pp. 2, 19, 20).

[27] Nicolas Roy, Frédéric Nolet, Frédérik Dubois, Marc-Olivier Mercier, Réjean Fontaine, and Jean-François Pratte. «Low Power and Small Area, 6.9 ps RMS Time-to-Digital Converter for 3-D Digital SiPM». In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 1.6 (2017), pp. 486–494. DOI: 10.1109/TRPMS.2017.2757444 (cit. on pp. 2, 19).

[28] J Pavia. «M., Scandini, M., Lindner, & S., et al."A 1× 400 backside-illuminated SPAD sensor with 49.7 ps resolution, 30 pJ/sample TDCs fabricated in 3D CMOS technology for near-infrared optical tomography,"» in: *Solid-State Circuits, IEEE Journal of solid-state circuits* 50.10 (2015), pp. 2406–2418 (cit. on pp. 2, 86).

[29] Scott Lindner, Sara Pellegrini, Yann Henrion, Bruce Rae, Martin Wolf, and Edoardo Charbon. «A High-PDE, Backside-Illuminated SPAD in 65/40-nm 3D IC CMOS Pixel With Cascoded Passive Quenching and Active Recharge». In: *IEEE Electron Device Letters* 38.11 (2017), pp. 1547–1550. DOI: 10.1109/LED.2017.2755989 (cit. on pp. 2, 25, 86).

[30] Jean-François Pratte et al. «3D Photon-To-Digital Converter for Radiation Instrumentation: Motivation and Future Works». In: *Sensors* 21.2 (2021). ISSN: 1424-8220. DOI: 10.3390/s21020598. URL: https://www.mdpi.com/1424-8220/21/2/598 (cit. on pp. 2, 18, 20, 86, 87).

[31] Albert Einstein. «Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]». In: *Annalen der Physik* 322.10 (1905), pp. 891–921. DOI: http://dx.doi.org/10.1002/andp.19053221004 (cit. on p. 3).

[32] Sergio Cova. *Sensori, segnali e rumore - Signal recovery*. 2016. URL: https://cova.faculty.polimi.it/elet/index.html (visited on 08/22/2021) (cit. on pp. 4, 8, 12, 13).

[33] G. Brida, S. Castelletto, C. Novero, and M. L. Rastello. «Quantum-efficiency measurement of photodetectors by means of correlated photons». In: *J. Opt. Soc. Am. B* 16.10 (Oct. 1999), pp. 1623–1627. DOI: 10.1364/JOSAB.16.001623. URL: http://josab.osa.org/abstract.cfm?URI=josab-16-10-1623 (cit. on p. 5).

[34] Simon M. Sze. *Dispositivi a semiconduttore*. Milano (IT): Ulrico Hoepli Editore S.p.A, 2017 (cit. on pp. 5, 11).

[35] Rongqing Hui and Maurice O'Sullivan. «Chapter 3 - Characterization of Optical Devices». In: *Fiber Optic Measurement Techniques*. Ed. by Rongqing Hui and Maurice O'Sullivan. Boston: Academic Press, 2009, pp. 259–363. ISBN: 978-0-12-373865-3. DOI: https://doi.org/10.1016/B978-0-12-373865-3.00003-3. URL: https://www.sciencedirect.com/science/article/pii/B9780123738653000033 (cit. on p. 5).

[36] V. Mackowiak and J. Peupelmann. «NEP – Noise Equivalent Power». In: 2015 (cit. on p. 5).

[37] CWJ Beenakker and Christian Schönenberger. «Quantum shot noise». In: (2003) (cit. on p. 6).

[38] Ya.M. Blanter and M. Büttiker. «Shot noise in mesoscopic conductors». In: *Physics Reports* 336.1 (2000), pp. 1–166. ISSN: 0370-1573. DOI: https://doi.org/10.1016/S0370-1573(99)00123-4. URL: https://www.sciencedirect.com/science/article/pii/S0370157399001234 (cit. on p. 6).

[39] B.K. Lubsandorzhiev. «On the history of photomultiplier tube invention». In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 567.1 (2006). Proceedings of the 4th International Conference on New Developments in Photodetection, pp. 236–238. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2006.05.221. URL: https://www.sciencedirect.com/science/article/pii/S0168900206009260 (cit. on p. 7).

