EPFL

Thèse n° 9529

Integrated electronics for time-of-flight positron emission tomography photodetectors

Présentée le 12 janvier 2023

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

pour l'obtention du grade de Docteur ès Sciences

par

Andrada Alexandra MUNTEAN

Acceptée sur proposition du jury

Prof. G. De Micheli, président du jury
Prof. E. Charbon, Dr C. Bruschini, directeurs de thèse
Dr D. Schaart, rapporteur
Dr C. Degenhardt, rapporteur
Prof. A. P. Burg, rapporteur

École polytechnique fédérale de Lausanne

2023

It pays to keep an open mind, but not so open your brains fall out.
— Carl Sagan

To Andrei...

Acknowledgements

A PhD is a rollercoaster, apart from all the scientific knowledge you gain over these years, it teaches you a life lesson... This would not have been possible without the help and support of many people and I will start by thanking my advisor, Professor Edoardo Charbon. I started working in Edoardo's lab when I was a master's student in Delft. I moved from Romania to the Netherlands to study microelectronics and Edoardo has continued to give me great opportunities to do it ever since. I want to sincerely thank Edoardo for this, for all the freedom he gave me to implement my ideas and for all the support during difficult periods in my PhD. Following, I want to thank Claudio Bruschini for all the support during these years and to say that I will never forget his attention to detail for every presentation or paper I had. I would also like to thank Brigitte Khan for all the help she gave me since I moved to Switzerland. Besides work, I am very happy that we share a common passion - yoga. We spent some time together talking about this practice and Brigitte gave me valuable insights which will stay with me forever. I would also like to thank Tora Begonia for taking care of all my administrative paperwork during the final months of my PhD and for not letting the croissants tradition die after Brigitte left. I would also like to thank Esteban Venialgo for his constant help and all the interesting scientific conversations we had during my master and PhD. I really appreciate all the personal time he took to discuss work with me. He always played the devil's advocate which made me challenge myself and grow. I would also like to thank Emanuele Ripiccini for all the support, all the knowledge he shared with me and the time spent in the lab. Emanuele joined the team later on, but he was always very helpful. During my first year of PhD I started a joint chip design with Francesco Gramuglia and I would like to thank him for this part of the experience. I have spent a lot of time in AQUA lab and a bunch of students left Delft to come to Neuchâtel. During this period, I spent more time with Preethi Padmanabhan. We started having coffee breaks together, went on hikes and then we discovered we shared a common passion for yoga. I also spent more time with Arin Can Ulku. I will never forget his dedication and integrity and I am happy for all the time spent together. I would like to thank all my other colleagues and I will list them here in no particular order: Kazuhiro Morimoto, Myung-Jae Lee, Ivan Michel Antolovic, Samuel Burri, Scott Lindner, Augusto Xiemenes, Augusto Carimatto, Andrea Ruffino, Bedirhan

Ilik, Feng Liu, Ermanno Bernasconi, Jad Benserhir, Baris Can Efe, Simone Frasca, Utku Karaca, Pouyan Keshavarzian, Ekin Kizilkan, Tommaso Milanese, Ming-Lo Wu, Halil Kerim Yildirim, Jiuxuan Zhao, Paul Moș, Chang Liu, Chufan Zhou, Yang Lin, Vladimir Pesic, Won Yong Ha, Kodai Kaneyasu. I would like to thank Michael and Fio for their friendship. I am very happy we had a chance to explore Europe together and I thank you for taking such good care of Yuki. I would also like to thank my close friends for their constant support and patience: Codri, Didi, Iza, Oana and Csilla. Thank you for being such a supportive and fun bunch for 17 years of my life. I would also like to thank Ioana Muscoi for all the help during the last two years. This journey would have been so much more difficult without your guidance. I will finish this long list by thanking my family. I want to thank my mum, Corina, for all the strength and courage she taught me during my life and for all the help along this very long journey. I want to also thank my dad, Eugen, for his support. I would have never been the person I am today if not for my grandparents, Otilia and Valentin, who were my constant support. Thank you for teaching me all the great values in this life and for all the sacrifices you made for me. In the end, I will shout out the biggest thanks to Andrei. Thank you for all the support, encouragement and for the constant help. This journey would not have been possible without you. Another important family member is Yuki, and I would like to thank her for all the emotional support and for constantly reminding me that I have to go home at a reasonable time. What matters the most is having people who love and support you no matter what, and I am very lucky to have all these people around me.

Neuchâtel, 07 December 2022

Abstract

Positron emission tomography is a nuclear imaging technique well known for its use in oncology for cancer diagnosis and staging. A PET scanner is a complex machine which comprises photodetectors placed in a ring configuration that detect gamma photons generated through annihilation between an electron and a positron. The accuracy with which the gamma photons are detected determines the quality of the extracted information, which is then further analysed through image reconstruction algorithms. Therefore, the design and optimization of photodetectors and their readout electronics is very important for the advancements of PET scanners. For more than a decade, silicon photomultipliers have been researched extensively and became the photodetectors of choice for this application. High performance readout electronics is essential in order to measure data with high precision and various readout schemes have been proposed over the years for PET photodetector modules. As PET scanners are complex systems, research is dedicated individually to different parts.

This thesis focuses on the design of integrated readout electronics for time-of-flight PET application. Consequently, the readout of three SPAD-based sensors was designed. The first sensor, Blumino represents the first fully integrated analog silicon photomultiplier with on-chip discrimination and time conversion. The design was implemented in 350 nm CMOS technology node. The chip serves as a prototype for future fully integrated A-SiPMs as PET photodetectors due to its simplicity and compactness. The second sensor, Blueberry, advances the previous design by exploring the benefits of multi-digital silicon photomultipliers and 3D integration. The sensor was designed in a 3D-stacked FSI CMOS technology node enabling features such as: improved spatial and timing resolution (a TDC design with on-chip error correction algorithm of 15 ps LSB). The third sensor, Smarty, is an on-chip fully reconfigurable neural network with 10 TDCs designed in 16 nm FinFET technology. The chip is capable of executing 363 MOPS and was designed for pre-processing and data compression at the sensor level. Various neural network configurations were explored and trained using genetic algorithms. The architecture was proven to be viable for reconstructing radioactive source positions in a coincidence setup.

Abstract

Key words: positron-emission tomography (PET), time-of-flight (ToF) , single-photon-avalanche diode (SPAD), time-to-digital converter (TDC), artificial neural network (ANN), coincidence timing resolution (CTR), reconstruction.

Résumé

La tomographie à émission de positons (TEP) est une technique d'imagerie nucléaire bien connue pour son usage en oncologie pour diagnostiquer les cancers et identifier leur progression. Le scanner TEP est un appareil complexe contenant des photodétecteurs placés dans un dispositif circulaire et qui détectent des photons gamma émis suite à l'interaction entre un électron et un positon. La précision avec laquelle les photons gammas sont détectés établit la qualité de l'information recueillie, laquelle est ensuite analysée par des algorithmes de reconstruction d'image. Par conséquent, la configuration et l'optimisation des photodétecteurs et leurs composants électroniques de lecture sont très importants pour le développement des scanners TEP. Pendant plus d'une dizaine d'années, les photomultiplicateurs en silicium ont été l'objet de nombreuses recherches et sont devenus des photodétecteurs de choix pour la TEP. Les performances élevées des composants électroniques de lecture sont essentielles dans la mesure de données de grande précision et des typologies variées de lecture ont été proposées depuis des années pour les modules photodétecteurs TEP. Etant donné que les scanners TEP sont des systèmes complexes, les recherches ciblent les différentes parties du scanner.

Cette thèse a pour objet le design des composants électroniques de lecture intégrés pour la TEP incluant le temps de vol. Pour ce faire, la lecture de trois capteurs basés sur la diode à avalanche de photon unique (SPAD) a été créée. Le premier capteur, Blumino, représente le premier photomultiplicateur analogique intégré avec un comparateur et un convertisseur de temps (TDC) sur puce. Son design a été réalisé avec un nœud de technologie de 350 nm CMOS. Le circuit sert de prototype avec un photomultiplicateur analogique en silicium (A-SiPM) entièrement intégrés pour les futurs photodétecteurs TEP en raison de sa simplicité et de sa compacité. Le deuxième capteur, Blueberry, a une conception plus complexe qui a fait avancer le design précédent en explorant les avantages des photomultiplicateurs multi-numériques en silicium et de l'intégration 3D. Le capteur a été conçu dans une technologie de nœud FSI CMOS empilé en 3D permettant la fonctionnalité d'une résolution spatiale et temporelle avancée (design des TDC sur puce possédant un algorithme de correction d'erreurs). Le troisième capteur, Smarty, est un réseau sur puce de neurones artificiels

entièrement reconfigurable avec 10 TDCs réalisé en technologie 16 nm FinFET. La puce est capable d'exécuter 363 méga opérations par seconde (MOPS) et elle a été configurée pour pré-traiter et compresser des données au niveau du capteur. Diverses configurations de réseaux neuronaux ont été explorées et entraînées en utilisant des algorithmes génétiques. La conception a démontré être viable pour reconstruire des positions de sources radioactives dans une configuration de coïncidence.

Mots clés : tomographie à émission de positons (TEP), temps de vol (ToF), diode à avalanche de photon unique (SPAD), convertisseur temps-numérique (TDC), réseau de neurones artificiels (ANN), résolution temporelle de coïncidence (CTR), reconstruction.

Contents

List of Figures

List of Tables

List of acronyms

AI	artificial intelligence
ANN	artificial neural network
APD	avalanche photodiode
A-SiPM	analog silicon photomultiplier
ASIC	application-specific integrated circuit
BSI	back-side illuminated
CMOS	complementary metal-oxide-semiconductor
cps	counts per second
CPGA	ceramic pin grid array
CT	computed tomography
CTR	coincidence timing resolution
DCR	dark count rate
DNL	differential nonlinearity
DOI	depth of interaction
DPC	digital photon counter
D-SiPM	digital silicon photomultiplier
DSR	dual-sided readout
FDG	fluorodeoxyglucose
FF	fill factor
FIFO	first in-first out
FMRI	functional magnetic-resonance imaging
FPGA	field-programmable gate array
FSI	front-side illuminated
FSM	finite-state machine
FT	fast terminal
FWHM	full-width-half-maximum
GUI	graphical user interface
HLS	high-level synthesis
INL	integral nonlinearity
LOR	line-of-response
LUT	lookup table

LSBs	least significant bits
MPW	multi-project wafer
MRI	magnetic-resonance imaging
MSBs	most significant bits
NN	neural network
OPU	output processing unit
PCB	printed circuit board
PDE	photon detection efficiency
PET	positron emission tomography
phe	photoelectron
PMT	photomultiplier tube
QE	quantum efficiency
RTL	register transfer level
SiPM	silicon photomultiplier
SNR	signal-to-noise ratio
SPAD	single-photon avalanche diode
SPECT	single-photon emission computed tomography
SPTR	single-photon timing resolution
ST	standard terminal
TDC	time-to-digital converter
ToF	time-of-flight
TSV	through-silicon via
VBD	breakdown voltage
VCM	common-mode voltage
VCO	voltage-controlled oscillator
Vth	threshold voltage
WTA	winner-take-all

1 Introduction

Over time, medical imaging techniques have consistently evolved with numerous high-quality systems being developed to facilitate diagnosis and treatment. Positron emission tomography (PET) scanners are key medical tools heavily used in oncology. Unfortunately, the global inequities in access of PET scanners is worrying, especially in underdeveloped countries. As recently reported in [1], based on the statistical model the authors developed, it was revealed that at least 96 countries should increase their number of available PET scanners. The model takes into consideration the needs of patients with 6 different types of cancer. The high production cost of PET scanners, along with their limited available number, especially in underdeveloped countries is an issue addressed in different socio-economic studies and significant effort is dedicated towards solving it [2]–[4].

This chapter offers an overview of PET and it serves as an introduction for different concepts presented in the thesis. The thesis goals and motivation, along with the thesis contributions and structure are further presented. The main focus of this thesis is related to electronics developed for photodetectors mainly used for positron-emission tomography application.

1.1 Medical imaging

Medical imaging techniques are used to image the human body for either clinical analysis, medical diagnosis, or monitoring health conditions. Conceptually, medical imaging techniques can be classified as either structural, functional or both [5], [6]. Structural imaging is used for the visualization and analysis of the anatomical properties of the body part that is imaged. Information about different geometric quantifications such as size, volume and thickness of a particular imaged structure

can be obtained.

Functional imaging is mainly used for detecting and monitoring the metabolic processes inside the object tissue or organ being imaged, such as blood flow, chemical composition, and absorption [7], [8]. Functional imaging modalities, such as PET and single-photon emission computed tomography (SPECT) make use of different positron-emitting radiopharmaceuticals (radiotracers). Fluorine-18 fluorodeoxyglucose ($^{18}F-FDG$) which is a glucose analog, is a frequently used radiotracer in clinical oncology due to its specificity in some specific cancer types [9], [10].

There are also medical imaging techniques which can be classified as either structural or functional, for example magnetic-resonance imaging (MRI). This technique is either functional or structural, depending on which kind of information is obtained from the performed scan. In the case of structural MRI scans, the anatomical structure of a body part is imaged, while in the case of a functional MRI (FMRI) scan, metabolic process information is obtained.

1.2 Introduction to positron emission tomography

Positron emission tomography is a functional imaging technique which is heavily used in clinical oncology to diagnose and monitor the evolution of different tumors and to search for metastases inside the body [9], [10]. PET, as an in vivo functional imaging technique makes use of tracers labelled with radioisotopes. In order to conduct a PET scan, the patient is injected with a short-lived radioactive tracer isotope, the most frequently used one being $^{18}F-FDG$. $^{18}F-FDG$ concentrates in areas with high metabolic activity, which includes cancerous tumors, and an image of the area of interest is obtained.

After the patient has been injected, the radioisotope undergoes positron decay, emitting a positron. The positron travels for a short distance within the tissue (positron range), losing its kinetic energy, and then interacts with an electron in a process called annihilation as depicted in Figure 1.1. An annihilation event results in two gamma rays with similar energetic (511 keV) and geometric profiles (180 degrees), emitted in almost opposite direction in a system.

The back-to-back emission pair of 511 keV annihilation gamma-rays travels in the body and is then absorbed by scintillators. A conceptual representation of a PET detector ring is depicted in Figure 1.2. The scintillators absorb the annihilation photons, converting their energy into visible and/or ultraviolet photons. The amount of light

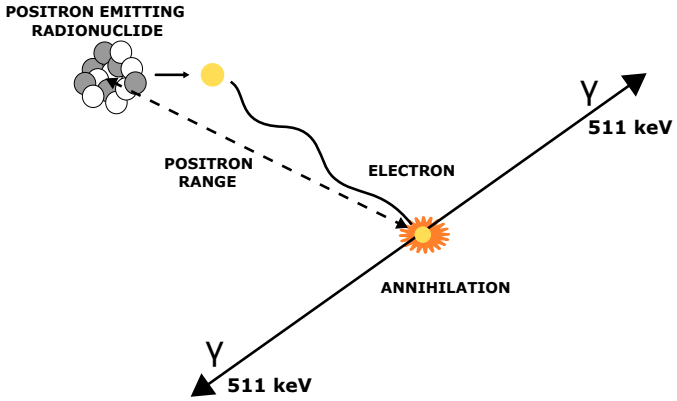


Figure 1.1: Conceptual representation of the positron range. The positron-emitting radionuclide emits a positron which travels for a short distance, called the positron range, until it annihilates with an electron and produces a pair of gamma rays. Adapted from [11].

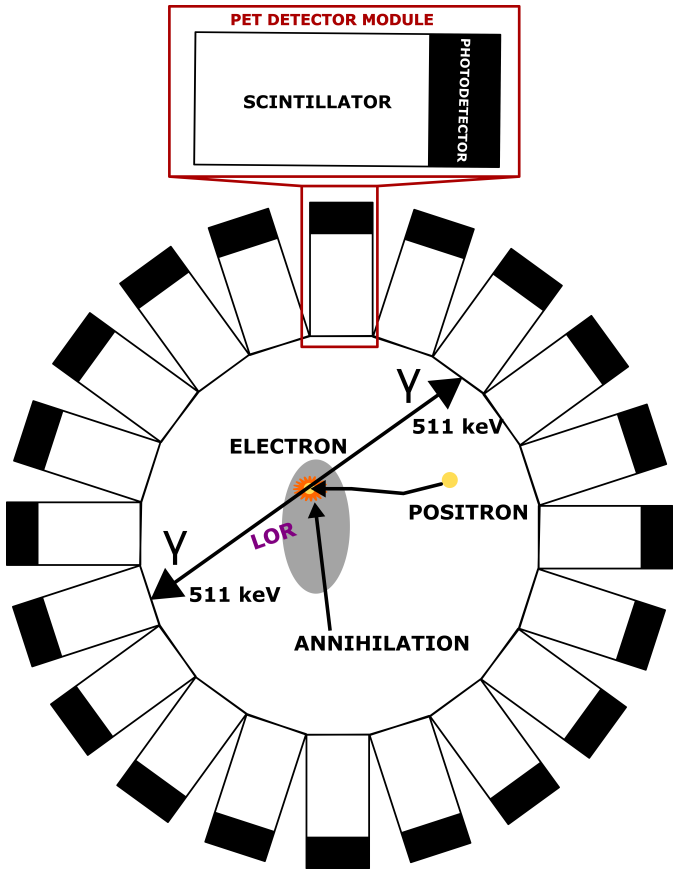


Figure 1.2: Conceptual representation of a PET ring detector. The PET detector module is comprised out of a scintillator and a photodetector.

produced by the scintillator is proportional to the amount of energy deposited in it by the incoming particle. On average, this process results in an optical pulse of approximately 10^4 photons, with a duration of tens of nanoseconds depending on the scintillator material [12]. The scintillator's emitted light corresponds to a specific emission spectrum [13]. The visible photons are further detected by the photodetectors coupled with scintillators whose role is to convert the scintillation light into electronic signals. A conceptual diagram of a scintillator coupled with a photodetector is depicted in Figure 1.3.

Considering the emission profile, if two gamma photons of 511 keV are detected in coincidence, i.e. within a time window of a few ns, the annihilation point is located along the line-of-response (LOR) between the two crystals whose detectors registered the electronic signal. A very large number of LORs has to be collected so that a tomographic image of the tracer distribution within the subject can be reconstructed [12]. Two sample images acquired in clinical studies with the Vereos PET/CT system are illustrated in Figure 1.4.

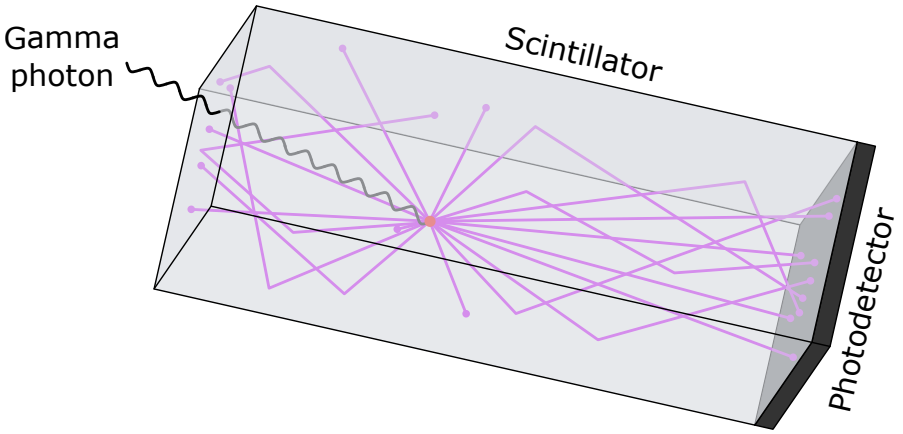
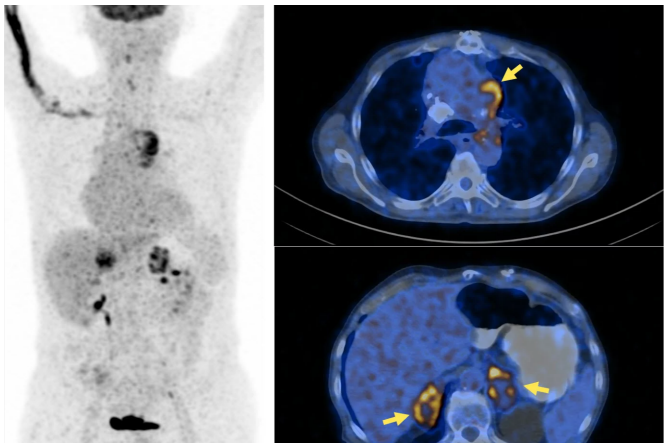


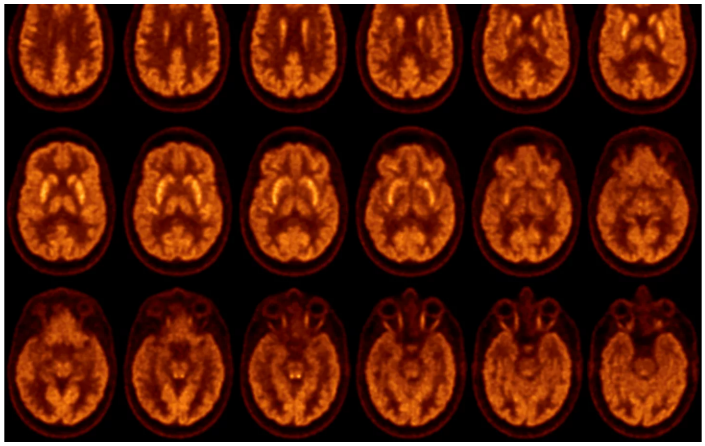
Figure 1.3: Conceptual representation of a scintillator coupled with a photodetector. The gamma photon strikes the scintillator and a large number of visible photons are produced.

1.2.1 Time-of-flight PET systems

In a non time-of-flight (ToF) PET system the annihilation point position is unknown over the entire LOR due to its uniform probability. Compared to a ToF-PET scanner where the emission point along the LOR is determined by the difference in the detection times of the annihilation photons ($t_2 - t_1$). PET scanners based on the non-ToF principle present a significant lack of precision compared to ToF-PET scanners. The



(a)



(b)

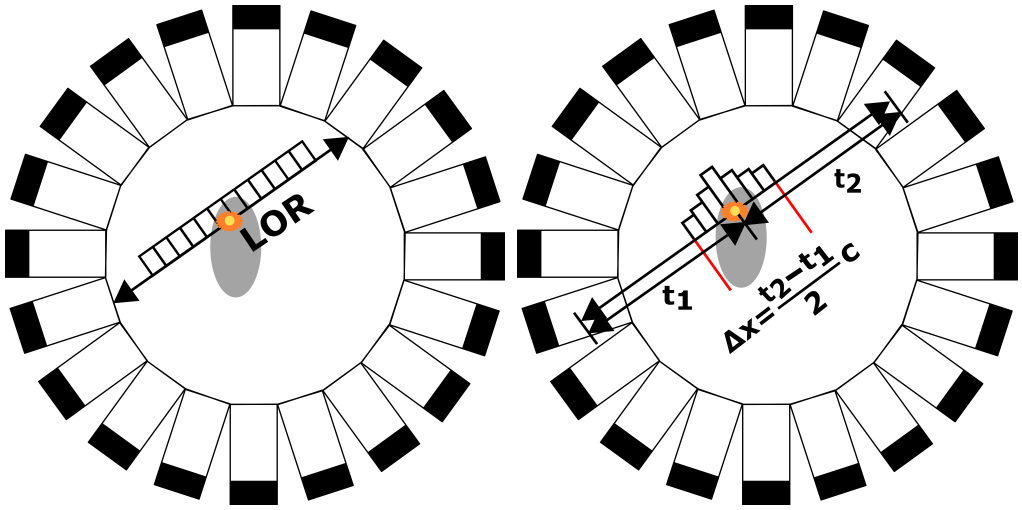
Figure 1.4: Sample images acquired in clinical studies with Vereos PET/CT system at The Ohio State University. Oncology Whole Body - Fast Acquisition a) 75 yo M, 58 kg, F-18 FDG, 81 sec total acquisition time , b) 102 kg, F-18 FDG, 15 min acquisition [14].

two principles are shown in Figure 1.5.

The position uncertainty Δx is given by the following equation :

$$\Delta x = c \times \frac{\Delta t}{2}, \tag{1.1}$$

where c is the speed of light in vacuum and Δt is the coincidence timing resolution (CTR) of the scanner [12]. CTR is one of the key parameter in a PET scanner as it determines the ability of the detectors to resolve the difference in interaction times of two gammas. CTR represents the full-width-half-maximum (FWHM) of the differences in time of the detected coincidences.



(a) (b)

Figure 1.5: a) non-ToF-PET. Emission point precision uniform distribution along the LOR. b) ToF-PET. Emission point precision along the LOR determined by $t_2 - t_1$.

Over time, CTR values have improved considerably, arriving at around 200 - 300 ps FWHM [15]–[17] in clinical systems by improving the photodetector technology as well as the scintillation material along with readout electronics. Impressive CTR values below 100 ps have been reported in experimental measurements in recent works such as [18]–[23].

1.2.2 Incertitude sources in a PET system

The main objective of a PET scanner is to produce high quality images, facilitating the correct diagnosis and treatment of different diseases. Various factors can contribute to PET image quality, such as positron range, photon non-collinearity, detector geometry, parallax error.

The positron range represents one of the physical limitations of spatial resolution in PET systems [24]–[26]. The distribution of the end points (the points where the positron-electron annihilation occurred) contributes to the spatial resolution of a PET system [25]. The positron range is also dependent on the medium where the positron is emitted. Depending on which radionuclide is used, the emitted positrons have different energies, for example ^{18}F has an endpoint energy of 0.64 MeV [26], [27]. $^{18}F - FDG$ is usually preferred due to its specificity. A conceptual representation of the positron range is depicted in Figure 1.1.

Another possible source of error that can affect the image quality is the non-collinearity of the annihilation photons. It is not always the case that after annihilation, the two photons travel in exact opposite directions (a LOR of 180 degrees). In fact, the initial two particles are often not completely at rest. In such cases, the photons will deviate slightly from a 180 degree LOR, as to obey conservation of momentum as illustrated in Figure 1.6. Even a small deviation can significantly deteriorate the image quality.

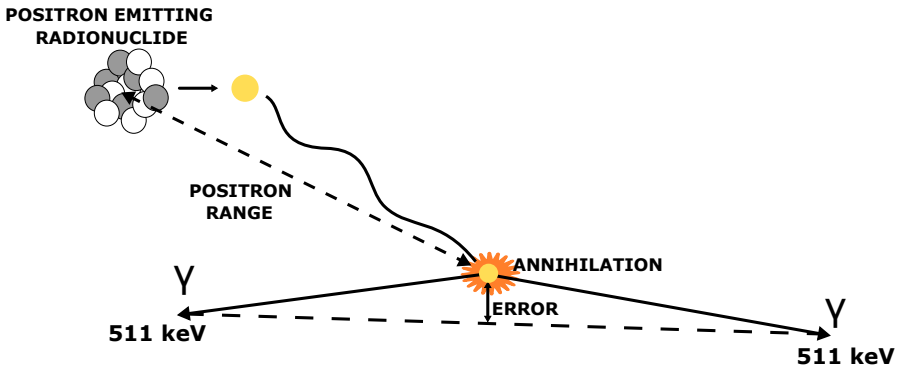


Figure 1.6: Non-collinearity representation. Two annihilation photons which travel in opposite direction with a deviation from the 180 degrees line. Adapted from [11].

Another PET source of error is the parallax error which represents the misplacement of the LORs. There are events that occur away from the center of the PET scanner [26]. As a result, these events do not arrive perpendicularly to the scintillators' entrance face as illustrated in Figure 1.7. The error of the correct estimation of the corresponding LOR depends on the depth-of-interaction (DOI) of the gamma inside the crystal.

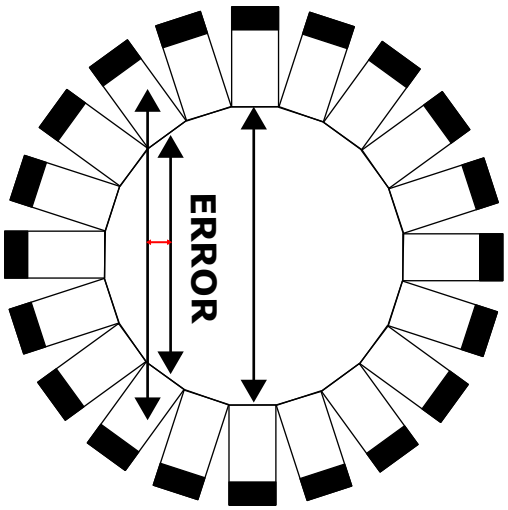


Figure 1.7: Parallax error. Misposition of LORs for events that occur away from the center of the scanner.

A low signal-to-noise ratio (SNR) can degrade the quality of an image. Assuming an analytical reconstruction algorithm, no random coincidences and a cylindrical phantom the gain in SNR is given by the following equation:

$$GAIN_{SNR} = \sqrt{D/\Delta x}, \tag{1.2}$$

where D is the size of the object being imaged [28]. The image SNR is affected by the scattered and random coincidence events [29], [30]. In some cases, the gamma loses partially its energy by scattering in the environment and this effect is called Compton scattering. Scattered events are events in which at least one of the photons has scattered inside the patient's body before being detected [12]. An energy filter is applied in order to select the 511 keV events and in some cases, the scattered events can be detected within the energy filter and registered as valid coincidences. An example of scattered events is illustrated in Figure 1.8.

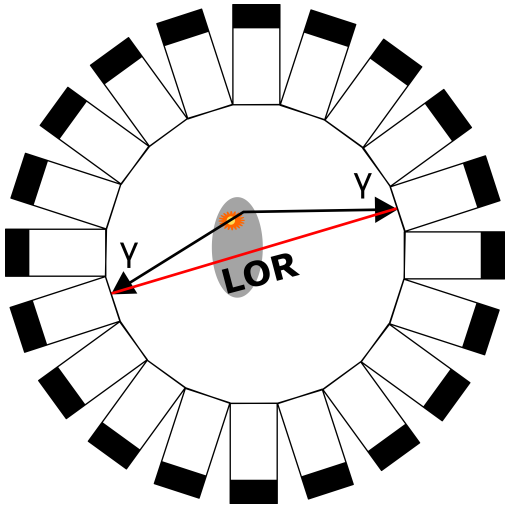


Figure 1.8: Example of a scattered event. The two gammas are deviated from a 180 degrees LOR due to scattering.

Another error source comes from the random coincidence events. Random coincidence events do not originate from the same annihilation event, however, they can be registered as being valid coincidences. An example of how random events can look like is depicted in Figure 1.9. Statistical fluctuations of these events worsen the image SNR in a PET scanner.

The image SNR can also be influenced by other factors, such as the scan acquisition time, the amount of injected radiotracer, and the system sensitivity. System sensitivity is defined as the detected fraction of coincidence events. A higher system sensitivity translates into a lower injected radiation dose; a very desirable attribute of

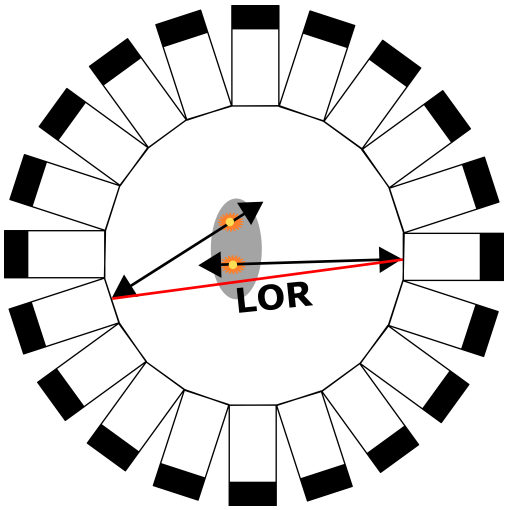


Figure 1.9: Example of two random events. The events have different origins but they are accounted for as being part of the same event, resulting in a wrong LOR being registered by the system.

PET scanners.

1.2.3 Inorganic scintillators for ToF-PET

Over the past several decades, significant effort has been placed into improving scintillator performance for medical imaging [31]–[35]. Inorganic scintillators are high-density transparent crystals that convert high-energy radiation to near visible or visible light. They can be cut into small sizes and partitioned into different array configurations matching the photodetector's geometrical profile, thus enhancing the spatial resolution. They can also exhibit very high light yield which is desired in order to increase the amount of collected photons, while a short scintillation decay time allows higher count-rate capability at system level [12].

The place where the gamma interacts inside the scintillator affects the timing resolution and two extreme cases are presented in Figure 1.10. In the first case, the interaction occurred near the scintillator surface and the photons travel a longer distance in order to reach the photodetector. The propagation speed of an optical photon at a particular wavelength (λ), in a medium with a refractive index n is given by the following equation:

$$v(\lambda) = c/n(\lambda). \tag{1.3}$$

In the second case, the interaction occured near the detection surface. Therefore, the gamma photon travels at the speed of light inside the scintillator until it reached

the photodetector. The uncertainty that is introduced by the point of interaction of gamma inside the scintillator is given by the DOI and it is dependent on the scintillators' optical properties and size [36]. Evidently, in the case of long scintillators, the effect of the timing jitter is much more evident than in the case of short scintillators. This difference translates into jitter that can be reduced by reducing the scintillator length or by using different detector geometries and DOI-correction algorithms [12]. Different techniques for DOI estimation have been researched over the years, for example, dual-sided readout (DSR) in which photodetectors are coupled with scintillators on their opposite faces or phoswich detector, which is a combination of scintillators with different pulse shape characteristics optically coupled with each other. The place where the event occurred is determined by analyzing the scintillators' resulting pulse shapes. Various techniques, which account for the DOI estimation are presented in [37]–[40].

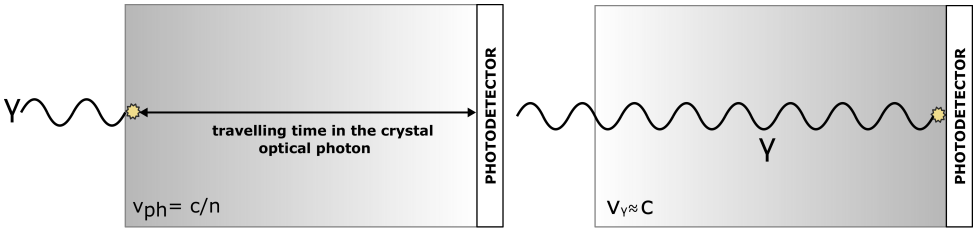


Figure 1.10: a) Gamma interaction occurs at the entrance in the crystal (longer travelling time in the crystal of the optical photons). b) Gamma interaction occurs at the output surface of the crystal. Travelling time in the crystal of the optical photons is much shorter. Figure concept derived from [41].

Although they have many advantages, PET scintillators are ultimately integrated into a system, with various physical constraints, one of which is the requirement that the emission spectrum should match the photodetector's photon detection efficiency profile. Moreover, the refractive index mismatch between the scintillator and photodetector influences the total number of collected photons, which in turn affects energy resolution. To improve the measurement precision, a large number of collected photons is required [42].

1.2.4 ToF-PET photodetectors

Ideally, ToF-PET photodetectors should present high internal gain, high photon detection efficiency, low noise and low timing jitter. All of these contribute to the CTR improvement.

The photodetector cost plays an important role at system level and it should be taken into consideration. Significant research has been dedicated towards making PET systems available and accessible worldwide. However, unfortunately the cost of a PET system is still very high. The cost per unit area is indeed an important consideration that should be accounted for in any PET system [43]. For a very long time, photomultiplier tubes have been the detectors of choice for PET systems. However, they have been gradually replaced by semiconductor based photosensors. In the following sections, these two types of photodetectors are presented.

Photomultiplier tubes

A photomultiplier tube (PMT) consists of a photocathode, several dynodes and an anode, as depicted in Figure 1.11.

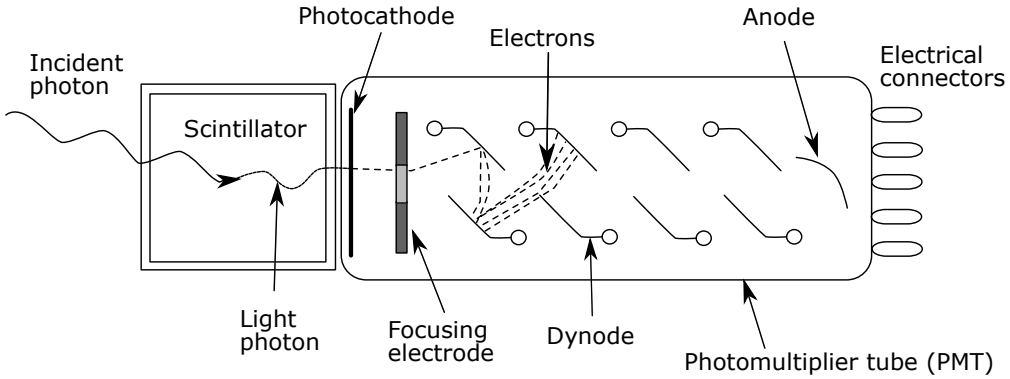


Figure 1.11: Photomultiplier tube block diagram [44].

Due to the photoelectric effect, electrons are produced upon the striking of incident photons into the photocathode region. The focusing electrode, which is placed right after the photocathode, directs the emitted photoelectrons towards the first dynode. Due to secondary emission, more and more low energy electrons are produced. The dynodes can be biased at different voltages and a large number of photoelectrons are produced. To improve timing performance, a higher voltage is typically applied to the first dynode so that the electric field is increased and only small variations are present in the trajectories of the first photoelectrons [42]. The last stage of a PMT is represented by the anode that collects all the multiplied secondary electrons emitted from the last dynode. The output is a sharp current pulse that is then detected by the readout electronics.

PMTs exhibit a large gain (approximately 10^6 or better) and quantum efficiencies (QE) (i.e. the ratio of the number of photoelectrons emitted from the photocathode to the number of incident photons [45]) in the range of 1 - 40% in the ultraviolet

and visible spectral region, with a significant drop in the near-infrared range. The PMTs which are compatible with PET scintillators' emission spectrum exhibit a QE of approximately 30% at 420 nm [42], [45]–[47].

As previously mentioned, the optical coupling between the PMT and scintillator is crucial because it introduces additional timing variations. Differences of the photoelectrons' path lengths and fluctuations in the photoelectron multiplication are factors that influence timing resolution. The PMT fabrication process has been greatly improved over time and fast PMTs have been developed by multiple manufacturers [48], [49].

PMTs have been gradually replaced over time by solid-state photodetectors. Their main limitations are given by the high sensitivity to magnetic fields, bulkiness and the operation at high voltages. Due to the high interest in combining PET and MRI scanners together, their sensitivity to magnetic fields represents a significant drawback.

Semiconductor-based photodetectors

Semiconductor based photodetectors gained interest during the last decade as a good replacement of conventional PMTs for ToF-PET photodetectors. Due to their robustness, high photon detection efficiency (PDE), low noise, low operating voltages, and, very importantly, insensivity to magnetic fields, solid-state photodetectors are excellent candidates for PET and PET-MRI systems.

Initially, p-i-n photodiodes and avalanche photodiodes (APDs) were explored as photodetectors for PET. However, due to their low amplification and large output capacitance, which results in an output signal rise time of tens of ns [43], a sufficiently high timing resolution could not be achieved for ToF-PET applications. During the last decade, single-photon avalanche diodes (SPADs) have proven to be a very attractive replacement of conventional PMTs.

SPADs are solid-state photodetectors, which are compatible with complementary metal-oxide-semiconductor (CMOS) technology nodes. A SPAD is reverse biased well above the breakdown voltage (VBD) with an excess bias voltage Vex so that it operates in Geiger mode. Due to the high biasing voltage, a single charge carrier arriving in the depletion region can generate a large avalanche that leads to an exponentially increasing current. The SPAD current-voltage characteristic is illustrated in Figure 1.12.

In order to stop the avalanche, which can damage the SPAD, quenching circuits are

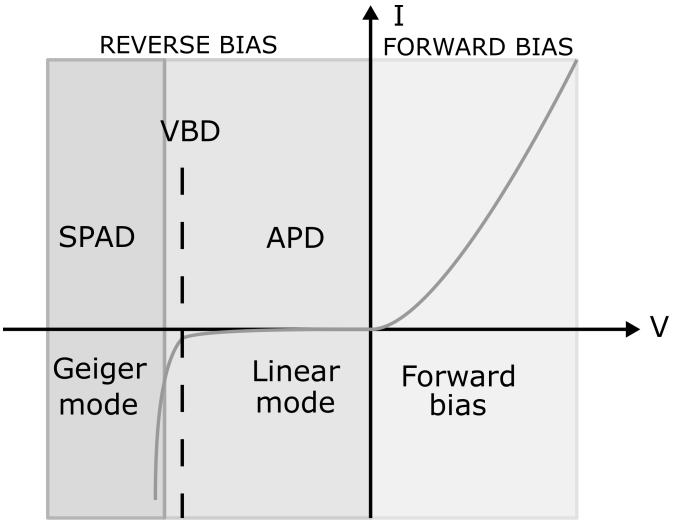


Figure 1.12: SPAD current-voltage characteristic.

used such as resistors connected in series with the SPAD. In this manner, the avalanche current self-quenches due to the voltage created across the ballast load resistor. More complex active quenching circuits can be used as well, which typically comprise discriminators that sense the high current rise and then automatically reduce the bias voltage below breakdown, thus stopping the avalanche.

There are different parameters that determine the SPADs' performance. These parameters are briefly introduced in the following paragraphs. Only the metrics that are of relevance will be introduced in the following chapters. This thesis makes use of two chips based on commercial analog silicon photomultipliers which are already characterized by the foundry and one chip based on a digital silicon photomultiplier. Therefore, the photodetector design is beyond the scope of this thesis, however, a short introduction is presented as a basis for further results which are mentioned in the next chapters.

Dark count rate

Dark counts are the spurious counts that occur in the absence of any photons impinging on the SPAD's active region. The rate at which these events occur represents the dark count rate (DCR) and it can be expressed counts per second (cps). Dark counts can deteriorate the timing resolution of SiPM based photodetectors and readout techniques have to accommodate for this, therefore, different DCR filtering methods have been proposed in literature [12], [50]–[53]. Significant research is dedicated towards improving the noise performance of SPADs in different technology nodes. A more

detailed explanation of the effects of dark counts in SPADs is presented in [54].

Photon detection efficiency

The photon detection efficiency defines the probability of detecting an impinging photon. The PDE is calculated as follows:

$$PDE = QE \times P_{avalanche} \times FF, \tag{1.4}$$

where QE is the quantum efficiency, $P_{avalanche}$ is the probability of having an avalanche and FF is the fill factor. The QE is the ratio between the number of photocarriers that are generated in the depletion region, to the total number of photons impinging on the active area. The FF is defined as the ratio between the sensitive area and the total device area. Silicon-based photodetectors are chosen differently taking into consideration the required wavelength sensitivity for each application. In PET, the photodetector's sensitivity should ideally match the emission spectrum of the scintillators. A high sensitivity is desired in order to increase the number of collected photons.

Crosstalk

Crosstalk is caused by the formation of secondary avalanches in the neighboring SPADs as a result of an initial avalanche. There are two types of crosstalk, optical and electrical. In the case of optical crosstalk, when a device detects a photon, secondary photons can be generated by the SPAD itself and can be further detected by the neighboring SPADs. The optical crosstalk increases with a decrease of the distance between the SPADs. Due to the high demand of producing more dense arrays, careful design is needed in order to minimize the optical crosstalk. Electrical crosstalk arises due to the capacitive coupling between anode or cathode traces of different SPADs. Through proper chip and PCB design, this type of crosstalk can be reduced. The crosstalk effect is described into more detail in [55]–[58].

1.2.5 From a single SPAD to an array of SPADs

Silicon photomultipliers (SiPMs) are semiconductor SPAD-based arrays. There are two main types of SiPMs: analog and digital (A-SiPMs and D-SiPMs). A-SiPMs are arrays of SPADs, commonly referred to as microcells, connected in parallel and whose output currents are summed up into one node as illustrated in Figure 1.13.

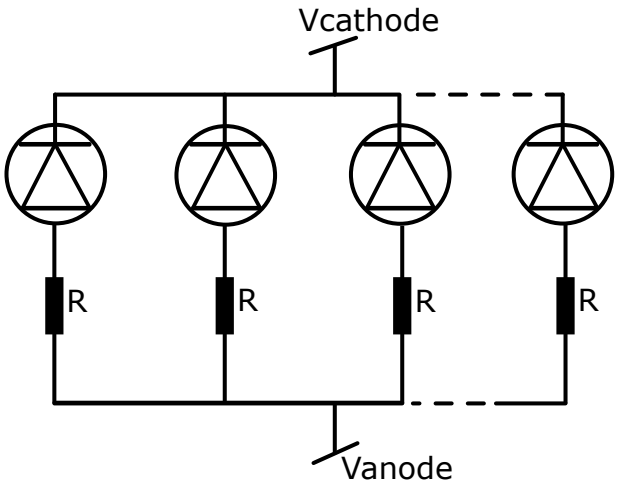


Figure 1.13: Schematic diagram of an analog silicon photomultiplier. All SPADs are connected in parallel with their corresponding quenching resistors R.

The amplitude of the output pulse is proportional to the number of photons impinging on the photosensitive area. A quenching resistor is integrated for each SPAD, which together with the additional spacing between cells reduces the FF compared to PMTs [42], [48], [49].

A-SiPMs exhibit several noise sources that need to be considered, such as dark counts, crosstalk and afterpulsing. Afterpulsing is a correlated noise source caused by trapped charge carriers in the silicon lattice that can be released after a few nanoseconds and can trigger another avalanche if the SPAD's excess bias voltage has been recharged. Despite all of this, an A-SiPM's transient response to a single-photon, which is tens of picoseconds, is significantly better compared to PMTs, but it degrades with an increase in the active area. Single-photon timing resolution (SPTR) state-of-the-art values smaller than 100 ps FWHM have been reported for analog SiPMs with an active area of 3×3 mm^2 and 4×4 mm^2 [20], [34], [59].

A-SiPMs are available in different sizes and a number of highly developed A-SiPMs can be sourced off-the-shelf. Because the output pulse has a great influence on the timing performance of the system, SiPMs which present a sharp rise time and short decay are preferred for ToF-PET photodetectors. However, in order to handle such a fast pulse, the readout electronics needs to be carefully designed. Newly developed ToF-PET scanners based on commercial SiPMs achieve a timing resolution of approximately 200 ps FWHM [43].

D-SiPMs represent a different approach of arrays of SPADs where the output sig-

nals are directly processed on-chip by additional readout circuits and converted into digital signals. A general representation of a D-SiPM is depicted in Figure 1.14.