[40] Marvin Chodorow et al. «16 - Electron Tubes». In: *Reference Data for Engineers (Ninth Edition)*. Ed. by Wendy M. Middleton and Mac E. Van Valkenburg. Ninth Edition. Woburn: Newnes, 2002, pp. 16-1-16–59. ISBN: 978-0-7506-7291-7. DOI: https://doi.org/10.1016/B978-075067291-7/50018-2. URL: https://www.sciencedirect.com/science/article/pii/B9780750672917500182 (cit. on p. 8).

[41] Robert H. Hadfield. «Single-photon detectors for optical quantum information applications». In: *Nature Photonics* 3 (Dec. 2009). DOI: 10.1038/nphoton.2009.230. URL: https://doi.org/10.1038/nphoton.2009.230 (cit. on p. 9).

[42] Michael F. L'Annuziata. «11 - SOLID SCINTILLATION ANALYSIS». In: *Handbook of Radioactivity Analysis (Second Edition)*. Ed. by Michael F. L'Annunziata. Second Edition. San Diego: Academic Press, 2003, pp. 845–987. ISBN: 978-0-12-436603-9. DOI: https://doi.org/10.1016/B978-012436603-9/50016-8. URL: https://www.sciencedirect.com/science/article/pii/B9780124366039500168 (cit. on p. 9).

[43] C. Bruschini, H. Homulle, I.M. Antolovic, et al. «Single-photon avalanche diode imagers in biophotonics: review and outlook». In: *Light Sci Appl* 8.87 (2018). DOI: https://doi.org/10.1038/s41377-019-0191-5 (cit. on pp. 10, 12–15).

[44] E Charbon. «Single-photon imaging in complementary metal oxide semiconductor processes». In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 372.2012 (2014), p. 20130100 (cit. on pp. 12, 13).

[45] Massimo Ghioni, Angelo Gulinatti, Ivan Rech, Piera Maccagnani, and Sergio Cova. «Large-area low-jitter silicon single photon avalanche diodes». In: *Proc SPIE* (Feb. 2008). DOI: 10.1117/12.761578 (cit. on p. 13).

[46] I. M. Antolovic. «SPAD imagers for super resolution microscopy». PhD thesis. Delft: TU Delft, Jan. 2018. URL: https://doi.org/10.4233/uuid:cc76e95c-b82e-4555-9110-348ad9989705 (cit. on p. 14).

[47] Federica Villa, Danilo Bronzi, Michele Vergani, Yu Zou, Alessandro Ruggeri, Franco Zappa, and Alberto Dalla Mora. «Analog SiPM in planar CMOS technology». In: *2014 44th European Solid State Device Research Conference (ESSDERC)*. 2014, pp. 294–297. DOI: 10.1109/ESSDERC.2014.6948818 (cit. on p. 15).

[48] Andrada Muntean, Ashish Sachdeva, Esteban Venialgo, Salvatore Gnecchi, Darek Palubiak, Carl Jackson, and Edoardo Charbon. «A Fully Integrated State-of-the-Art Analog SiPM with on-chip Time Conversion». In: *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*. 2018, pp. 1–3. DOI: 10.1109/NSSMIC.2018.8824662 (cit. on pp. 15, 28).

[49] *Introduction to the Silicon Photomultiplier (SiPM),(AND9770/D)*. ON Semiconductor (cit. on pp. 15, 16).

[50] Q. Xie N. D'Ascenzo V. Saveliev and L. Wang. «The Digital Silicon Photomultiplier». In: *Optoelectronics: Materials and Devices*. Ed. by S. L. Pyshkin and J. Ballato. 2015. Chap. 18. DOI: 10.5772/59334 (cit. on pp. 17, 18).

[51] Neale A. W. Dutton, Salvatore Gnecchi, Luca Parmesan, Andrew J. Holmes, Bruce Rae, Lindsay A. Grant, and Robert K. Henderson. «11.5 A time-correlated single-photon-counting sensor with 14GS/S histogramming time-to-digital converter». In: *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*. 2015, pp. 1–3. DOI: 10.1109/ISSCC.2015.7062997 (cit. on p. 19).