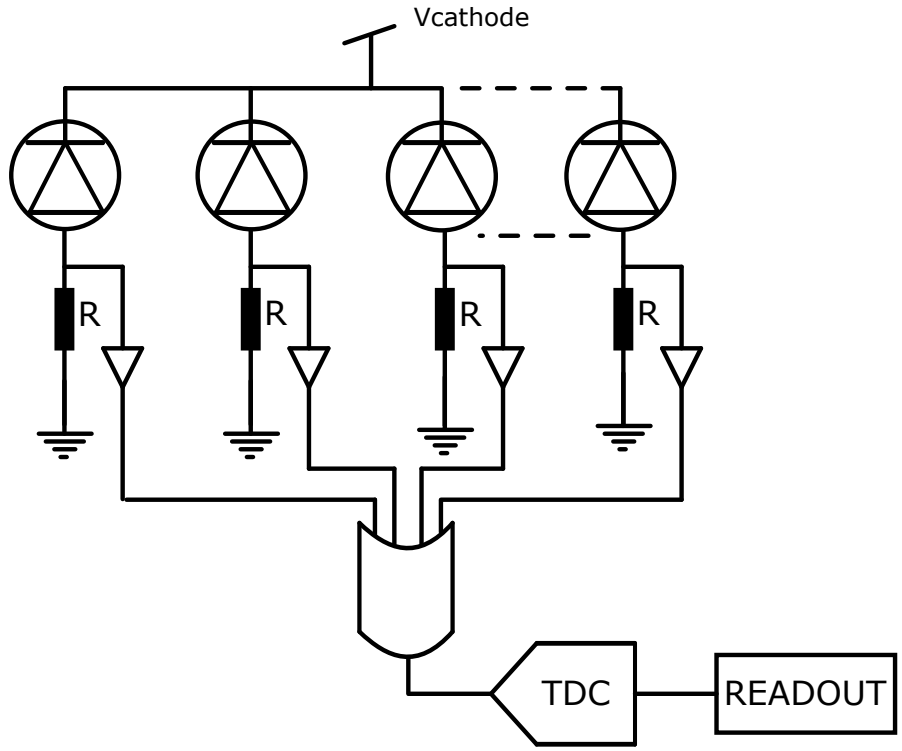


Figure 1.14: Schematic diagram of a digital silicon photomultiplier. The output of each SPAD is digitized and sent to a TDC.

The most common readout circuit encountered in D-SiPMs is the time-to-digital converter (TDC) which measures the time distance between a start and a stop signal. Different sensors based on D-SiPMs present various topology architectures with TDCs coupled in different ways with SPADs: TDCs coupled with a cluster of SPADs, one TDC per pixel or TDCs coupled with SPADs by column as depicted in Figure 1.15.

D-SiPMs can possess additional noise sources, such as jitter caused by the clock distribution network. Due to the geometrical arrangement, signal skews in the signal distribution trees can greatly degrade the timing performance. However, the propagation delay is deterministic and can be compensated for.

Depending on its architecture and design technology, TDC can occupy a large area. In 2D designs, this can lead to a large reduction in fill factor. One possible solution is the design of a 3D architecture, for which the fill factor is preserved, as the readout electronics and SPADs are placed on different tiers. Philips has developed a commercially available PET scanner based only on D-SiPMs [60]. The digital photon counter

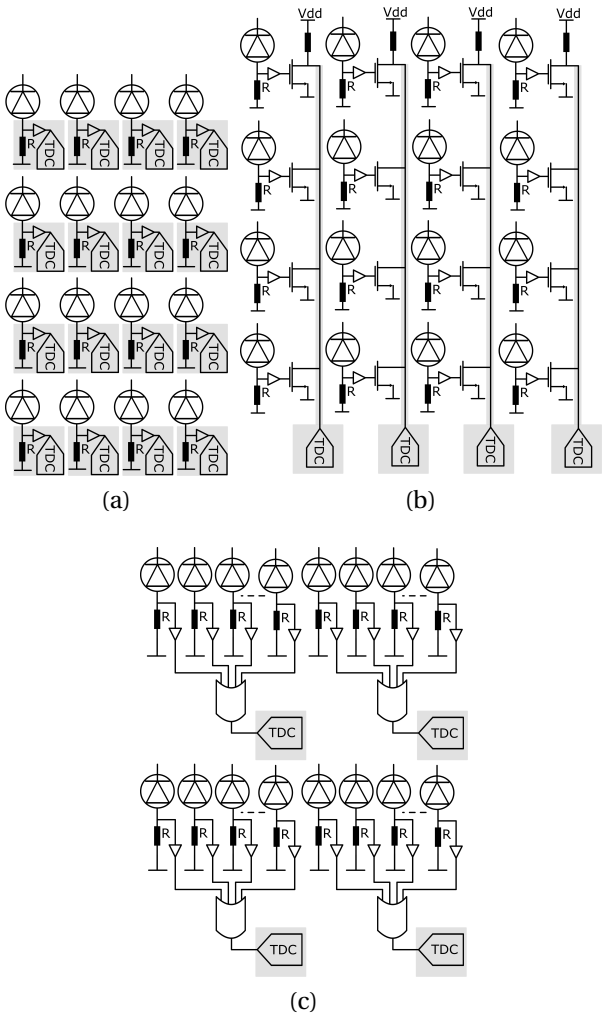


Figure 1.15: 2D D-SiPM architectures with a) one TDC per pixel, b) one TDC per column, c) one TDC per cluster.

(DPC) technology is capable of detecting and counting individual SPADs on-chip. The sensor contains four pixels arranged in a 2 × 2 matrix and each pixel contains 3200 or 6400 cells. A photon counter is present for each of the four pixels and each sensor has a pair of TDCs which timestamp the first arriving photon. A conceptual diagram of the Philips DPC is presented in Figure 1.16.

Another attractive sensor implementation based on solid-state photodetectors is 3D-integrated CMOS sensors. Recently, 3D-CMOS integration gained a lot of interest due to the capability of integrating highly optimized dedicated SPAD CMOS technologies with fast timing electronics developed in smaller CMOS technology nodes. A conventional 3D CMOS sensor comprises the top tier, which is dedicated

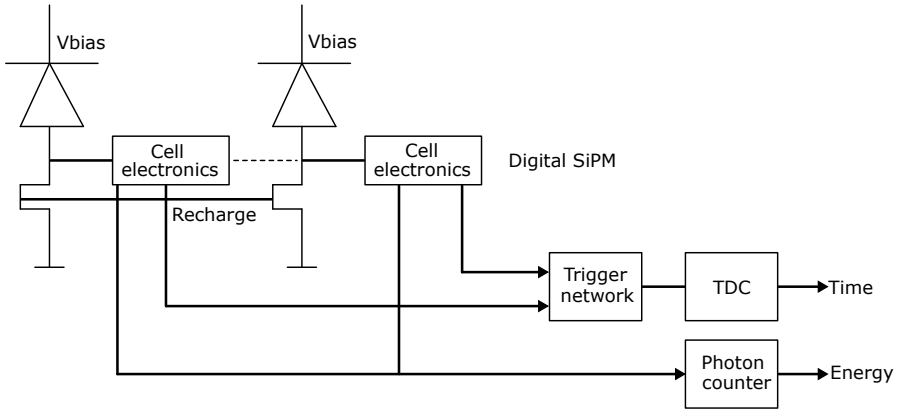
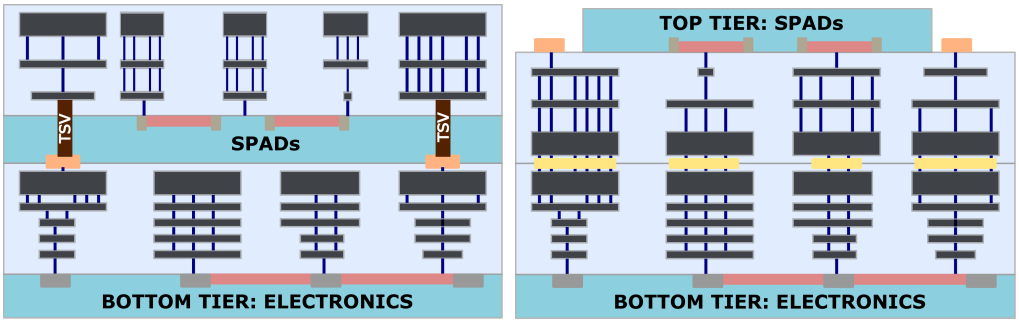


Figure 1.16: Schematic diagram of the DPC. Figure reproduced from [61].

to the photosensor only and the bottom tier, which is solely dedicated to electronics. Because all the SPAD circuits are located on the bottom tier, the FF is greatly improved.

There are two ways of implementing 3D CMOS SPAD-based sensors, namely, front-side illuminated (FSI) and back-side illuminated (BSI) topologies. A conceptual representation of the two topologies is depicted in Figure 1.17.



(a) (b)

Figure 1.17: a) Cross-section of 3D-stacked front-side illuminated technology. b) Cross-section of 3D-stacked back-side illuminated technology.

In the case of the BSI technology, the connection between top tier and bottom tier is realized through hybrid bonding. This means that the top tier is placed upside down and the electrical connections are made between the top metals of both tiers [62]. In this configuration, the incoming light crosses the bulk before it is absorbed in the depletion region. Because of the absorption coefficient of silicon, only long wavelengths can reach the photosensitive area and as such, these types of detectors are well suited for red and infrared imaging.

In the case of the FSI technology, both the top and bottom tiers have the same orientation and their connections are done with through-silicon vias (TSVs) implemented in the top tier. Due to the manufacturing process of the TSVs that fixes a limit on their form factor, the top tier has to be thinned down significantly if a small pitch is desired [63]. In this case, the incoming photons need to cross the multiple metal and dielectric layers before being absorbed in the photosensitive area, unless the layer stack on top of the SPAD is etched away and a cavity is formed. As a result, higher energy photons can be detected, which makes this approach better suited for blue wavelengths, therefore for PET photodetectors.

Both analog and digital silicon photomultipliers have advantages that make them suitable PET photodetectors. However, A-SiPMs are frequently used in PET systems due to their off-the-shelf availability. The specifications of A-SiPMs in different research fields, not only PET, have been greatly improved over time along with the performance of readout circuits. All these aspects frequently make the A-SiPMs the photosensors of choice for PET systems.

1.3 Main parameters of a PET system

The main goal of a PET scanner is to deliver high quality images of the tracer distribution inside the body. The final image quality is influenced by the quality of the collected data. As previously described, false coincidences can significantly degrade the image quality, therefore, the system should be able to discriminate and discard false events. Several parameters contribute to this target and will be described in the following subsections.

1.3.1 Energy resolution

The energy resolution represents the scanner's capability of discriminating the 511 keV incident gamma photons and discarding the rest. The energy resolution is measured by acquiring the energy spectrum of a positron-emitting radioactive source and determining the ratio between the FWHM of the energy peak and its position as illustrated in Figure 1.18. A narrow energy filter is applied so that lower energy events are rejected.

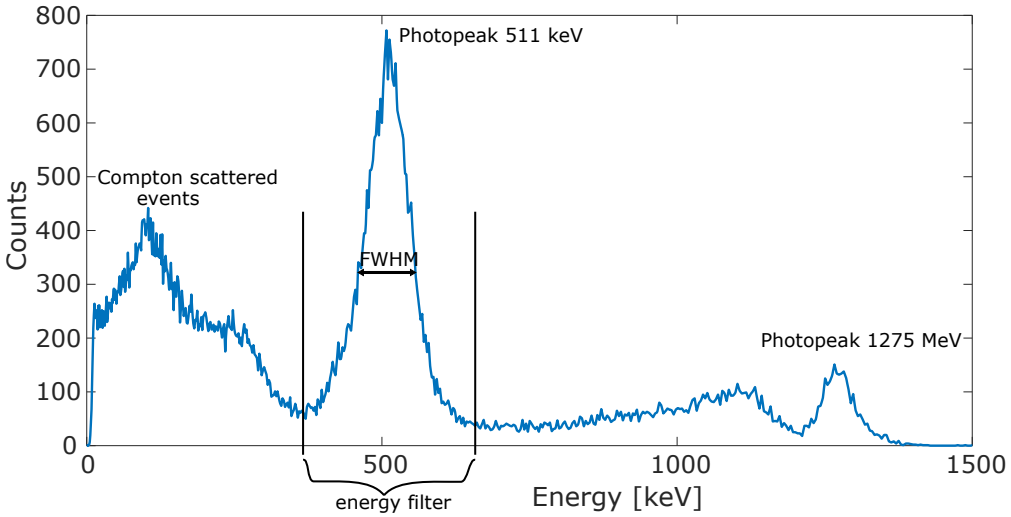


Figure 1.18: ^{22}Na spectrum acquired with an analog silicon photomultiplier presented in this thesis in Chapter 2. The 511 keV energy peak is illustrated, along with the energy filter.

1.3.2 Spatial resolution

The spatial resolution is limited by different individual contributors and it is generally assumed that all the contributors add in quadrature, although some of the effects, such as the positron range, are not described by a Gaussian and they may not be statistically independent. The spatial resolution for a point source located at a radius r from the center of the ring is described by the equation below:

$$\Gamma = 1.25 \times \sqrt{(d/2)^2 + s^2 + (0.0044R)^2 + b^2 + (12.5 \times r)^2/(r^2 + R^2)}, \tag{1.5}$$

where d is the crystal width, s is the positron range, b is the crystal decoding error factor, r is the distance between the source position to the center of the ring and R is the radius of the detector ring [64], [65]. The positron range contributes differently to the spatial resolution depending on the radioisotope that is used as presented in [66]. The non-collinearity effect adds a Gaussian blurring effect that is proportional to the radius of the ring, R. The 0.0044R factor represents the magnitude of this effect. The parallax error also degrades the spatial resolution as a function of the distance from the center of the scanner. This is mainly evident in the case of small-animal PET scanners which have a small ring diameter and use long and narrow crystals in order to enhance the scanner sensitivity.

1.3.3 Timing resolution

The PET scanner measures the time of arrival of coincidence photons within a coincidence timing window, whose FWHM is determined in such a way as to not cut out valid events. A conceptual diagram of the coincidence measurement technique is illustrated in Figure 1.19.

1.3.4 Count-rate

The scanner dead time limits the maximum operation rate. It is mainly influenced by the readout electronics and the time needed by the scintillator to convert the incoming gamma photons into visible photons and transfer them to the photodetector. Scintillators which present short decay times enhance the count-rate capability if they are coupled with fast photodetectors and readout circuits.

In a system, the dead time can be classified into two types: paralyzable and non-paralyzable. A paralyzable system is capable of processing additional events while processing of previous events is done in the background. An example of a paralyzable dead time in a PET system is the scintillator's dead time determined by the decay time constant. Conversely, a non-paralyzable system is not capable of processing multiple events at the same time and the processing is instead done sequentially and not in real time. This significantly limits the system count-rate capability. Readout electronics, in some cases, presents a non-paralyzable dead time. However, for example, a multi-shot TDC or multiple TDCs that share a detector have paralyzable dead time.

1.4 The importance of timing resolution in PET

At present, available PET scanners already achieve an impressive CTR, for example, Siemens Biograph Vision which achieves 214 ps [67], [68]. Another example is the Penn PET Explorer designed for clinical and research uses which achieves a resolution of 256 ps FWHM [69]. The design and performance characteristics of three different commercially available PET scanners are presented in Table 1.1.

The timing performance of the detectors is an important parameter in PET. Currently, it is possible to obtain a better localization of the gamma annihilation point along the LOR by measuring the difference in time of arrival of a pair of annihilation photons. Significant research is dedicated towards improving the CTR value which will lead to better image SNR, and therefore, higher image quality. The timing resolution in PET is dependent on the timing jitter introduced by different components such as the scin-

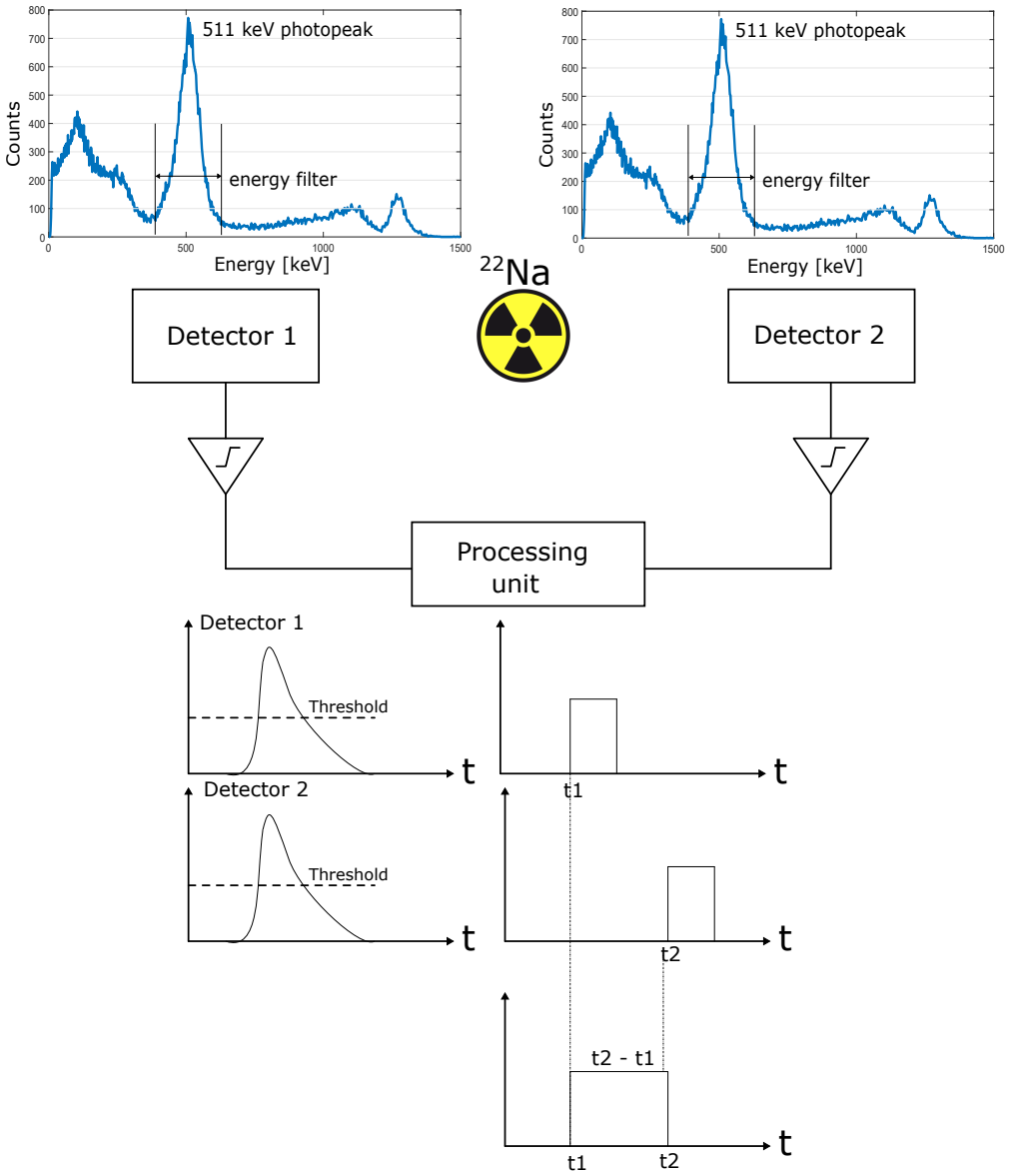


Figure 1.19: ^{22}Na source is placed between two back-to-back detectors (Detector 1 and Detector 2). The spectrum of the radioactive source is measured with each detector and an energy window which selects the 511 keV is applied. A comparator is used for each detector to digitize the output pulse and send it to the processing unit. The processing unit marks the events that correspond to the energy window and calculates the difference in time between the events.

Table 1.1: Selected commercially available PET scanners.

PET scanner	BiographVision Quadra [70]	Vereos [71]	uExplorer [72]
Photodetector	A-SiPM	DPC	A-SiPMs
Scintillator size	$3.2 \times 3.2 \times 20$ mm^3	$4 \times 4 \times 19$ mm^3	$2.76 \times 2.76 \times 18.10$ mm^3
Scintillator type	LSO	LYSO	LYSO
Axial FoV	106 cm	164 mm	194 cm
ToF	$\leqslant$ 228 ps	310 ps	409 ps
Nr. of detectors	NA	23040	53760

tillator, the photodetector and the readout electronics. Improvements in all of these will result in a better CTR value. The replacement of PMTs by SiPMs in commercial ToF-PET systems already brought a timing resolution improvement. In addition, the readout electronics plays a major role as well. In the end, the timestamps are read-out through different electronic circuits such as pixels, comparators, time-to-digital converters and PCB design. This entire chain is subject to noise and it can deteriorate the measured result.

In this thesis, three different TDCs architectures are presented with a focus on improving timing resolution. The TDCs are based on ring oscillator architectures and they are implemented in three different technology nodes, such as 350 nm, 180 nm and 16 nm. Their architectures are discussed extensively in each chapter, along with the sensors' design.

Following, the TDC's metrics that are used in the next chapters are presented.

1.4.1 TDC metrics

In this section, the TDC's metrics that are discussed in the next chapters will be briefly introduced.

LSB

The least significant bit (LSB), also known as bin size, represents the minimum time difference that can be measured in one shot. The LSB representation is depicted in Figure 1.20 and is an important parameter when determining the overall performance. The LSB of the TDC can be improved through different design techniques such as pulse shrinking [73], [74], Vernier delay line [75]–[77] or designs based on ring topologies

which achieve sub-gate delay resolution [78]–[80]. Another important aspect is the technology in which the TDC is designed. Advanced technology nodes enable much faster TDCs with simpler architectures, therefore, better resolution. On the contrary, in the case of older technology nodes, the transistors are much slower and in order to achieve a good resolution, more complex designs are necessary.

Measurement range

The range of the TDC is given by its number of bits in combination with LSB. The TDC range represents the maximum time interval that the TDC can measure and it is determined as follows:

$$TDC_{range} = (2^n - 1) \times LSB, \tag{1.6}$$

where n represents the TDC's number of bits.

Transfer function

The TDC transfer function shows the corresponding TDC output for all possible values of the input as presented in Figure 1.20.

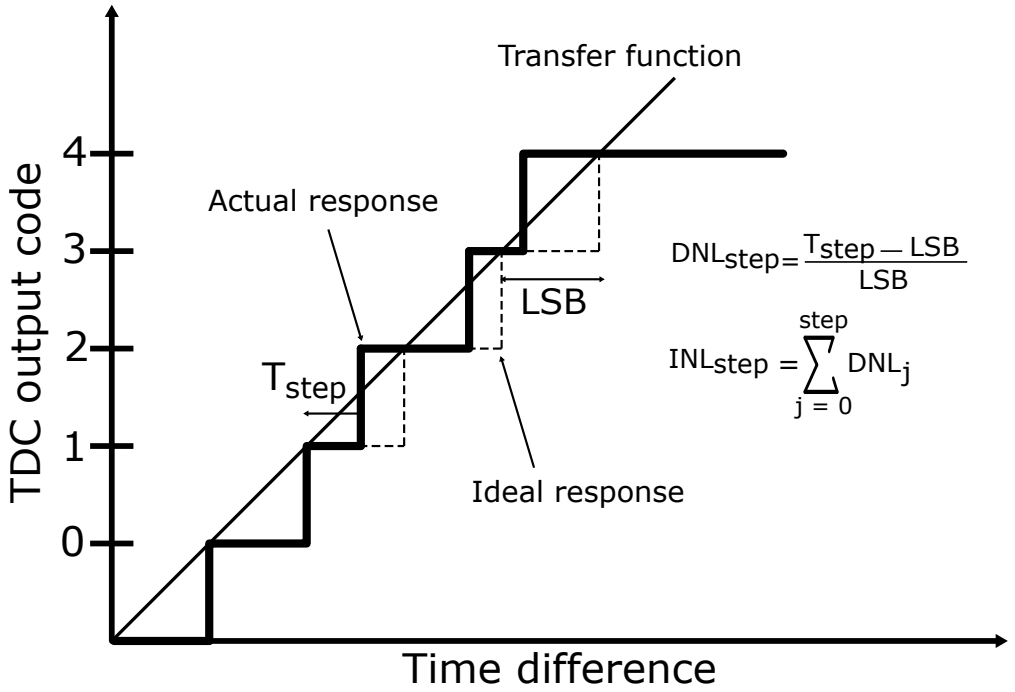


Figure 1.20: Conceptual diagram of the TDC step-plot with highlights on the DNL, INL and transfer function.

Non-linear imperfections

The TDC's nonlinearities represent all the deviations from its expected characteristics and are expressed in terms of differential nonlinearity (DNL) and integral nonlinearity (INL). The DNL describes the deviation of the value between two consecutive converter digital codes from the ideal step, while the INL is the cumulative sum of the DNL and represents the deviation of the TDC's transfer function from the ideal one. A DNL and INL representation is shown in Figure 1.20.

1.5 Artificial intelligence benefits for PET systems

The following section offers an overview on different neural network aspects and how PET could benefit from artificial intelligence based approaches.

In recent history, artificial intelligence (AI) has evolved as a very valuable asset in many areas such as economics, biology, speech recognition, facial recognition and automotive, among others [81]–[83]. AI makes use of different tools depending on the problem at hand, namely, search and optimization algorithms, logic, probabilistic methods, deep learning, artificial neural networks and others [84]–[86]. Artificial neural networks (ANNs) were initially inspired by the structure of the human brain and they were created as an attempt to solve problems that conventional algorithms struggle with. Neurons are the nervous system's fundamental structures and the connection between neurons is essential in order to exchange information between one neuron and another. In general terms, this is the main concept behind ANN.

Over time, neural networks have started being used in many fields, including the medical one. Healthcare systems deal with large amounts of data that need to be processed and neural networks are a great tool to perform this task. Neural networks for image reconstruction have been used for MRI, computed tomography (CT), and SPECT. Considering for example the emission tomography applications, artificial neural networks have already been used for a very long time for image processing [87], [88].

Image reconstruction is a completely different field in which complex algorithms are used in order to enhance the final image quality and it is a traditional way of using NNs in PET. Artificial intelligence methods based on deep networks for PET image reconstruction have been proposed in various works [89]–[92]. The improvement of the PET image quality is important in applications such as small lesion detection or early diagnosis of different neurological diseases [92] and intense research is carried out in this perspective. This will not be further discussed because it requires resource

intensive implementations that are not fit for on-chip design and therefore, it is beyond the scope of the thesis.

In what concerns the process of acquiring the ToF-PET information, conventional techniques such as leading edge discrimination or constant fraction discrimination are frequently used [93]. Estimating algorithms based on a statistical approach have been developed for time-of-flight estimation [94]. Research work is carried out using neural networks as time-of-flight estimators in PET [95]. For example, more recently, the authors in [93] present a convolutional neural network (CNN) which uses digitized waveforms as an input through constant-fraction discriminator in order to estimate the ToF information. A nine-layer CNN is trained in MATLAB with approximately 1 million coincidence events from a ^{22}Na point source at different timing delays. Cross-sectional images of a positron-emitting radionuclide were obtained directly from the coincidence annihilation photons with an average timing precision of 32 ps. In this approach, the authors make use of digitized waveforms in order to obtain the ToF information and an off-chip CNN neural network.

The authors in [96] present a study on the current limits of monolithic crystals concerning the timing resolution with the use of AI. The proposed architectures are implemented in FPGA and the NN input is provided by charges and timestamps produced by silicon photomultipliers. The proposed NN topologies utilize thousands of coefficients (10 million for event timestamping NN and 2000-20000 for position estimating NN) along with thousands of neurons (5000 - 8000 for event timestamping NN and 8000 for position estimation NN) while the readout electronics and programming is provided by custom developed ASICs.

In the context of the thesis, a different approach is used. An on-chip fully reconfigurable feed-forward neural network which makes use of TDC timestamps is used. Due to its on-chip integration, the NN implementation is limited by the available resources to 1024 coefficients and 128 neurons with the ability of changing its topology (decide the number of needed layers or neurons) within the limits. The goal of this design was to advance the use of NN in ToF PET by using raw TDC timestamps to estimate the ^{22}Na source position along one axis and minimize data throughput. The implementation of this task is not trivial because the feed-forward NN has to be able to use the raw TDC timestamps, filter them, and provide the final answer. In addition, in PET, the processing electronics handles large datasets. NNs are good candidates that can serve as pre-processing blocks in order to significantly reduce the throughput. The implemented NN is discussed in detail in Chapter 4. In the following paragraphs, a short introduction of the NN main parameters that are also used in this thesis is

presented.

A conceptual diagram of a fully connected ANN is depicted in Figure 1.21. The term *topology* will be used in this thesis very often and it refers to the neural network's structure. A typical neural network consists of several layers which are usually classified as input, output and hidden layers. The input layer is represented by the neurons which receive the external data (as presented in Figure 1.21, the input neurons are O_3, O_4, O_5, and the external data is represented by O_0, O_1, O_2), while the output layer delivers the final NN result (the output layer is formed by O_{10}, O_{11}, and the output result is OUT_0, OUT_1). All the layers in between these two layers are called hidden layers (hidden layer formed by O_6, O_7, O_8, O_9). There are different types of neural networks which come in different topologies; the one presented in this thesis focuses on a fully-connected neural network. A fully connected NN presents fully connected layers, in which each neuron of a certain layer is connected to all the neurons from the previous layer or the neurons can be organized in groups in such a way that a group of neurons is connected to one layer, while another group is connected to another layer.

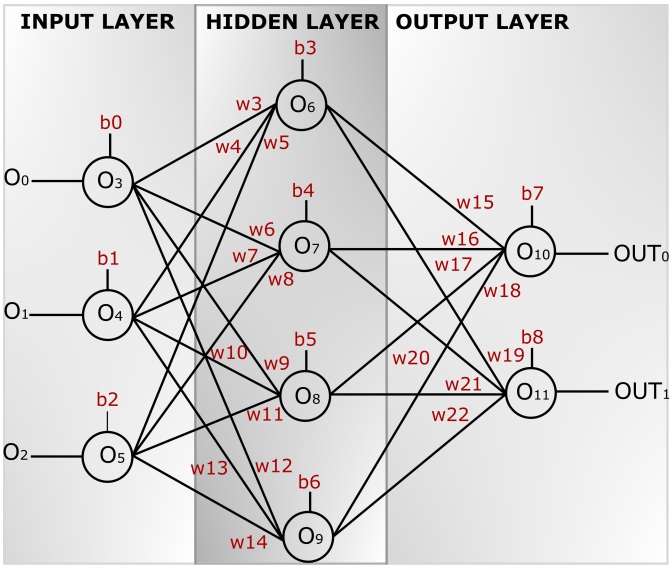


Figure 1.21: Conceptual representation of a fully connected feed-forward artificial neural network. This example depicts a three-layer NN: one input layer with three neurons (O_3, O_4, O_5), one hidden layer with four neurons (O_6, O_7, O_8, O_9) and one output layer with two neurons (O_{10}, O_{11}). The weights are represented by w3 to w22, and biases are b0 to b8.

As in a physiological brain, each neuron's function is to process the information received from all the other neurons through synapses, which are generally called connections. In a biological brain, the synapses are responsible for sending information from one neuron to another. Each neuron processes the signal by using a non-linear

function and computes an output that is further transmitted to other neurons in the ANN or is the final ANN result depending on where the neuron is located in the network. Each neuron has a bias and each connection in the ANN has a weight. The weights' role is to decide how much influence the input has on the output value, while the biases correspond to each neuron and they assure that the neuron is activated even if its value is 0. The input values are multiplied by their corresponding weights, accumulated with the neuron bias (pre-activation function), and then passed through the neuron's activation function, as depicted in Figure 1.22.

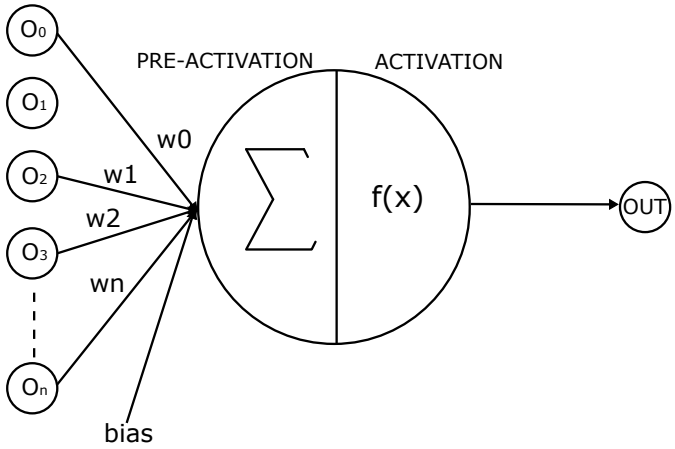


Figure 1.22: Conceptual representation of an artificial neuron. It contains the input vector $(O_0, O_1, O_2, O_3, ..., O_n)$, the bias value, the weights vector corresponding to the weights from each connection arriving at the neuron, and the pre-activation and activation functions.

There are different activation functions that are frequently used, such as: sigmoid, rectifier linear unit and hyperbolic tangent [97]–[99]. Mathematically, the neuron output can be expressed as:

$$OUT = f\left(bias + \sum_{i=1}^{n} O_i \times w_i.\right) \tag{1.7}$$

Neural networks are capable of performing different learning tasks based on the knowledge they have from datasets where the solutions are provided beforehand. Learning is the capacity of a NN to handle a problem and make observations based on the existing external data. During the learning phase, the weights are adjusted in order to improve the accuracy of the final result as much as possible. NN learning is a complex process which is crucial for the NN performance. The learning performance is evaluated with a loss function defined by the user depending on the application and it is a measure of the deviation of the NN output from the desired value.

The learning rate determines how much to change the model in response to the estimated error in each iteration and, unfortunately, the model is not trivial to choose. A high learning rate value shortens the computation time, however, it can lead to unfavorable results. A lower learning rate takes longer time but with a greater chance of delivering good results. In general, stochastic gradient descent algorithms that support adaptive learning are available in almost all the NN dedicated libraries.

While there are different types of NN learning, the author focuses on the type which is applicable on the NN described in Chapter 4, namely, supervised learning. Supervised learning uses a set of data where the inputs are paired with their corresponding outputs. This is a widely used learning technique in different applications such as pattern and gesture recognition. The goal of the NN is to make a correct prediction for each input. In the following, some NN main parameters will be briefly presented due to their frequent appearance in Chapter 4 of the thesis.

Hyperparameters, which are variables that determine the NN's structure, are important players in the NN's learning phase. In general, they are determined before learning starts, such as the number of hidden layers. However, some of them can be adjusted during the learning process itself, namely, the learning rate. These parameters have a direct impact on the NN performance. There are no strict rules or values that apply when quantifying the hyperparameters. Each problem is unique and has to be handled in different ways and the user has to decide and find out how to set the correct NN parameters. There are many available libraries and programs, which can be used in an effort to optimize the NN performance.

1.6 Thesis goals and motivation

Functional imaging technologies have an important role in the medical field serving for the early diagnosis and treatment follow-up of different diseases, such as cancer. The major role of functional imaging is to deliver good quality images that can later help doctors in the diagnostic process. Positron emission tomography is a frequently used functional imaging technique. PET scanners are complex systems and a complete design of such a system requires expertise from different fields. Therefore, research in PET can be tackled from different perspectives such as, detector design, readout, image processing, mechanical design and so on. However, PET is still an expensive medical tool with limited accessibility. During the last years, significant research has been devoted to make PET systems available and more accessible. In order to deliver qualitative information, a PET detector module (crystal and photodetector)

should have a good energy, timing and spatial resolution capability. All these can be achieved through different design approaches either at the crystal or photodetector level. Another issue is that a PET system is required to handle a large amount of data, which further has to be processed, therefore the integration of the readout could help to pre-process the data and reduce the throughput.

This thesis explores different readout and pre-processing techniques for the time-of-flight sensors required for the PET application. The primary focus has been kept on the full integration of the readout and processing electronics in order to reduce the complexity, therefore the cost.

1.7 Thesis contributions

Three different sensors have been designed, implemented and described in this thesis. Follwing, all the author's contributions of the thesis are described.

The design of the first fully-integrated analog silicon photomultiplier with on-chip time conversion was implemented in 350 nm CMOS technology node. A custom analog silicon photomultiplier with enhanced sensitivity in blue spectrum was provided by On Semiconductor. This sensor represents a combination between a custom SPAD dedicated CMOS process with a standard CMOS process. The author is responsible for developing each library component, along with its characterization, the design and implementation of the time-to-digital converter, the integration of the entire system and its full characterization. The comparator was a design and implementation work of a TUDelft master student, Ashish Sachdeva. The sensor is presented in Chapter 2.

A 3D-front-side-illuminated multi-channel digital SiPM for PET was implemented in 180 nm CMOS technology node. This sensor serves as a prototype towards future 3D-FSI sensors for PET by exploring the advantages of the 3D integration. The design of this sensor is a collaborative work between the author and Francesco Gramuglia. The author's work is related to the design, implementation, integration and readout of the 64 time-to-digital converters present in the sensor as well as the design, implementation and integration of the reset trees and SPAD address arbitration tree. The author was responsible for the characterization of the time-to-digital converter stand-alone structure. Therefore, it should be noted that the focus of this thesis is not on the SPAD design itself and pixel circuit, but the control and readout of the previously mentioned electronic blocks. The design, implementation and testing of the SPADs, pixel circuits, readout and full chip integration was a work implemented by Francesco Gramuglia. The full description of this sensor is presented in Chapter 3.

The first on-chip fully reconfigurable artificial neural network with TDC input channels for PET was developed in 16 nm FinFET technology node. The neural network is capable of working with time-of-flight information provided by the TDCs in order to reconstruct the position of the radioactive source. The first step towards the design of this sensor is an analytical model which studies the behavior of different feed-forward neural networks topologies as readout circuits for time-of-flight PET detectors. The author is responsible for the design, implementation and measurements of this entire chip. The sensor is fully described in Chapter 4.

1.8 Thesis structure

The thesis is organized as follows: Chapter 2 presents Blumino, a fully integrated A-SiPM, with a discriminator and a TDC. This chapter details the architecture, the measurements of each design block, and fully characterizes the entire sensor. Chapter 3 describes Blueberry, a 3D integrated FSI sensor. An overview of the entire architecture is given, followed by a more detailed description of the TDC's design and functionality. Following, measurement results with the TDC are presented. Chapter 4 presents Smarty, the fully integrated and reconfigurable NN with 10 TDCs. The chapter describes each design block and all measurement results for each component, along with final source position reconstruction measurements. In Chapter 5, the conclusions of this thesis are drawn along with future work recommendations.

2 Blumino: A fully integrated analog silicon photomultiplier with on-chip time conversion

2.1 Motivation

The timing performance of a PET detector module is influenced by several factors and it is one of the key parameters in a ToF-PET system, due to its direct influence on the image reconstruction quality. Although timing improvements can be observed by, for example, combining the photodetectors with custom application-specific integrated circuits (ASICs), the large capacitance between the two systems can significantly degrade the overall performance [100].

One alternative solution is to integrate the front-end circuits directly on-chip. This is very common in the case of D-SiPMs, where all the electronics is placed around the photodetectors, or in 3D system configurations where the photodetectors are placed on the top tier and the electronics on the bottom tier [101], [102]. In the case of A-SiPMs the approach of combining them with discrete ASICs, which comprise the necessary circuits to read them out, is more frequent [103]–[107]. However, one of the main limitations of such systems is the compactness, along with long development and testing cycles. Moreover, the power dissipation of ASICs is large compared to a fully integrated system [42].

Blumino has been designed in order to investigate the impact of integrating custom and standard CMOS processes together, improve timing resolution, with the additional goal of preserving the full original photodetector performance. By integrating the main optimized electronic blocks necessary to capture the timing information on-chip, e.g. discriminator and TDC, the overall timing performance of the system is expected to be improved.

The A-SiPM employed in Blumino has a third terminal in addition to the anode and

cathode, called fast terminal (FT), which is suitable for fast timing. Another advantage is represented by the reduction of the capacitive load on the fast output, which, in turn, is expected to improve the timing performance. The standard terminal (anode) is used to measure the energy of the gamma photon. The system's backward compatibility is preserved through the standard terminal (ST) which is available off-chip.

In order to implement the design, an entire standard cell library had to be created. It comprises 110 components that were used to design all the blocks of the Blumino sensor, such as: delay cells, buffers, inverters, load adaptation cells, buffer chains, latches, counters, IO pads, etc. In addition, test benches were created for each cell in order to extract their essential properties such as propagation delays and rising and falling times before utilizing them in the circuit itself. Compared to other design technologies, where the standard cells and their characterization are already provided by the foundry or third party vendors, this case required a rigorous characterization of each design cell in order to assure a predictable circuit performance post-silicon. All the cells and their layouts were implemented solely by hand, starting with the MOS transistors. This represented a significant part of the workload, which should be taken into consideration when analyzing the results. In the following subsections each of Blumino's main components is described, along with relevant measurement results. Part of the work presented in this chapter was published in [80].

2.2 System architecture

As shown in Figure 2.1, Blumino's architecture comprises three main blocks: an A-SiPM, a discriminator, and a TDC. The FPGA (Opal Kelly XEM7360 with Xilinx Kintex-7) is off-chip and serves as an external post-processing unit.

The FT of the A-SiPM is connected to the input of the discriminator through AC coupling while the ST is routed to an exposed pad and it can be coupled with external electronics. The TDC is based on a START-STOP architecture, where the output from the comparator serves as a START signal, while the STOP is provided by the FPGA. In addition, a MUX is provided to allow for the START signal to be applied externally.

A FLAG signal is triggered after every A-SiPM event, which allows the user to select only the events which trigger the TDC, thereby reducing the readout time. The TDC's range can be extended in post-processing by adding an additional counter in the FPGA that takes as an input the most significant bit of the TDC's counter.

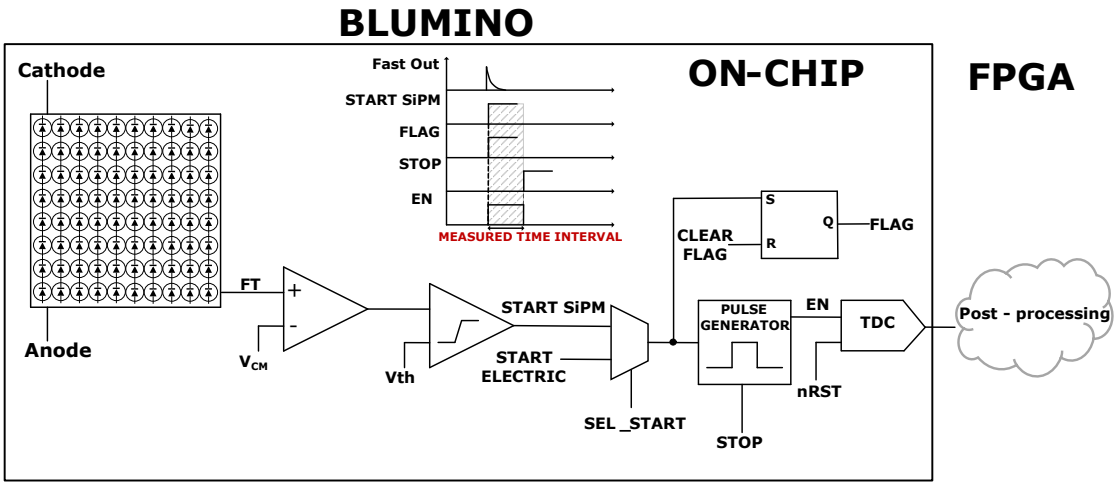


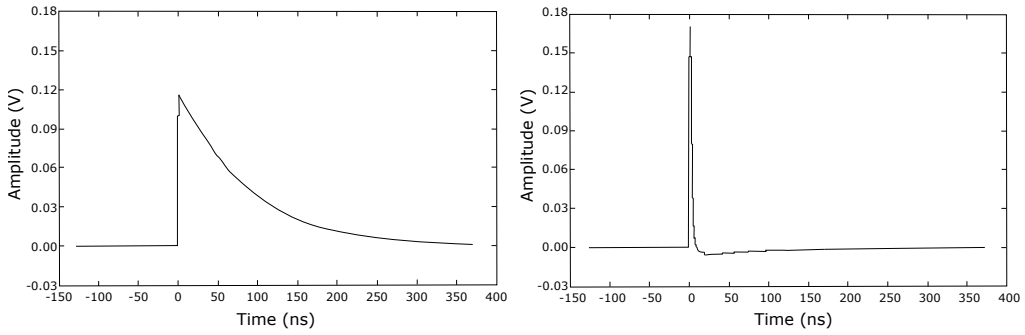
Figure 2.1: Blumino block diagram. The three main blocks are A-SiPM, comparator and TDC. The A-SiPM's fast terminal is connected to a preamplifer. The output of the preamplifier triggers the comparator whose threshold can be externally changed (by adjusting the threshold voltage (Vth)). The common-mode voltage (V_{CMA}) is 1.65 V. The comparator's output triggers the TDC. An external trigger is provided through the START_ELECTRIC signal. The FLAG signal is asserted upon the TDC trigger. Off-chip post processing is carried out by the FPGA.

2.2.1 Analog SiPM with fast output

On Semiconductor, formerly SensL, developed a unique modification to the standard A-SiPM by adding a third terminal in addition to the anode and cathode terminals, called fast terminal. Representing the time derivative of the standard terminal, the FT is proportional to the number of cells that have fired in the A-SiPM. Its amplitude is therefore proportional to the number of detected photons. The FT presents a lower output capacitance (and faster rising edge) than the ST, making it suitable for precise timing measurements and fast readout systems [108]. Compared to the ST, a signal with a very sharp rising edge is present at the output of the A-SiPM without the need of adding extra circuits, as depicted in Figure 2.2.

Blumino is populated with a C-series 3 mm × 3 mm A-SiPM with 4774 microcells (SPADs) of 35 μm each. The system features 48% PDE at 420 nm wavelength and 6.0 V excess bias. For D-SiPM systems, the associated electronics is placed next to the photodetectors. In A-SiPM systems, however, the circuits are placed around the detectors and the fill factor is not affected, as shown in Figure 2.3.

The fill factor, which is the ratio between the photosensitive area and the total chip area, is 57% due to the presence of test structures around the sensor. If only Blumino is considered (A-SiPM, comparator and TDC), the fill factor is increased to 71%.



(a) (b)

Figure 2.2: a) A-SiPM standard terminal pulse shape (anode -cathode) with 16 Ω series resistor. b) Fast terminal pulse shape. Results obtained with MicroFJ - 60035 - TSV series [108].