[52] Niccolò Calandri, Mirko Sanzaro, Lorenzo Motta, Claudio Savoia, and Alberto Tosi. «Optical Crosstalk in InGaAs/InP SPAD Array: Analysis and Reduction With FIB-Etched Trenches». In: *IEEE Photonics Technology Letters* 28.16 (2016), pp. 1767–1770. DOI: 10.1109/LPT.2016.2570278 (cit. on p. 20).

[53] Kateryna Kuzmenko, Peter Vines, Zoe Greener, Jarosław Kirdoda, Derek Dumas, Muhammad M. A. Mirza, Ross Millar, Douglas J. Paul, and Gerald S. Buller. «Planar geometry Ge-on-Si SPAD detectors for the short-wave infrared (Conference Presentation)». In: *Advanced Photon Counting Techniques XIII*. Ed. by Mark A. Itzler, Joshua C. Bienfang, and K. Alex McIntosh. Vol. 10978. International Society for Optics and Photonics. SPIE, 2019. URL: https://doi.org/10.1117/12.2518858 (cit. on p. 20).

[54] M.-A. Tétrault, É. Desaulniers Lamy, A. Boisvert, J.-F. Pratte, and R. Fontaine. «Real-time discreet SPAD array readout architecture for time of flight PET». In: *2014 19th IEEE-NPSS Real Time Conference*. 2014, pp. 1–3. DOI: 10.1109/RTC.2014.7097479 (cit. on p. 20).

[55] Zickus V., Wu ML., Morimoto K., et al. «Fluorescence lifetime imaging with a megapixel SPAD camera and neural network lifetime estimation». In: *Sci. Rep.* 10 (Dec. 2020). DOI: 10.1038/s41598-020-77737-0 (cit. on p. 20).

[56] Yuki Maruyama, Jordana Blacksberg, and Edoardo Charbon. «A 1024×8, 700-ps Time-Gated SPAD Line Sensor for Planetary Surface Exploration With Laser Raman Spectroscopy and LIBS». In: *IEEE Journal of Solid-State Circuits* 49.1 (2014), pp. 179–189. DOI: 10.1109/JSSC.2013.2282091 (cit. on p. 20).

[57] Scott Lindner et al. «A Novel 32 × 32, 224 Mevents/s Time Resolved SPAD Image Sensor for Near-Infrared Optical Tomography». In: *Biophotonics Congress: Biomedical Optics Congress 2018 (Microscopy/Translational/Brain/OTS)*. Optical Society of America, 2018, JTh5A.6. URL: http://www.osapublishing.org/abstract.cfm?URI=BRAIN-2018-JTh5A.6 (cit. on p. 20).

[58] Terry Jones and David W. Townsend. «History and future technical innovation in positron emission tomography». In: *Journal of Medical Imaging* 4.1 (2017), pp. 1–17. DOI: 10.1117/1.JMI.4.1.011013. URL: https://doi.org/10.1117/1.JMI.4.1.011013 (cit. on pp. 21, 23).

[59] R. Boellaard, R. Delgado-Bolton, W.J.G. Oyen, et al. «FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0». In: *Eur J Nucl Med Mol Imaging* 42 (2015), pp. 328–354. DOI: 10.1007/s00259-014-2961-x (cit. on p. 21).

[60] Joseph Lau, Etienne Rousseau, Daniel Kwon, Kuo-Shyan Lin, François Bénard, and Xiaoyuan Chen. «Insight into the Development of PET Radiopharmaceuticals for Oncology». In: *Cancers* 12.5 (2020). ISSN: 2072-6694. DOI: 10.3390/cancers12051312. URL: https://www.mdpi.com/2072-6694/12/5/1312 (cit. on p. 21).

[61] Rajesh Ganai, Shaifali Mehta, Mehul Shiroya, Mitali Mondal, Zubayer Ahammed, and Subhasis Chattopadhyay. «A Proof- of-Principle for Time of Flight-Positron Emission Tomography Imaging». In: June 2018, pp. 125–128. ISBN: 978-3-319-73170-4. DOI: 10.1007/978-3-319-73171-1_27 (cit. on pp. 21, 22).