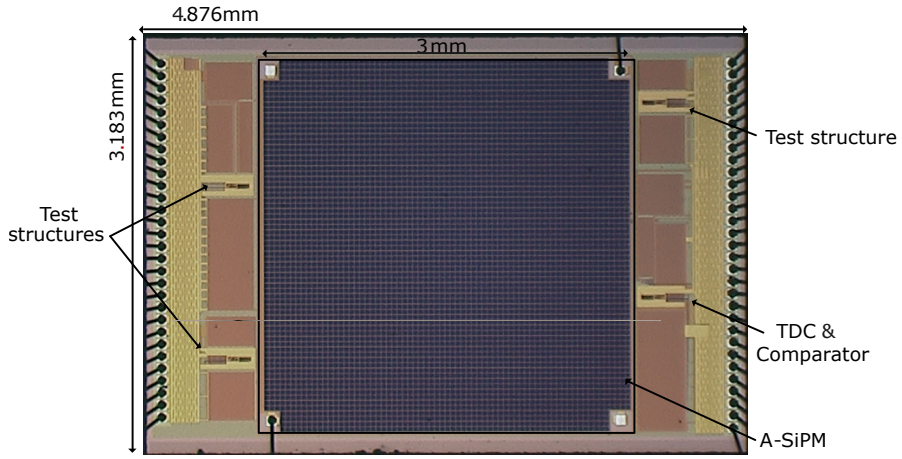


Figure 2.3: Blumino micrograph. It comprises an A-SiPM, a comparator, and a TDC. Three TDC structures are placed around the periphery of the A-SiPM for test purposes.

2.2.2 Discriminator

The interface between the A-SiPM and TDC is a preamplifier, followed by a complementary self-biased differential amplifier which acts as a comparator. It is a modified version of the design presented in [109], whereas the concept was introduced in [110] and [111]. The FT's output is connected to the preamplifier in order to increase the absolute threshold resolution with respect to the non-amplified input signal range as presented in Figure 2.4. These two blocks are described in full detail in [112].

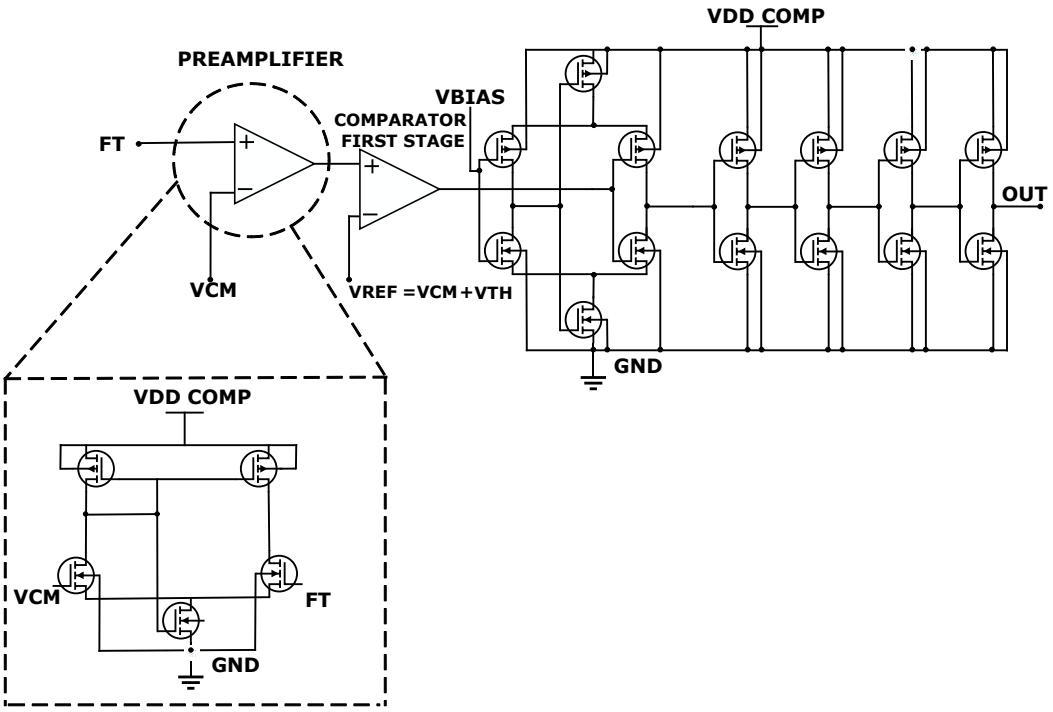


Figure 2.4: High-speed asynchronous comparator with an amplifying stage, a comparator, and a multi-stage buffer. The common-mode voltage can be changed through VCM and the reference voltage through Vref. A more detailed description of the preamplifier and comparator can be found in [112].

2.2.3 Time-to-digital converter

The TDC comprises three main structures: a voltage-controlled ring oscillator (VCO), an asynchronous ripple counter, and nine transparent phase detectors. The asynchronous ripple counter keeps track of the number of oscillations through the ring and determines the most significant bits of the TDC. The nine transparent phase detectors capture the state of each phase when the ring freezes, thus determining the least significant bits (LSB). The TDC's main components are illustrated in Figure 2.5.

The transparent phase detectors are implemented as represented in Figure 2.6. The asynchronous ripple counter architecture along with the counter cell are depicted in Figure 2.7 and Figure 2.8.

The VCO is based on a multi-path gated ring oscillator (MGRO) topology [80], [113], [114]. Each MGRO stage consists of a tri-state inverter with three inputs (three parallel inverters of the same size) as shown in Figure 2.9.

Each delay stage has multiple inputs, i.e., each one is connected to the previous delay

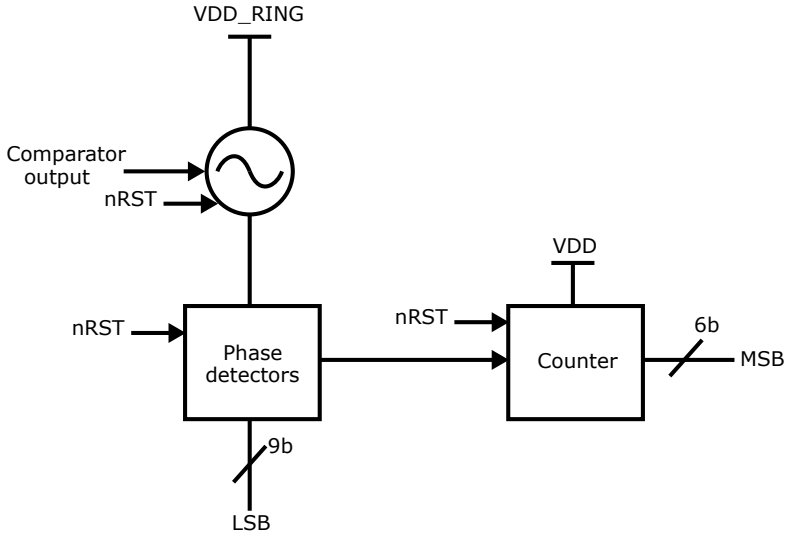


Figure 2.5: TDC block diagram. The three main components of the TDC are illustrated: the VCO, the phase detectors and the asynchronous ripple counter. The phase detectors provide 5 bits that represent the LSBs of the TDC and the counter returns 6 bits which represent the MSBs. A reset signal (nRST) is used in order to reset the TDC after each measurement cycle.

stage, and the others are connected to different stages along the ring. This structure allows each delay stage to start transitioning ahead of time, reducing the delay per stage, and increasing the maximum oscillation frequency, as depicted in Figure 2.10. The use of tri-state inverters with three inputs limits the minimum number of delay stages in the ring to nine. As a consequence, 18 phases are represented on 5-bit (LSB) after decoding, which together with the 6-bit counter (MSB) create a 10-bit result, with redundancy. As there is no decoder present on-chip, all 15 bits are transferred outside. The final TDC result is calculated as:

$$N_{result} = 18 \times N_{coarse} + N_{fine}, \tag{2.1}$$

where N_{coarse} is the counter value and N_{fine} is the decoded fine bits value.

The VCO itself comprises nine oscillation phases (Q_0 to Q_8) that are detected by nine phase detectors. The last oscillation phase (Q_8) represents the input of the ripple counter as depicted in Figure 2.5. The TDC starts running only when the VCO receives the START signal from the comparator, which together with a STOP signal sent from the FPGA forms an enable (EN) signal. The VCO starts oscillating on the rising edge of the EN signal. At the falling edge, the ring freezes in its current state which is saved in the phase detectors. After a measurement cycle, the TDC is read out in one of two ways (serial or parallel). Parallel readout results in higher speed and was used for all

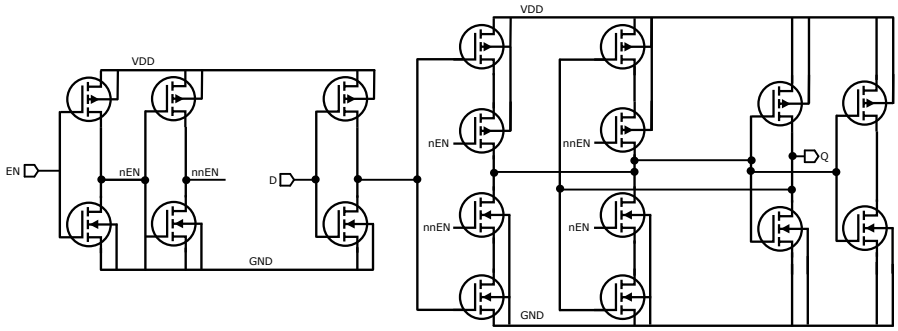


Figure 2.6: TDC phase detector.

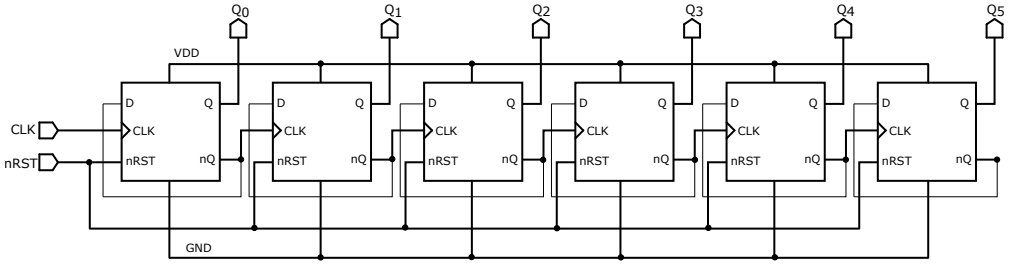


Figure 2.7: Asynchronous 6 bit ripple counter .

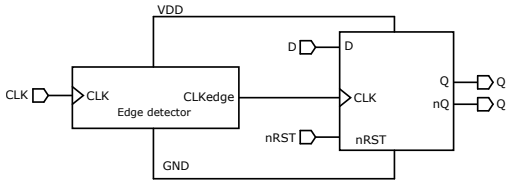


Figure 2.8: Counter cell.

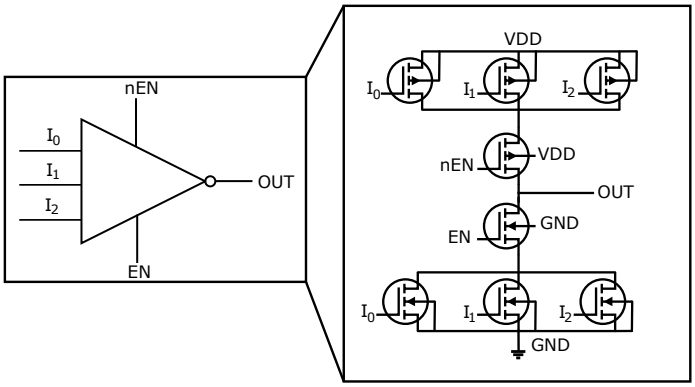


Figure 2.9: VCO delay stage. Tri-state inverter with three inputs.

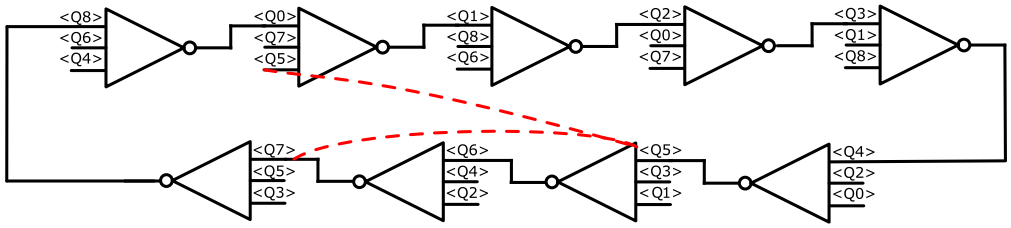


Figure 2.10: MGRO with nine, three input delay stages. Each input is connected to three different other locations along the ring.

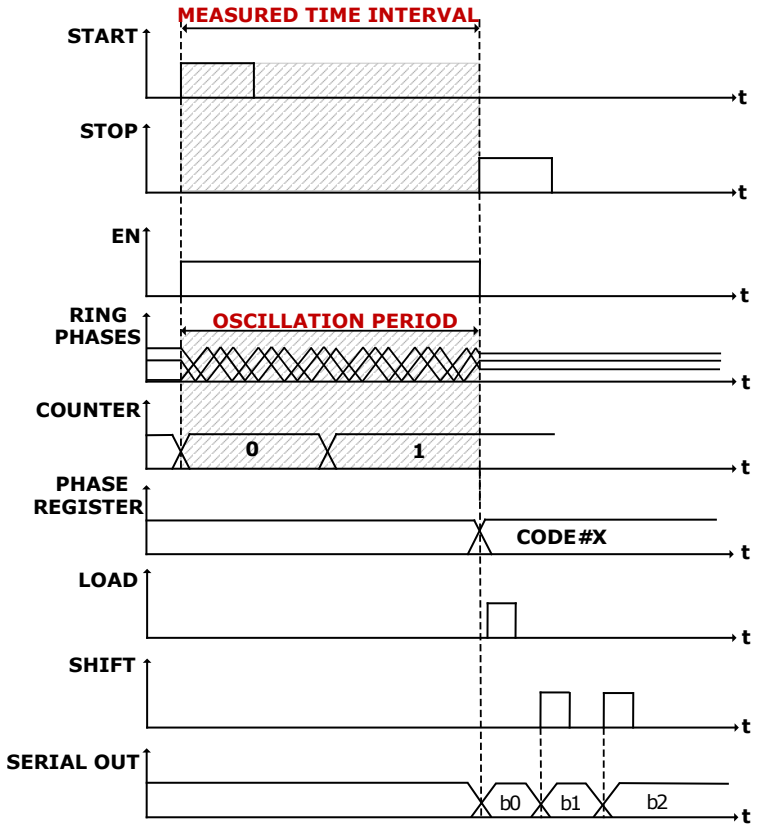


Figure 2.11: TDC operating principle. The measured time interval is illustrated along with the read-out sequence.

performend measurements. The TDC's operating principle is illustrated in Figure 2.11.

2.3 Design considerations

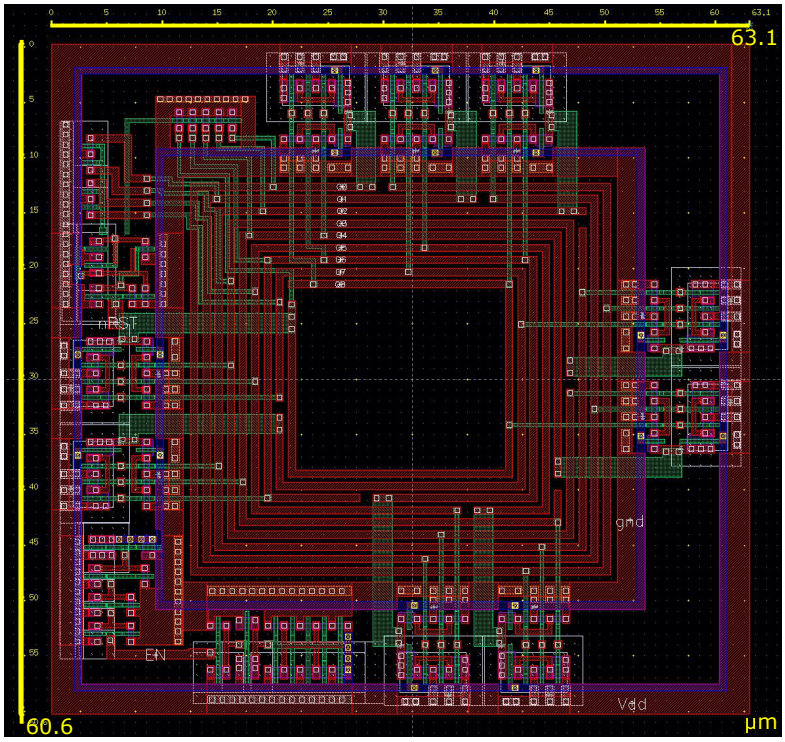
The connections of the delay stages are complex due to their placement at different points in the ring, which makes the design layout cumbersome. Because the MGRO is very different from the traditional ring oscillators, whose outputs are directly connected to the following delay stage, one of the main design challenges was matching the propagation delay between all the ring nodes.

The first attempted approach consisted of all the delay stages being arranged in a conventional ring manner, as depicted in Figure 2.12.

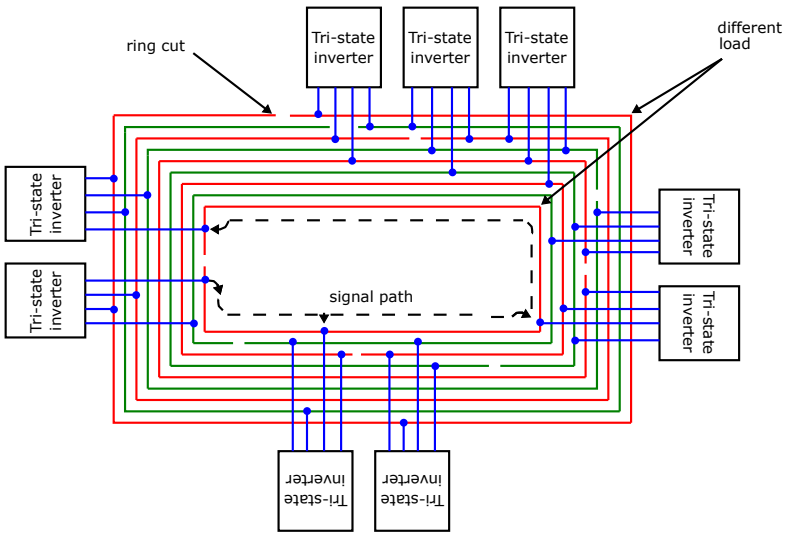
The main advantage is that all three inputs of the inverter arrive in the same order. However, this topology comes with a few disadvantages. Firstly, the area occupancy of such a topology is too large. Secondly, and perhaps most importantly, the delay stages do not present the same load. This results in an unbalanced layout which can introduce non-linearities in the circuit. Thus, the main focus of this layout was to carefully size the connections by using two different metal layers, of which only three in total were available for this particular tapeout. Layers metal 1 and metal 3 were chosen, with the benefit that the parasitic capacitance between the connections was reduced due to the larger distance between metal 1 and 3.

Unfortunately, the aforementioned layout design exhibited significant disadvantages, and a different layout design was implemented. Ultimately, the approach used in Blumino consists of the delay stages being placed in a line and connected through paths of almost equal length and same width, as shown in Figure 2.13.

In this way, an almost constant load is kept for each stage of the delay ring. Compared to the previous approach, this ring topology is much more compact and optimized for small area, the main difference being the balanced load between the delay stages. One of the disadvantages is the unequal propagation length between different groups of delay stages. However, this can be mitigated to some extent by reordering the inverters. The final TDC layout is presented in Figure 2.14.

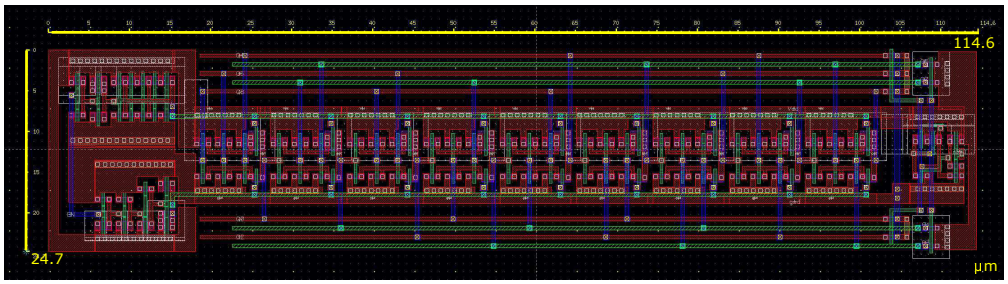


(a)

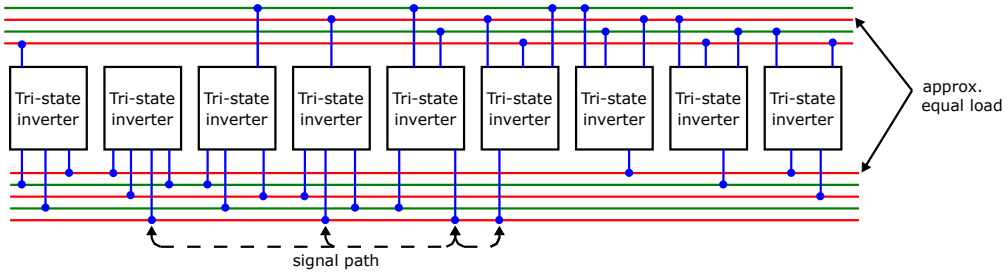


(b)

Figure 2.12: a) Layout of the ring topology with unbalanced load. b) Conceptual reprezentation of the layout in a) for a better visualization of the organization of the delay stages along the ring [115].



(a)



(b)

Figure 2.13: a) Layout of the ring topology with balanced load. b) Conceptual reprezentation of the layout in a) for a better vizualisation of the organization of the delay stages along the ring [115].

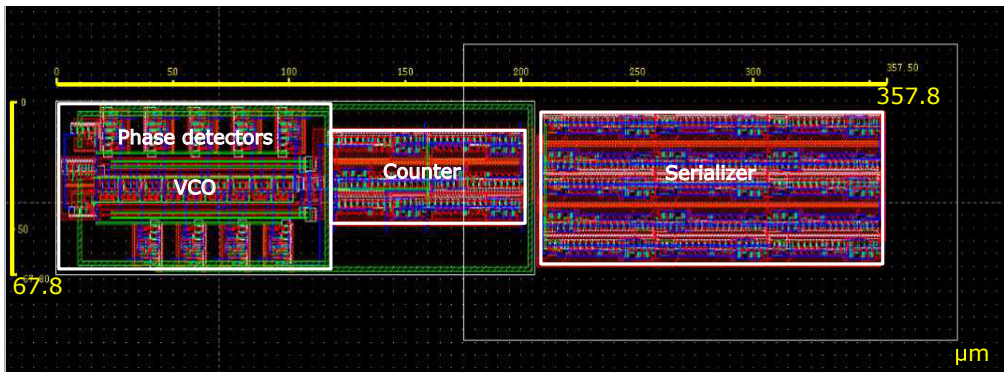


Figure 2.14: Final TDC layout without decoupling capacitors.

2.4 Characterization

2.4.1 Testing platform

The chip was encapsulated in a ceramic pin grid array (CPGA) package with a quartz window. This approach was chosen due to the convenience of changing the chips on the printed circuit board (PCB) very easily without the need of discarding PCBs in case one of the chips is damaged. A custom PCB was designed to provide the power and bias voltages required to operate the chip, as well as to serve as an interface between the photodetector and field-programmable gate array (FPGA) as presented in Figure 2.15.

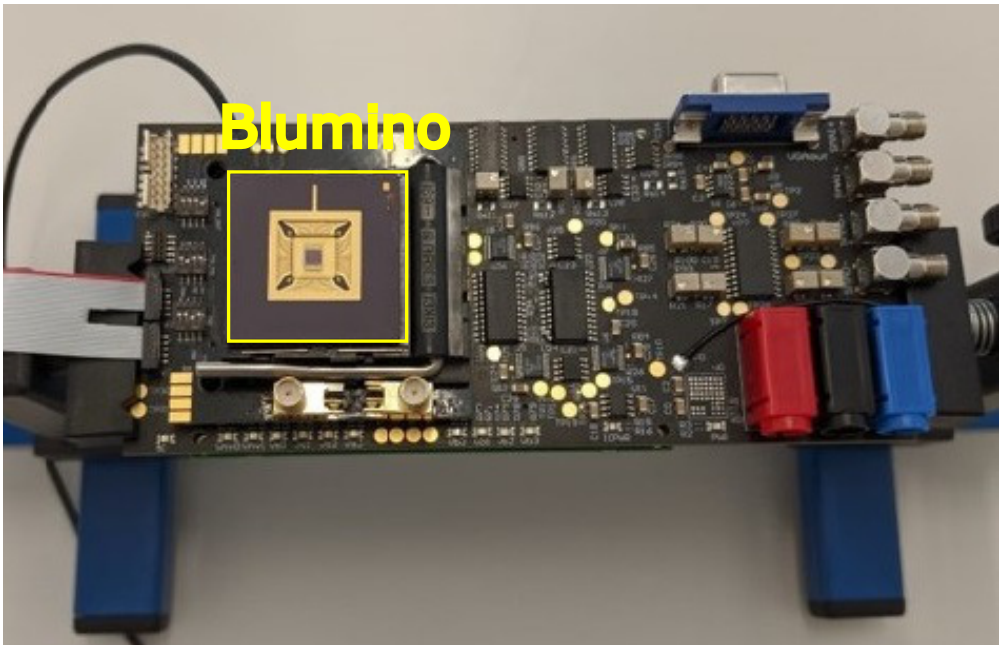


Figure 2.15: Blumino testing platform, composed of a motherboard, which serves as a control board for the sensor, along with a small daughterboard where the sensor is placed.

The custom PCB connects to an Opal Kelly XEM7360 Kintex-7 FPGA board. The chip input-output pads are directly connected to the FPGA along with the control signals for the power supplies. The PCB makes use of digital potentiometers which allow the possibility of sweeping bias voltages for both TDC and comparator. In addition, the power supply voltages can be digitally adjusted. This is needed especially for the TDC, whose oscillation frequency is voltage-controlled, as well as for the comparator, whose threshold voltage is controlled externally. During electrical testing, an additional START signal is generated with the same frequency but with an adjustable phase with

respect to the STOP signal; this allows different impulse widths to be fed to the TDC. The FLAG is issued every time the START arrives at the TDC and it is read together with the output data. The FLAG was used in post-processing to eliminate invalid TDC readings during optical test, where there is a possibility of not receiving the START signal (no photon detection). At the end of the measurement cycle, the data is accumulated on the FPGA in a large 65535 words first in-first out (FIFO) register and transferred to the PC vias USB 3.0. The data is then further analysed in Matlab.

2.4.2 Electrical characterization

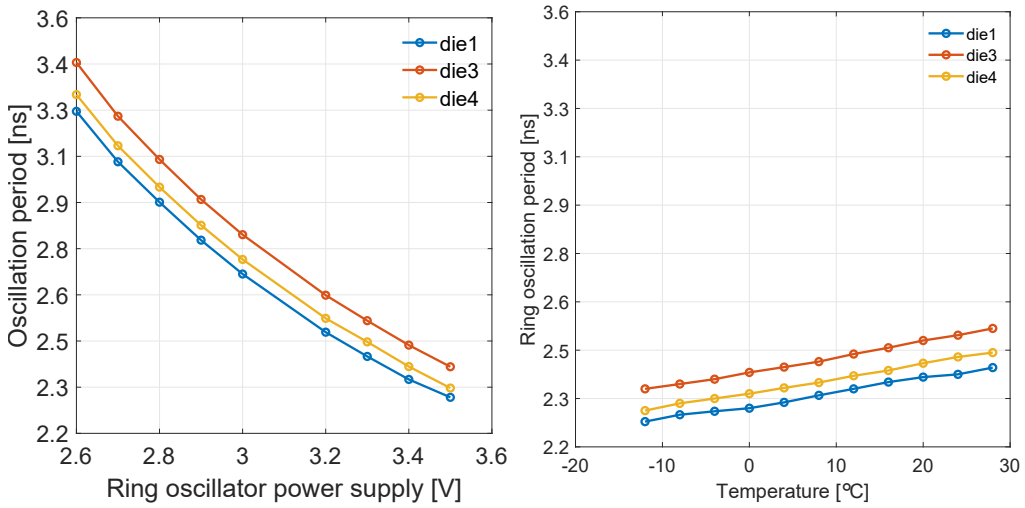
The main building blocks of the sensor were first characterized individually. An internal multiplexer allows the FPGA to communicate directly with the TDC and bypasses the analog front-end, allowing for the TDC to be characterized separately. In order to measure the VCO's oscillation frequency, the TDC is set in a free-running mode. The dependency of the oscillation on the power supply variation is measured by sweeping the VCO's supply voltage. The TDC's LSB also changes with power supply. The oscillation frequency (T_{osc}) is read out through the sixth counter bit while manually changing the power supply of the ring oscillator. T_{osc} decreases with an increase in power supply at an average rate of -1.24 ns/V, improving the TDC's least significant bit. The test was performed on three different dies, all of them presenting the same behavior as depicted in Figure 2.16a.

The oscillation period's temperature dependence was measured with a temperature chamber. The temperature was varied in steps of 4 °C from -12 °C to 28 °C. The oscillation period thermal drift was determined to be 0.3125 ns/°C as depicted in Figure 2.16b by using the following equation:

$$thermal\_drift = \frac{T_{warm} - T_{ref}}{temp_{warm} - temp_{ref}} \tag{2.2}$$

where T_{warm} is the oscillation period at a warm temperature, T_{ref} is a reference oscillation period which can be any random chosen point on the diagram, and $temp_{warm}$ and $temp_{ref}$ are their respective temperatures.

According to the measurement results the temperature variations can be compensated for by changing the power supply of the ring oscillator within a narrow range of around 300 mV.



(a) (b)

Figure 2.16: a) VCO oscillation frequency dependence on power supply. Measurement performed at room temperature. b) VCO oscillation frequency dependence on temperature @3.3 V supply voltage.

2.4.3 Optical characterization

The transfer function of the TDC was determined through measurements by generating START and STOP signals with the FPGA. The rising edge of the START signal triggers the TDC, while the rising edge of the STOP signal stops it. Consequently, the time interval between the START and STOP was varied over a range of 80 ns, with 16384 iterations per point which were determined through the FPGA code. The transfer function of the TDC is shown in Figure 2.17 and it indicates an average TDC LSB of 128 ps. Compared to post-layout simulation of the TDC where the LSB value was 65 ps, measurements indicate a LSB almost twice as large as the simulated one. This difference is almost certainly due to the presence of inaccuracies in the transistor models, because the LSB variation is too large in order to be attributed solely to the layout mismatches and parasitic capacitances. The LSBs of different TDCs in multiple dies were measured and all the results indicate the same behavior.

The functionality of the entire A-SiPM-comparator-TDC chain was determined through single-shot precision measurements at multi-photon level by using a 375 nm PiL037-FC laser for an A-SiPM excess bias voltage of 2.5 V. The laser was synchronized with the FPGA and the results of three different delays between START (laser trigger) and STOP signal (from FPGA) are depicted in Figure 2.18.

Due to the comparator's minimum threshold which, by design, corresponds to 114

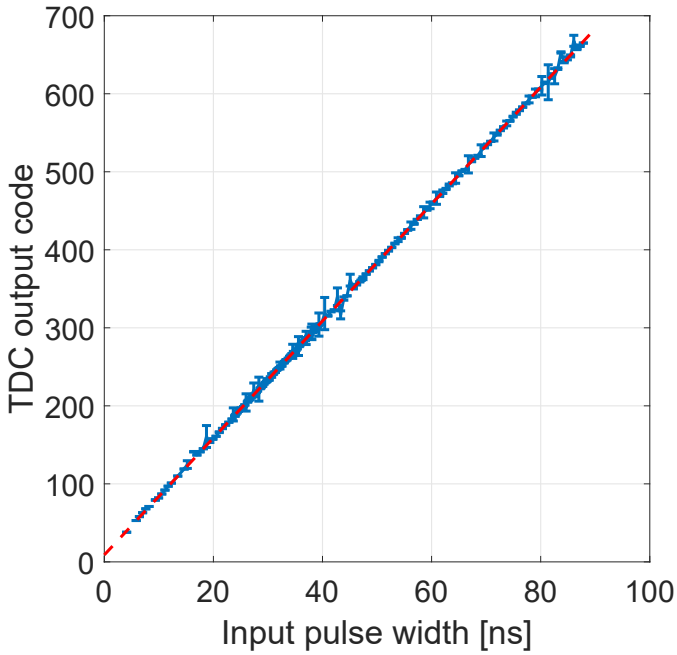


Figure 2.17: TDC transfer function determined through electrical characterization.

fired cells out of 4774 A-SiPM microcells in total, the measurements were performed in the multi-photon burst detection regime. As depicted in Figure 2.18 a small sigma variation is present in the TDC output code.

The DNL of the TDC was measured using a code density test where the photodetector is illuminated with white light, thus the photons are randomly distributed in time. When reading out all the timestamps from the TDC over a long period of time, each code of the TDC should be present in the histogram. In the ideal case scenario, the resulting histogram is uniform, and any deviations from this behavior are caused by the non-linearities of the TDC. The DNL can thus be determined by the difference between a specific bin in the histogram and the expected uniform value. The DNL and INL results are outlined in Figure 2.19 and present a DNL and INL of -1/+5 LSB and -2.4/+0.9 LSB respectively, after compensation.

The INL was compensated for by using a lookup table (LUT) created from a large dataset where

$$D_{compensated} = D_{measured} - INL_{D_{measured}}. \tag{2.3}$$

$D_{measured}$ is the measured code, $INL_{D_{measured}}$ is the INL of the measured code and $D_{compensated}$ is the compensated value of the respective code.

This method is acceptable for a limited range of temperature changes, however, if

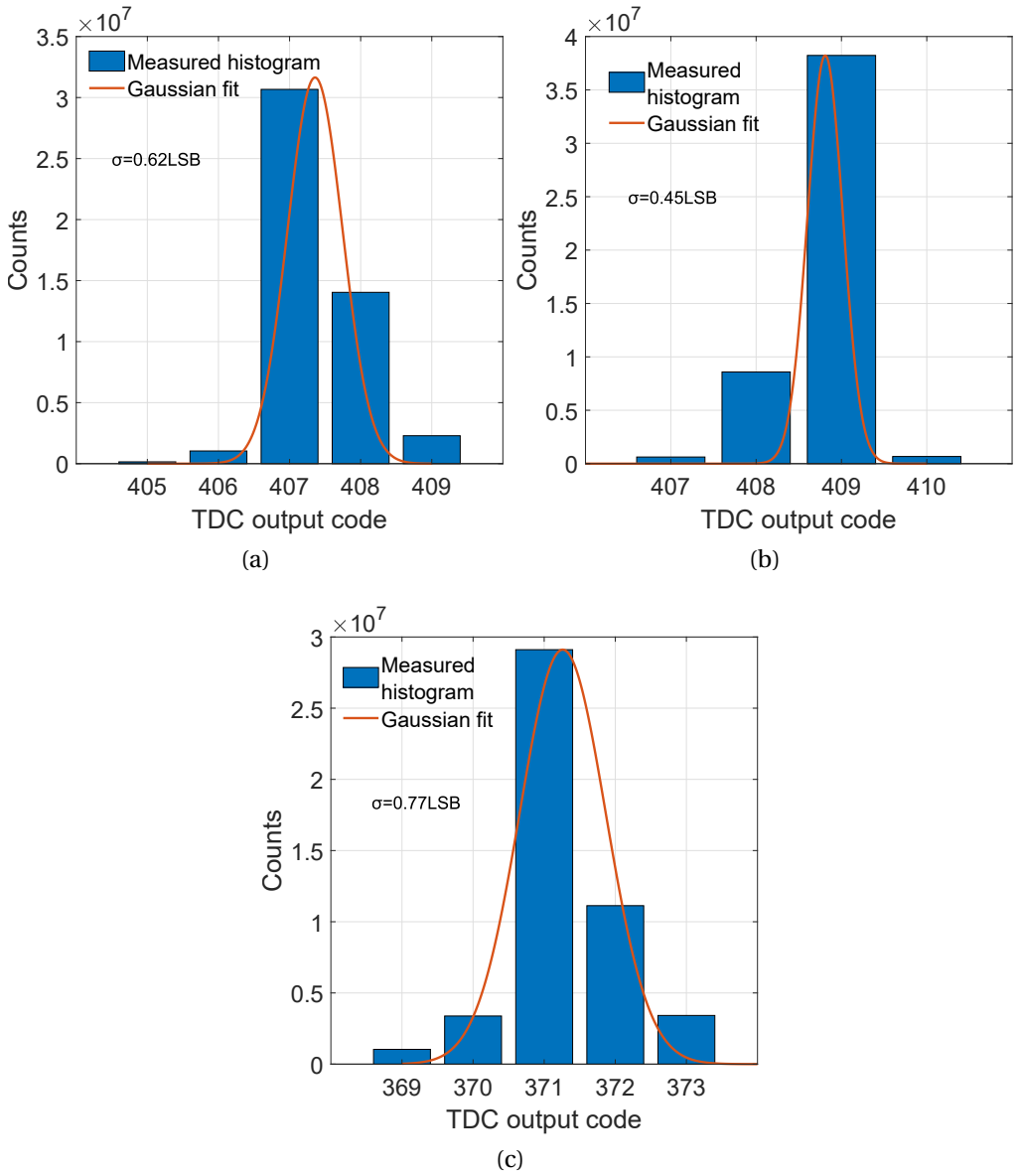


Figure 2.18: Full system single-shot precision of different TDC output codes. a) TDC output code 407 which corresponds to 52.096 ns measuring time interval. b) TDC output code 409 which corresponds to 52.352 ns measuring time interval. c) TDC output code 371 which corresponds to 47.488 ns measuring time interval. Measurements performed with a 375 nm picosecond laser.

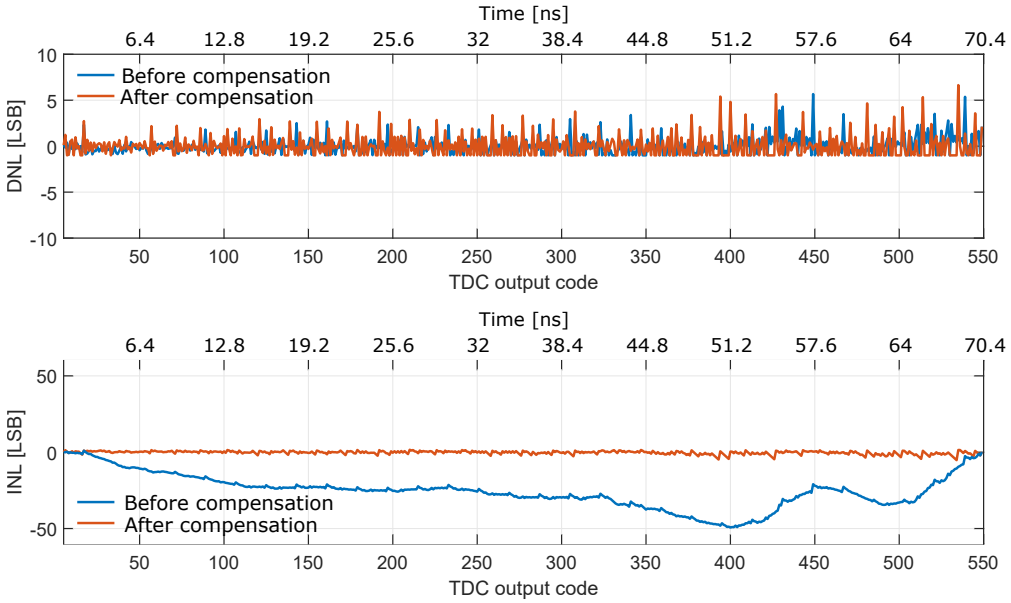


Figure 2.19: TDC's non-linearities - DNL and INL. Before compensation (blue) and after compensation (red).

a wide range of temperatures is required, other compensation methods need to be implemented [116]. This has been demonstrated by measuring at two different temperatures, 26 °C and 16 °C, where the VCO power supply was set to 3.3 V and 3.25 V, respectively, to compensate for the change in the oscillation frequency. By using the same LUT, the compensated INL was noticeably degraded. The DNL was however almost the same.

The PDP was measured on Pandion, a 400 × 100 SPAD sensor which uses the same SPAD device as in Blumino [117]. Both sensors, Pandion and Blumino were fabricated in the same multi-project wafer (MPW), therefore, fabrication variations should not significantly affect the PDP results , shown in Figure 2.20.

Compared to off-the-shelf standalone SiPMs of the same kind as the ones integrated on Blumino and developed by On Semiconductor, the PDP shows a slightly decrease after integration.

Another possible degradation due to the integration of a custom A-SiPM process with standard electronics process on the same chip was investigated in terms of DCR as well. The DCR was measured for six different dies at different excess bias voltages by accumulating hundreds of frames with a 40 GS/s LeCroy oscilloscope that were further analysed in Matlab. The DCR was measured using the standard A-SiPM terminal connected to a wideband (0.1 to 1000 MHz) amplifier from Mini-Circuits with an

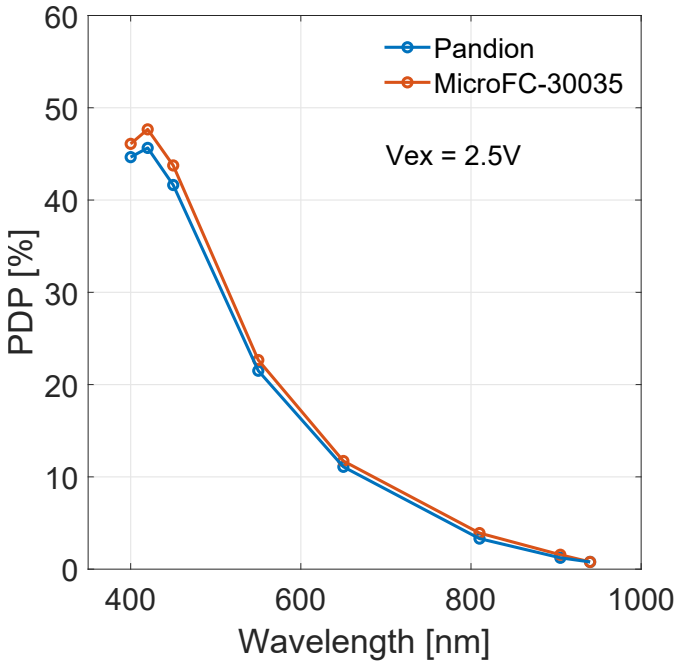


Figure 2.20: Pandion PDP comparison with off-the-shelf MicroFC-30035 SiPM (this series comprises the same SPAD size as the one implemented in Pandion) from On-semi, implemented in the same CMOS technology node.

amplification of 10. By plotting the DCR as a function of the oscilloscope threshold voltage, a distinct plot with two plateaus as seen in Figure 2.21a is obtained. The high plateau corresponds to the 0.5 photoelectron (phe) level and contains pulses originating from both the DCR and crosstalk, whereas the lower plateau contains pulses that correspond to 1.5 phe level. These originate as a result of crosstalk and have higher amplitude than the DCR pulses. The DCR was calculated by taking into account the pulses at the 0.5 phe level and the measurement results are depicted in Figure 2.21b. Figure 2.21c presents the crosstalk probability defined as the ratio between the count rate at the 1.5 phe and 0.5 phe levels [118].

2.4.4 CTR measurements performed with standard terminal

The CTR value is critical to evaluate the timing performance of a ToF-PET detector module. A ^{22}Na source which produces back-to-back 511 keV gamma photons, was placed between two Blumino sensors in coincidence. The energy of the two gamma photons is absorbed in the scintillators and is converted into visible photons which are then detected by the A-SiPM. Two 2.5 × 2.5 × 20 mm ^{3} LYSO crystals were wrapped with Teflon tape about 0.5 mm thick and glued to the A-SiPM using optical grease. The measurement setup is presented in Figure 2.22.

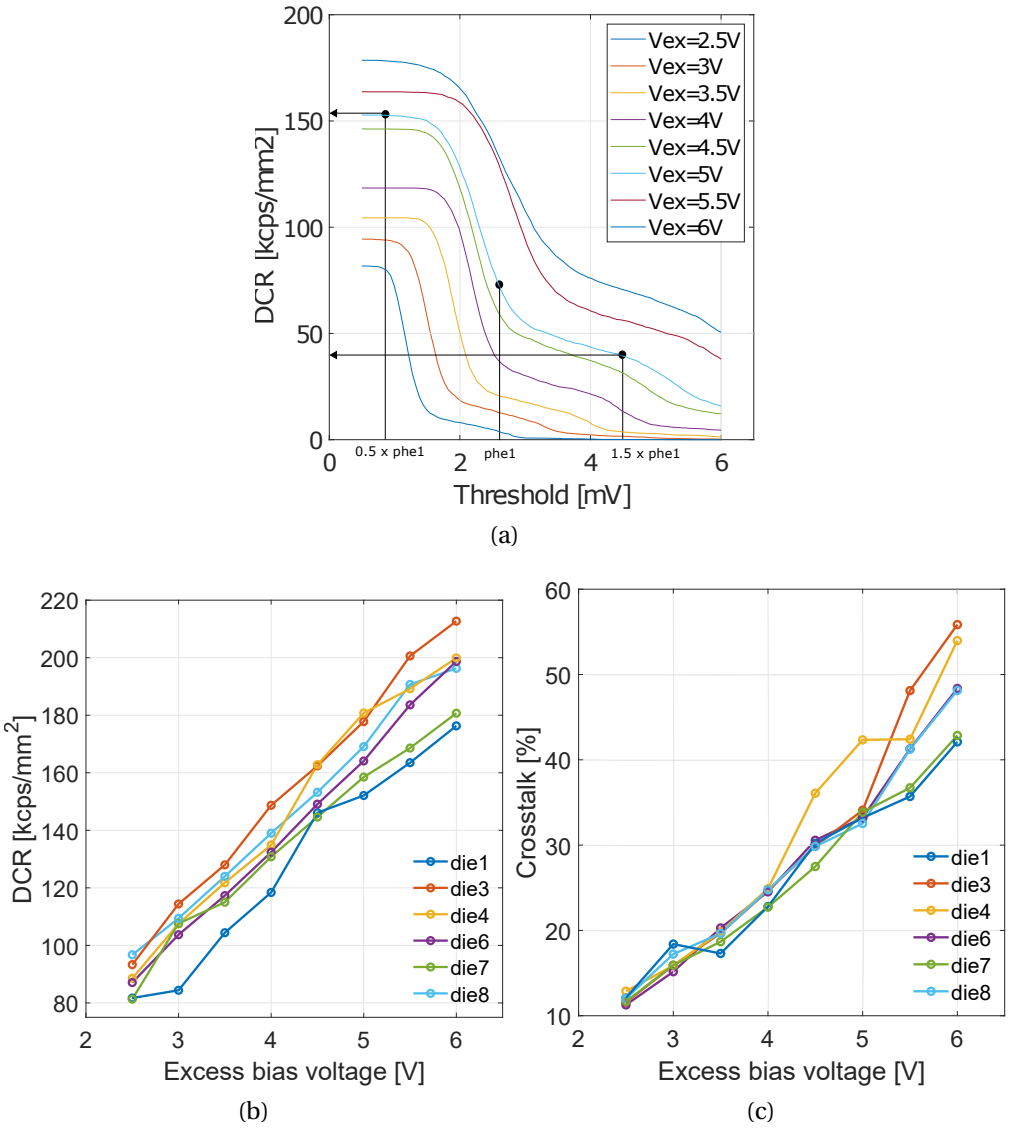


Figure 2.21: a) DCR as a function of oscilloscope threshold voltage. The first photoelectron level, phe1, is taken as the trigger level at the middle of the drop from one step to the next. The count rate at the 0.5 × phe1 level is considered as DCR and the rate at 1.5 × phe1 divided by the DCR value is the crosstalk probability. b) A-SiPM dark count rate versus excess bias voltage. Measurements performed at room temperature using the standard terminal. c) A-SiPM crosstalk versus excess bias voltage. Measurements performed at room temperature using the standard terminal.