[62] W.W. Moses. «Time of flight in PET revisited». In: *IEEE Transactions on Nuclear Science* 50.5 (2003), pp. 1325–1330. DOI: 10.1109/TNS.2003.817319 (cit. on pp. 22, 23).

[63] Victor Westerwoudt, Maurizio Conti, and Lars Eriksson. «Advantages of Improved Time Resolution for TOF PET at Very Low Statistics». In: *IEEE Transactions on Nuclear Science* 61.1 (2014), pp. 126–133. DOI: 10.1109/TNS.2013.2287175 (cit. on p. 22).

[64] C.M. Pepin, P. Berard, A.-L. Perrot, C. Pepin, D. Houde, R. Lecomte, C.L. Melcher, and H. Dautet. «Properties of LYSO and recent LSO scintillators for phoswich PET detectors». In: *IEEE Transactions on Nuclear Science* 51.3 (2004), pp. 789–795. DOI: 10.1109/TNS.2004.829781 (cit. on p. 23).

[65] Francesco Gramuglia et al. «Light Extraction Enhancement Techniques for Inorganic Scintillators». In: *Crystals* 11.4 (2021). ISSN: 2073-4352. DOI: 10.3390/cryst11040362. URL: https://www.mdpi.com/2073-4352/11/4/362 (cit. on p. 23).

[66] S. E. Brunner and D. R. Schaart. «BGO as a hybrid scintillator / Cherenkov radiator for cost-effective time-of-flight PET». In: *Phys. Med. Biol.* 62.11 (2017). DOI: 10.1007/s00259-014-2961-x (cit. on p. 23).

[67] Massimo Ghioni, Angelo Gulinatti, Ivan Rech, Franco Zappa, and Sergio Cova. «Progress in Silicon Single-Photon Avalanche Diodes». In: *Selected Topics in Quantum Electronics, IEEE Journal of* 13 (Aug. 2007), pp. 852–862. DOI: 10.1109/JSTQE.2007.902088 (cit. on p. 24).

[68] Ivan Michel Antolovic, Claudio Bruschini, and Edoardo Charbon. «Dynamic range extension for photon counting arrays». In: *Opt. Express* 26.17 (Aug. 2018), pp. 22234–22248. DOI: 10.1364/OE.26.022234. URL: http://www.opticsexpress.org/abstract.cfm?URI=oe-26-17-22234 (cit. on p. 25).

[69] Przemyslaw Mroszczyk and Piotr Dudek. «Tunable CMOS Delay Gate With Improved Matching Properties». In: *Circuits and Systems I: Regular Papers, IEEE Transactions on* 61 (Sept. 2014), pp. 2586–2595. DOI: 10.1109/TCSI.2014.2312491 (cit. on p. 26).

[70] Andrada Muntean, Esteban Venialgo, Salvatore Gnecchi, Carl Jackson, and Edoardo Charbon. «Towards a fully digital state-of-the-art analog SiPM». In: Oct. 2017, pp. 1–4. DOI: 10.1109/NSSMIC.2017.8533036 (cit. on p. 28).

[71] Zeke van Sachez. *Opal Kelly Documentation Portal*. 2021. URL: https://docs.opalkelly.com/display/HOME/Opal+Kelly+Documentation+Portal (visited on 08/27/2021) (cit. on p. 33).

[72] *FrontPanel SDK*. Opal Kelly Incorporated. 2018 (cit. on pp. 34, 36).

[73] *7 Series FPGAs SelectIO Resources User Guide (UG471)*. Xilinx. May 2018 (cit. on p. 40).

[74] *Kintex-7 FPGAs Data Sheet: DC and AC Switching Characteristics*. Xilinx, Mar. 2021 (cit. on p. 40).

[75] *okCFrontPanel Class Reference*. Opal Kelly Incorporated. 2019. URL: https://library.opalkelly.com/library/FrontPanelAPI/classokCFrontPanel.html (cit. on p. 52).

[76] *Oriel product training - Spectral Irradiance*. Oriel Instruments - Newport. Stratford (CT) (cit. on pp. 82, 84).