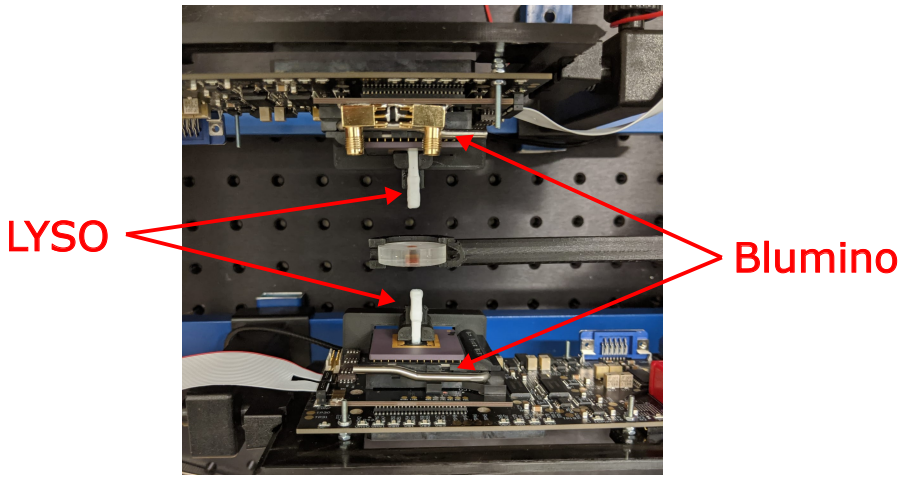


Figure 2.22: CTR measurement setup using the A-SiPM standard output. Two measurement platforms for the Blumino sensor are placed in coincidence with LYSO scintillators placed on top of the A-SiPMs and a ^{22}Na placed in the middle.

As a first step, the energy resolution was determined by using the charge integration of the A-SiPM's standard output, without any additional amplification, at 3 V excess bias. The energy resolution is defined as the ratio of the FWHM of the energy peak and the energy value corresponding to the energy peak maximum [119]. The energy spectrum was analysed and calibrated using Matlab and it shows an energy resolution of 17.1% as depicted in Figure 2.23.

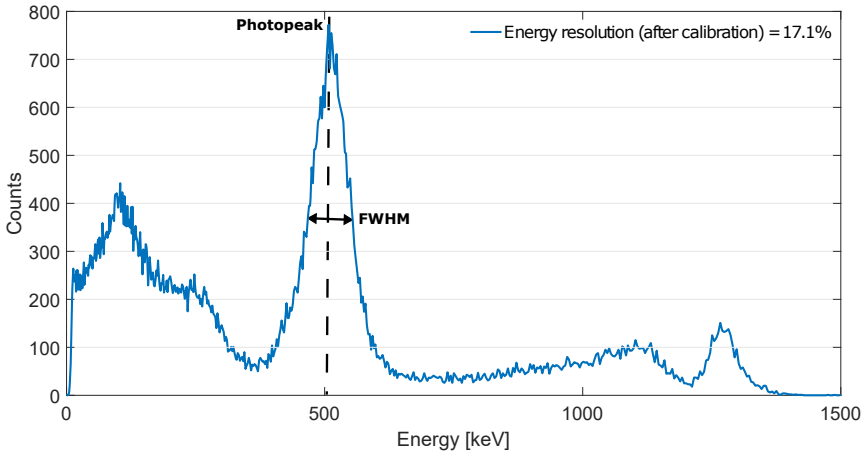


Figure 2.23: ^{22}Na spectrum measured with Blumino's standard output.

Two Blumino sensors were then operated in coincidence and 100.000 frames were accumulated using a 40 GS/s LeCroy oscilloscope triggered by one of the A-SiPMs. Only the frames where a pulse was present on both channels were kept for analysis. All the waveforms were post-processed in Matlab. The frames containing pulses that

did not correspond to the 511 keV photopeak (using an energy window of 408 keV to 613 keV) were discarded. The remaining data was analysed for different voltage thresholds by extracting the absolute timestamps of the impulse for both channels and computing the difference between them. The FWHM of the resulting distribution for a specific threshold after the error introduced by the timewalk was accounted for represents the CTR and is depicted in Figure 2.24.

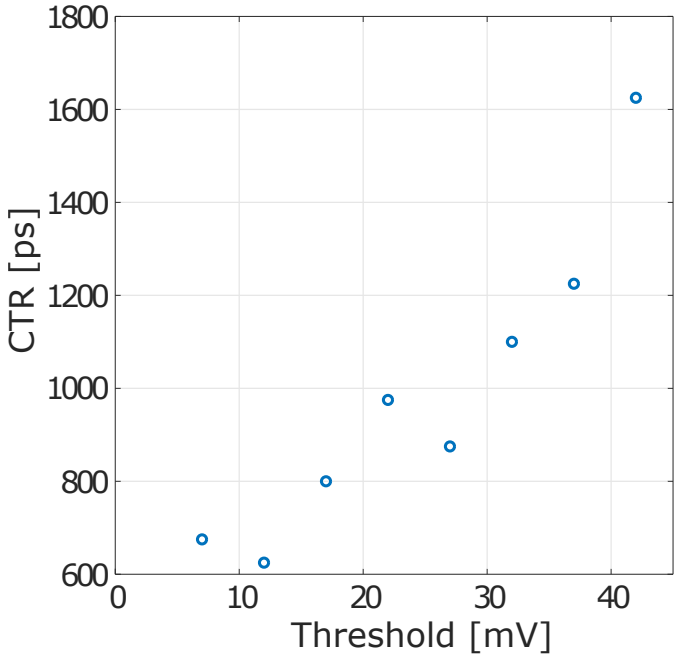


Figure 2.24: CTR measured with A-SiPM's standard output @3 V excess bias using 2.5 × 2.5 × 20 mm ^{3} LYSO crystals wrapped with Teflon. Timewalk correction was applied.

The timewalk error is introduced when the signal time is measured by using a constant threshold. As a consequence, the measured times between two different events with different energies that occur at the same true time are different because the slope of the pulse is not the same as depicted in Figure 2.25.

Timewalk correction is required in order to improve the timing resolution [120], [121]. The timewalk correction was implemented as follows:

- For each coincidence event, $dt = t_2 - t_1$ was measured, where t_1 and t_2 are the arrival times of the first detected photons in the two detectors placed in coincidence.
- Their respective charges q_1 and q_2 were calculated.
- The dt vs. q_1 distribution was plotted and a 5^{th} degree polynomial function $coeff(q_1)$ was fitted as in Figure 2.26.

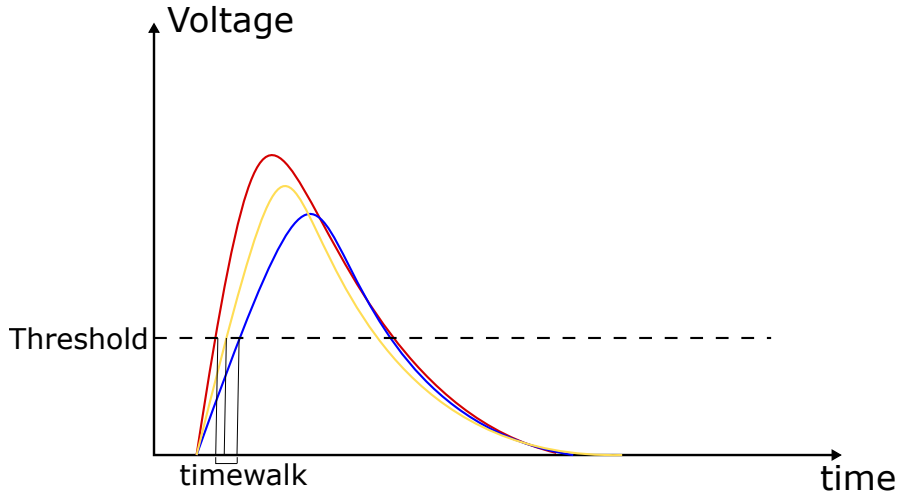


Figure 2.25: Conceptual representation of the timewalk. The threshold value is reached at different times depending on the pulse amplitude, even if all three pulses start at the same moment in time.

- A new variable was defined as: $dt' = dt - coeff(q1)$. The dt' vs. q_2 distribution was plotted as depicted in Figure 2.27 and another polynomial function $coeff(q_2)$ was fitted to it.
- The corrected timestamp difference is given by $dt'' = t_2 - t_1 - coeff(q_1) - coeff(q_2)$.

2.4.5 New CTR measurement platform

A dedicated small sensor board which implements time-over-threshold on the ST for the energy measurement and reads out the timing information from the TDC was designed as illustrated in Figure 2.28.

In order to perform CTR measurements, timing and energy information needs to be acquired at the same time. Therefore, a synchronisation between the ST and FT is necessary. A conceptual block diagram of the small sensor readout board is depicted in Figure 2.29.

The time-over-threshold technique is implemented on the ST by using discrete components. The A-SiPM's ST signal is amplified by a transimpedance amplifier featuring a bandwidth of 100 MHz. The output of the transimpedance amplifier provides a negative pulse, due to its transfer function. Then, the amplifier's output was inverted so that a positive threshold could be provide by the microcontroller (μC). The com-

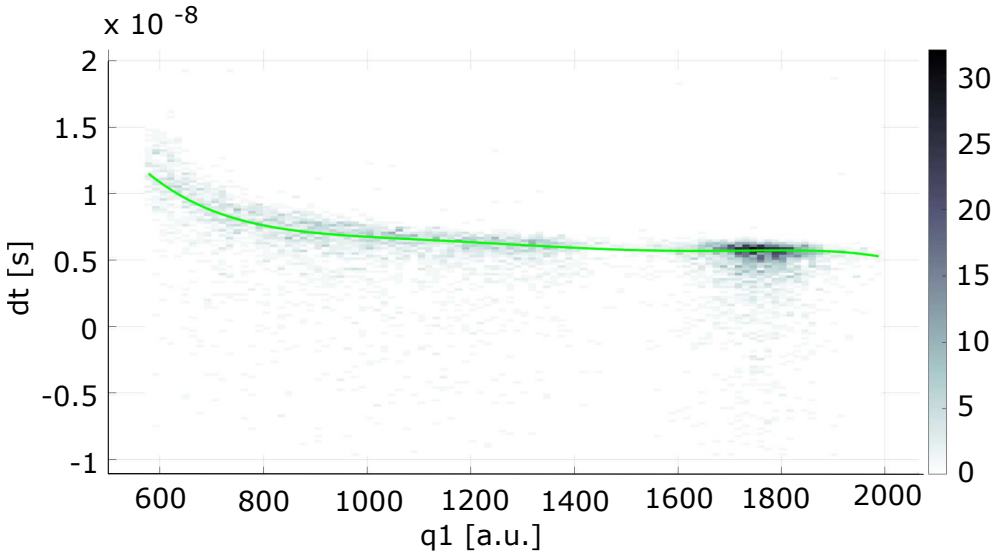


Figure 2.26: Two dimensional histogram of the measured time difference $dt = t_2 - t_1$ of the two timestamps registered by two detectors placed in coincidence as a function of the energy q_1 of the pulse from the first detector. The green line is a 5^{th} degree polynomial fit. Measurement performed at 3 V excess bias.

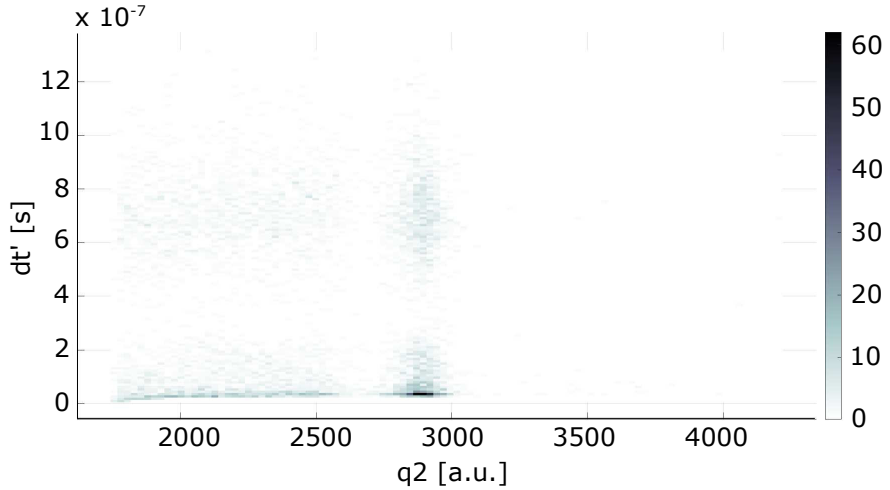


Figure 2.27: Two dimensional histogram of the measured time difference minus the polynomial fit function for the second detector (dt') as a function of the energy q_2 of the pulse from the second detector. Measurement performed at 3 V excess bias.

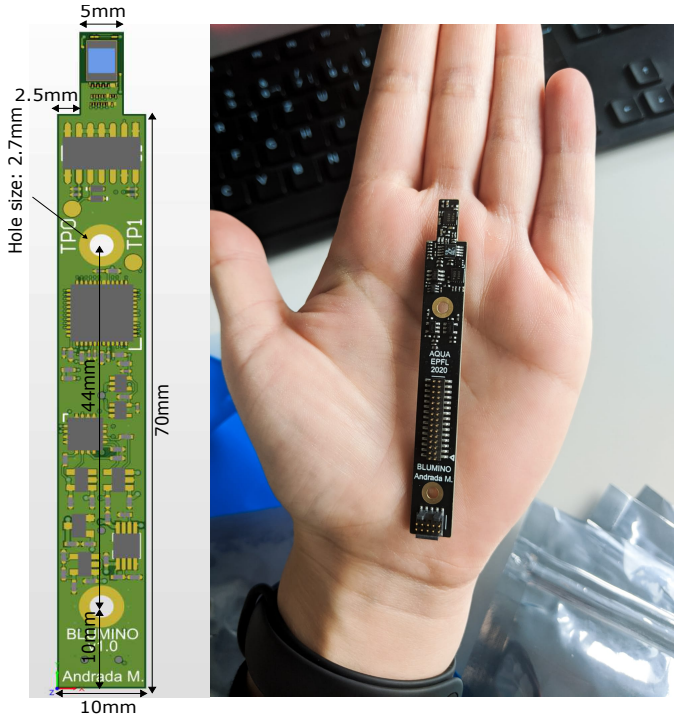


Figure 2.28: Blumino small readout sensor board for CTR measurements with the integrated TDC on the fast terminal. The time-over-threshold technique is implemented on the standard terminal. The board contains pre-processing on the microcontroller.

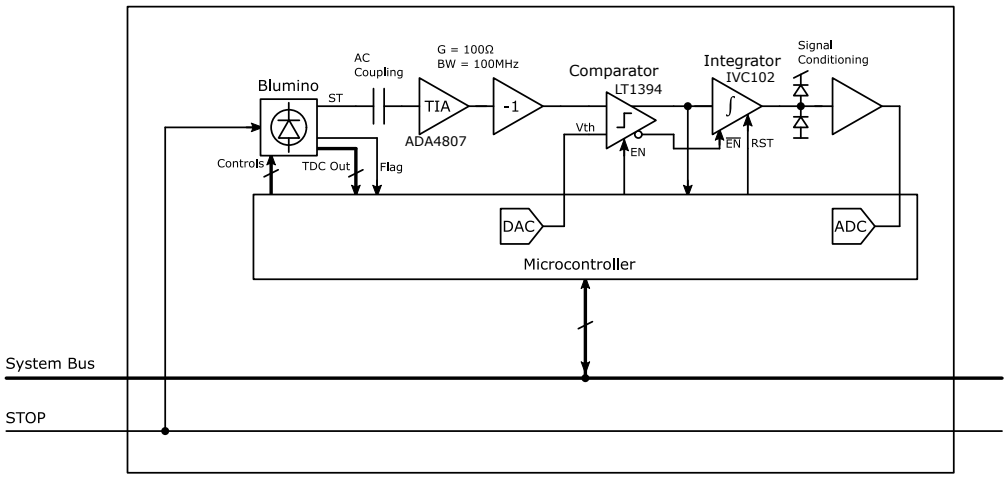


Figure 2.29: Blumino small readout sensor board conceptual diagram. The time-over-threshold technique is implemented with discrete components on the ST. The integrated TDC timing information is directly sent to the μC. The μC's internal DAC is used to control the comparator's threshold while the ADC is used to read out the integrator's value.

parator's threshold voltage is set by the μC's DAC. The comparator has a latch enable (LE) which is activated after the first photon arrival which is kept low until the first photon arrives and then set to high so that the comparator is latched. Therefore, the on-chip integrator is stopped from measuring the following incoming events. A μC was used in order to facilitate the data handling among multiple boards and due to its integrated components (data converters). The μC's DAC and ADC present a 12 bit resolution, with an LSB of 0.8 mV and a reference voltage of 3.3 V. The ADC's input conditioning limits the voltage range between 0 and 3.3 V in order to avoid damaging the input due to the integrator's -16 V supply voltage. At the end of an acquisition cycle, the energy and timing information is readout by the μC. The small sensor readout board is coupled to an interface board, as depicted in Figure 2.30, which is further coupled with a motherboard and then a FPGA in order to read out the information coming from multiple sensors.

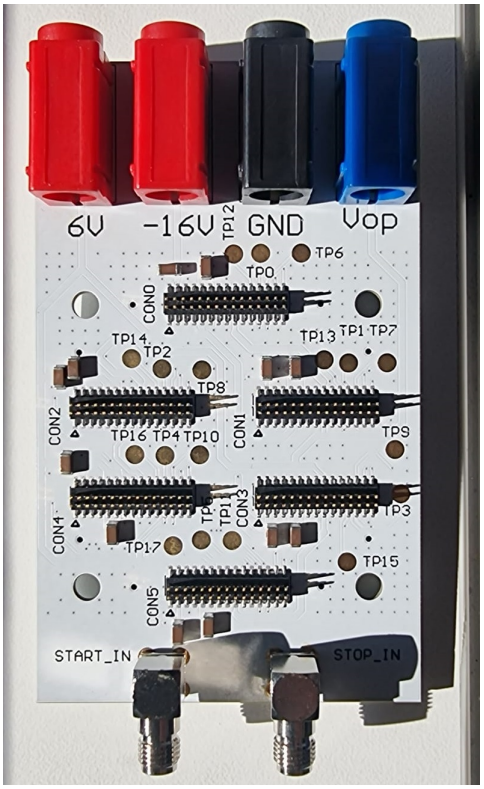


Figure 2.30: Blumino interface readout board. It accomodates 6 ribbon cables each of them capable of handling tens of mini sensor boards to speed up the readout process.

The interface board provides three different supplies for the mini PCBs: 6 V (mother board), -16 V (integrator) and SPAD Vop. The width of the traces has been chosen in such a way that it can handle all the needed currents and the power traces have also been enlarged. A special tool included with Altium Designer that can equalize the

lengths of the START_IN and STOP_IN signal was used. The board was designed on 4 layers only and in order to speed up the readout process and minimize the current that passes through a single ribbon cable, 6 ribbon cables have been used instead of one. The power consumption of a single small sensor board (considering the interface board as well) is 277 mW.

The sensor readout board was designed in a small configuration in order to improve the timing performance by providing amplification and readout right next to the ST. In addition, the board was designed as well for a future small PET ring prototype based on the Blumino sensor as presented in Figure 2.31.

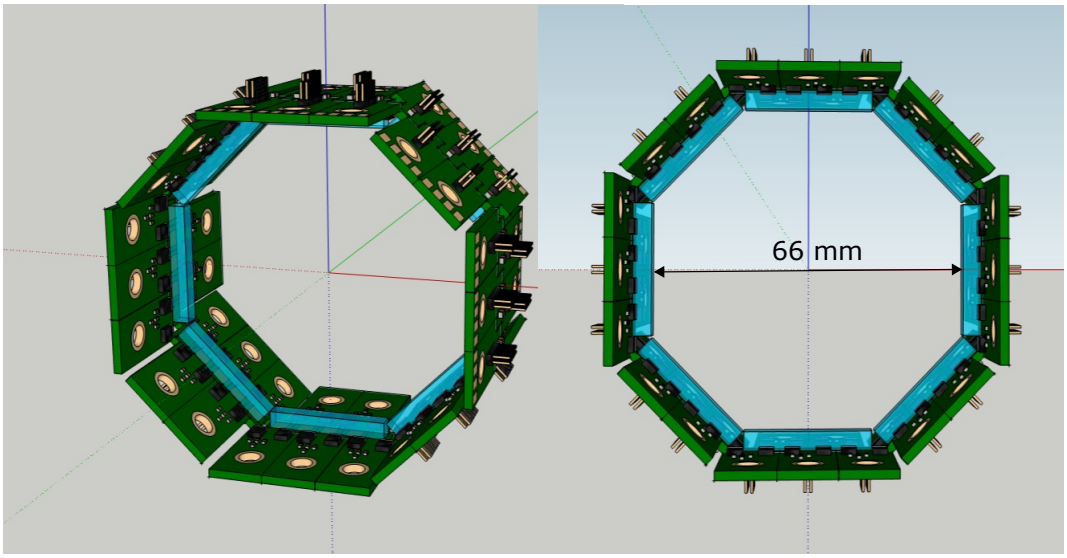


Figure 2.31: Future small PET ring prototype based on Blumino sensors. All sensors are controlled by a single interface board. 6 independent channels are available, each one accomodating tens of small readout boards. The ring diameter is 66 mm and it is limited by the number of available Blumino sensors. The configuration presented above is based on $3 \times 3 \times 25$ mm ^{3} scintillators.

In this small PET ring future prototype, all the boards are controlled by a single interface board as mentioned above. The configuration presented in Figure 2.31 is based on $3 \times 3 \times 25$ mm ^{3} scintillators for a 66 mm diameter ring. The ring configuration is flexible while being limited by the available number of Blumino sensors.

A graphical user interface (GUI) has been designed in order to control all the supply voltages, thresholds and send various commands to the μC as depicted in Figure 2.32.

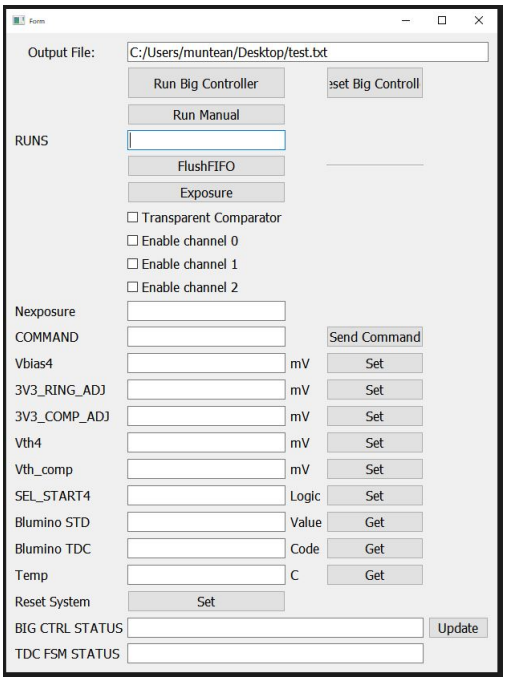


Figure 2.32: Small sensor board GUI interface. The threshold and voltages for all the components are digitally controlled from the GUI interface. The exposure window can be set.

CTR measurements

The initial important step is to investigate the CTR measurement performed with only two small sensor boards placed in coincidence as depicted in Figure 2.33.

Previously, the ST was connected directly to the oscilloscope through long cables and the threshold was set in the oscilloscope, therefore, the timing performance was deteriorated. Measurements were also performed with shorter cables in order to investigate their influence on the CTR value. In the end, no significant change was observed. An amplification and discrimination right next to the ST implemented on the small readout sensor board improved the timing performance. The measurements were performed using the 40 GS/s LeCroy oscilloscope by monitoring the output of the comparators from both boards. Two different crystal configurations: 2.5 × 2.5 × 3 mm ^{3} LYSO scintillator coated with $BaSO_4$ and 2.5 × 2.5 × 20 mm ^{3} LYSO scintillator wrapped with Teflon were analysed. The CTR measurement results are illustrated in Figure 2.34.

The results obtained with 2.5 × 2.5 × 20 mm ^{3} LYSO that correspond to Figure 2.34 b) can be compared with the previous CTR measurements performed in the same configuration on the ST but without the comparator and amplifier on PCB. As seen in

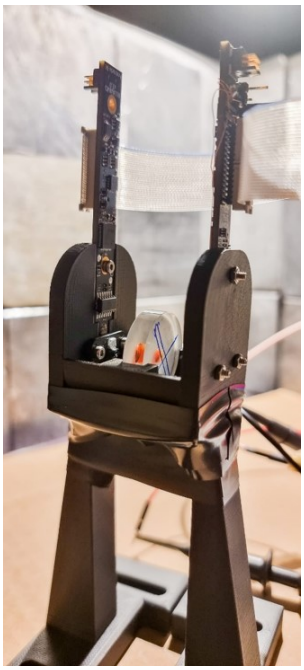


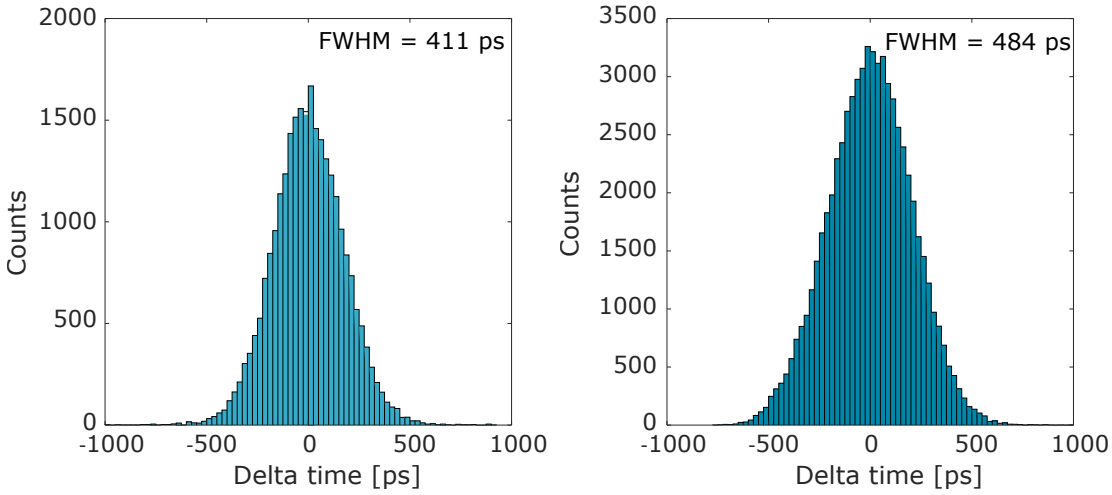
Figure 2.33: CTR measurement setup with Blumino small readout sensor boards. A ^{22}Na radioactive source is placed between the two detectors. The photodetectors are coupled with 2.5 × 2.5 × 3 mm ^{3} LYSO scintillators and placed at a distance of 32 mm.

Figure 2.24, a CTR of 630 ps was measured in the first case while the current CTR is 484 ps, presenting a 23.17% improvement.

The exposure-based readout of the architecture resulted in a low event rate when using only two detectors. As a result, the system was inefficient and CTR measurements with the FT could not be performed. A revision of the board is currently under development, to change the architecture to an event-driven acquisition type which will drastically increase the efficiency of the system.

2.5 Conclusions

This chapter discussed the design and characterization results of the first fully-integrated analog SiPM with on-chip discrimination and time conversion. A simple, yet fully functional and compact sensor was designed by combining a SPAD CMOS dedicated process with standard CMOS in 350 nm technology node. While such old technology nodes could be beneficial for the photodetector performance, more complexity needs to be added in order to develop faster electronics. The designed TDC proposes a simple architecture based on a multi-path gated ring oscillator capable of speeding



(a) (b)

Figure 2.34: CTR measurements performed on ST with the integrated amplifier and comparator on the small readout board. a) CTR results after timewalk correction obtained with 2.5 × 2.5 × 3 mm ^{3} LYSO scintillator coated with $BaSO_4$. b) CTR results after timewalk correction obtained with 2.5 × 2.5 × 20 mm ^{3} LYSO scintillator wrapped with Teflon. Measurements performed @ 3V excess bias.

up the oscillation frequency, hence, improving the LSB. However, there are multiple contributors to the overall timing resolution. Blumino aimed to accommodate for this by integrating the timing circuitry together with the A-SiPM. The sensor's performance was determined through optical, electrical and radiation measurements. A measuring platform suitable for tiling which makes use of both the integrated electronics and the ST has been implemented for performing CTR measurements and to further serve as a PET module in a future small PET ring demonstrator. CTR measurements performed with the new measuring platform on ST indicate an improvement of 23.17% by using amplification and comparison in close proximity on PCB. The TDC's LSB is 128 ps and the system exhibits an energy resolution of 17.1% when coupled to a 2.5 mm × 2.5 mm × 20 mm LYSO scintillator. Minor degradations due to the integration of custom CMOS SPAD process with standard CMOS were observed in the DCR with a measured value of 81.7 kcps / mm ^{2} at 2.5 V excess bias. Single-shot precision measurement results have a small standard deviation of 0.45 LSB in the best case. Further improvements such as the design of faster timing circuitry as well as the design of an event driven measuring platform in order to speed up the measurement time would improve the overall performance. A comparison between Blumino sensor and other SPAD-based sensors designed in the same technology node is presented in Table 2.1.

Compared to the designs presented in [122] and [123], Blumino's fill factor is sig-

nificantly higher due to the use of an A-SiPM. In addition, Blumino exhibits a larger PDP of 46% @425 nm for a lower excess bias of just 2.5 V. Compared to the other works, Blumino has a larger TDC LSB of 128 ps due to the lack of accurate transistor models during the design phase. Due to the nature of the MGRO TDC, it was expected to have higher nonlinearities with increases in power consumption for this particular design, as presented in the comparison table. The lower dynamic range, compared to [122] and [123], is not a limiting factor because Blumino's TDC has a feature that allows the unlimited extension of the dynamic range using off-chip components.

Table 2.1: Blumino comparison with sensors fabricated in the same CMOS technology node.

Parameters	Blumino	[122]	[123]
Technology [nm]	350 CMOS	350 HV CMOS	350 HV CMOS
SiPM type	A-SiPM	D-SiPM	D-SiPM
Microcell size [μm^2]	35× 35	30 × 50	30 × 50
Microcells #	4774	416	416
Fill factor [%]	75	57	57
PDP	46% @ 425 nm Vex = 2.5 V [a]	30% @ 425 nm Vex = 4 V	32.6% @ 420 nm Vex = 3.5 V
TDC type	MGRO	GRO coarse + fine	GRO coarse + fine
Total # TDCs	1	192	432
TDC LSB [ps]	128	51.8	48.5
Dynamic range [ns]	128	3390	6360
DNL, INL [LSB]	-1/+5, -2.4/+0.9	1.97, 2.39	±0.75, 4/-2.1
TDC area [mm^2]	0.0242	0.0136	NA
Power/TDC [mW]	< 9	1.65	1.65

[a] determined based on Pandion measurements [117].

3 Blueberry: A CMOS 3D-Stacked FSI Multi-Channel Digital SiPM

3.1 Motivation

Time-of-flight positron emission tomography is an imaging technique used in a wide range of medical applications. To achieve the best performance, a PET detector module should have a large sensitive area that can be coupled with existing scintillators, high spatial resolution, high timing resolution to measure the arrival times of the coincidence gamma photons, large energy resolution to discriminate the 511 keV events and reduced readout time to increase the count-rate capability [124], [125]. Although these requirements are demanding, they can be readily achieved with modern electronic circuits. The associated electronics can be implemented with either discrete components or integrated electronics; however, each modality presents its own advantages and disadvantages. In the case of using discrete components, these are usually easily available and modules of different configurations can be implemented (for example, the small measurement board of Blumino as presented in Chapter 2). In addition, modifications can be performed easily and debugging is much easier than in the case of integrated systems because each signal can be accessed. The use of discrete components can in general result in an increase of the overall power consumption. By using an integrated solution, the designs are more compact and additional features can be implemented. While in Chapter 2 we explored the full integration of an A-SiPM with a single integrated TDC, the goal of the Blueberry project was to implement a compact, fully integrated sensor that can accommodate multi-channel digital SiPMs with multi-timestamping capability through a fully integrated solution. The use of multiple timestamps is important for the order statistics which has been explored extensively over the years in [126]–[128].

Blueberry is a 3D-stacked multi-channel photodetector implemented in 180 nm standard CMOS technology. In order to implement the 3D-stacked FSI process, a

minimum size for the chip of 6 mm × 3.5 mm was imposed by the foundry. Therefore the final chip dimension is 7.5 mm × 4.2 mm. In order to decrease the design complexity of such a large detector, the design was divided into two identical parts. The top tier comprises a detector array, consisting of 64 × 64 front-side illuminated SPADs. The readout and processing electronics are located on the bottom tier, and the two layers are connected through TSVs [36], [129]. The 3D-stacked approach was chosen based on the need of using complex and diverse electronic circuits which process all the information from all the SPADs. Additionally, a 2D architecture would have been cumbersome to implement, significantly lowering the final fill factor of the SPAD array. The use of 3D integration facilitated the implementation of a large sensitive area which presents a fill factor of approximately 67%. In addition, a large area was available for the implementation of all the electronic blocks on the bottom tier. Finally, the TSV manufacturing process represents a challenge in itself. As inserting the TSVs can oftentime fail, Blueberry aims to serve as a prototype circuit for a fully integrated 3D-stacked FSI ToF-PET photodetector along testing the TSV process as well.

3.2 Core architecture

The Blueberry sensor comprises two independent 64 pixel × 64 pixel SPAD arrays, as shown in Figure 3.1.

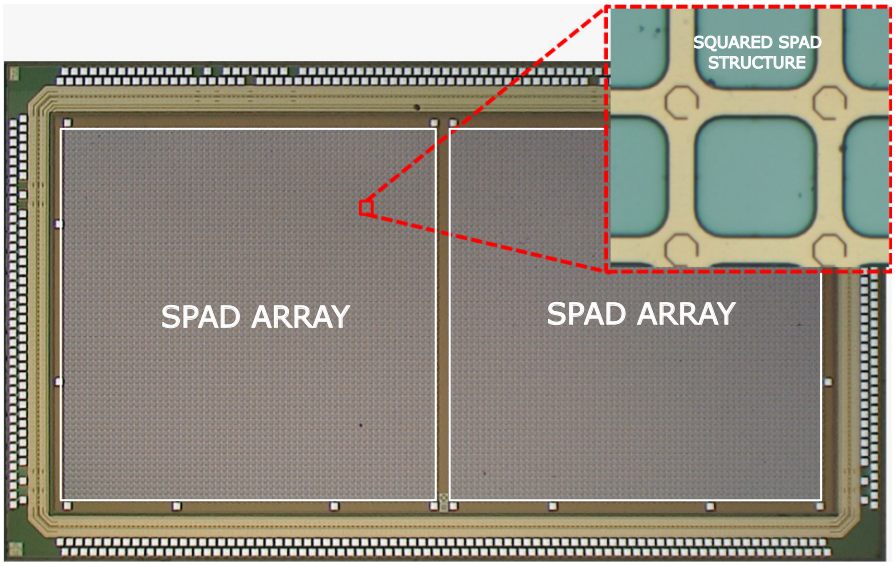


Figure 3.1: Blueberry micrograph. Two large independent SPAD arrays of 4096 SPADs each are present on the top tier. Chip size: 7.5 mm × 4.2 mm. Inset: squared SPAD structure with TSV landing site outlines in the corner [36].

Each SPAD is based on a p-i-n structure, has a square shape with round corners and a

pitch of 50 μm. Measurement results indicate a photon detection probability of 55% at 480 nm at 6 V excess bias. More information regarding the implementation and performance of the SPAD itself is presented in [36]. Each SPAD's anode is connected to the bottom tier through a TSV resulting in a total of 8192 TSV connections for both SPAD arrays.

As shown in Figure 3.2, the SPADs and their corresponding circuits are grouped into 64 clusters. Each cluster contains a TDC which timestamps the time of arrival of the first firing SPAD. As a result, Blueberry has a total of 64 TDCs.

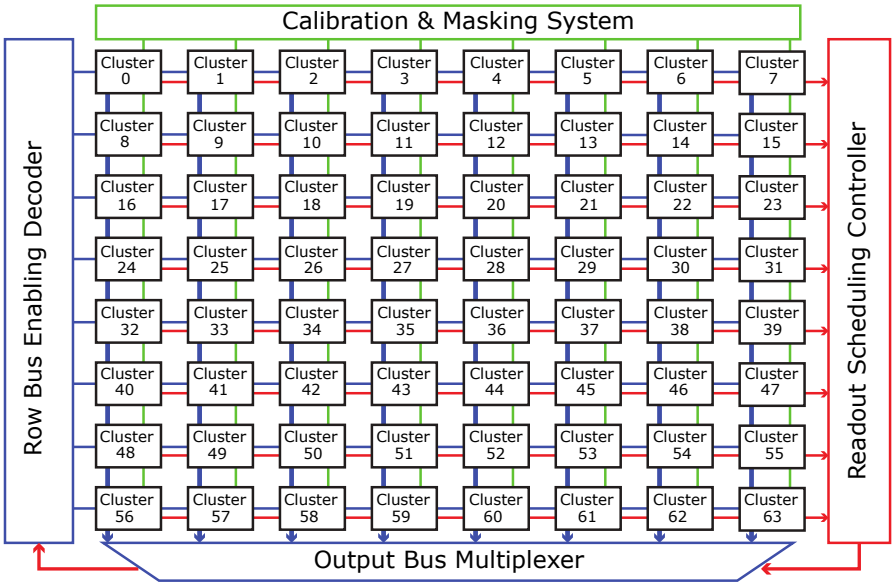


Figure 3.2: Core architecture. The core is divided in 64 clusters. The row bus enabling decoder is used to access the clusters' data. The calibration circuit is used to calibrate all 64 TDCs in the array. A masking system is used to mask noisy SPADs. The readout scheduling controller enables the event driven readout of the chip [36].

Data from the clusters is accessed using a row and a column addressing scheme. The entire system can be read out in two different modes: user or event driven readout. In the first case, the user can freely select which cluster to read while in the second case, the readout scheduling controller is used for reading out only the clusters with data. In addition, the sensor comprises a calibration and masking system which allows the calibration of each TDC and the masking of noisy pixels.

3.3 Cluster architecture

Each cluster in the array contains four main blocks: pixel circuits, photon counting systems, a TDC and a SPAD address circuit. The cluster comprises 8 × 8 pixels. Each

pixel is based on passive quenching and active recharge circuit with the output of each pixel connected through an OR tree to the photon counting system. The latter consists of ripple counters which count the number of pulses generated in a frame. In each cluster, four 4-bit counters are present in order to increase the counting capability. The outputs of all of the four counters in a cluster are summed up in an adder and the final result is a 6-bit word. The final result is then saved in a register and sampled by the global STOP signal. The output of the OR tree also triggers the TDC and a SPAD address tree based on a winner-take-all (WTA) approach is implemented to determine the address of the first firing SPAD in the cluster. The data for the timing, address and counting systems in each cluster is sampled by a global STOP signal and then saved in a memory buffer. A conceptual block diagram of the cluster is presented in Figure 3.3. More detailed information regarding the SPADs' design, pixel architecture, counting and readout systems is presented in [36].

3.4 Antiphased time-to-digital converter

The TDC is based on a reconfigurable antiphased topology and contains five main structures: a VCO, an asynchronous ripple counter, 36 phase detectors, 4 reconfigurable delay cells, and an output processing unit (OPU) as depicted in Figure 3.4. The motivation for choosing this architecture is to compare the performance of the multi-path gated ring oscillator implemented in 180 nm with the design from Chapter 2 implemented in 350 nm CMOS technology node. I also wanted to investigate the factor of two decrease in the raw resolution discrepancy between post-layout simulation and silicon. In this implementation, no such effect was observed. By using the antiphased structure, the resolution of the TDC was increased by a factor 4 without changing the original design. As a consequence, the design occupies a larger area and the use of multiple sets of delay cells and phase detectors results in higher non-linearities.

The most significant bits of the TDC are given by the asynchronous ripple counter which counts the number of oscillations in the ring. The least significant bits are determined by the non-transparent phase detectors which capture the state of each phase. The TDC's resolution can be tuned by using the reconfigurable delay cell which will be described later on in this chapter.

The VCO is based on a multi-path gated topology. Each delay stage of the MGRO is composed out of symmetric tri-state inverters with three inputs as illustrated in Figure 2.9. The inputs of each delay stage are connected to different places along the ring in order to help increase the oscillation frequency as depicted in Figure 3.5.

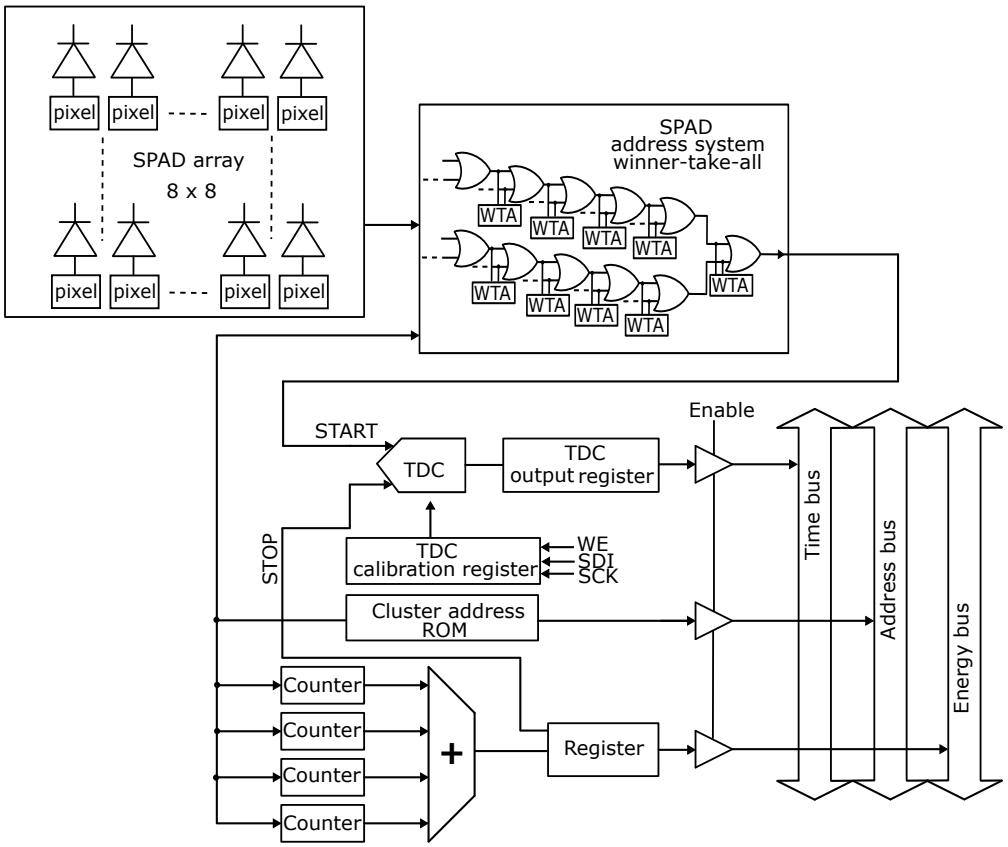


Figure 3.3: Cluster block diagram. Each cluster is an array of 8 × 8 SPADs with their corresponding pixel circuits. The output from all of the pixels is processed by a SPAD address system based on an OR-tree. The SPAD address tree is based on a winner-take-all implementation. Four 4-bit counters are implemented in each cluster in order to count the number of pulses. The output of the SPAD address systems triggers the TDC. The time, energy and address information from each cluster is written on a bus.

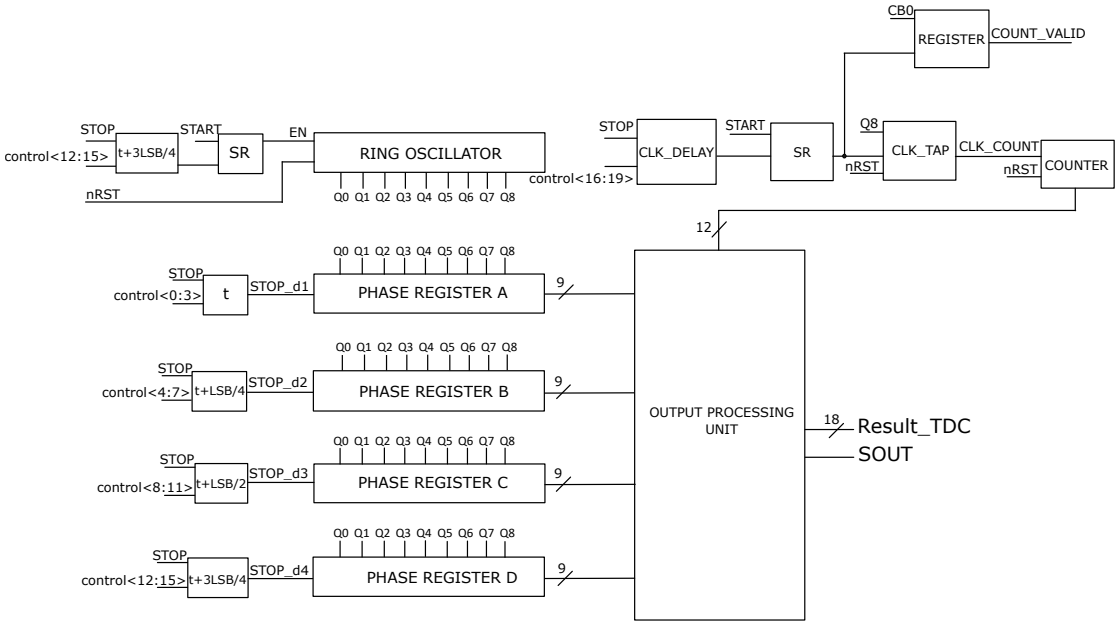


Figure 3.4: Time to digital converter block diagram. The 9 phases of the VCO are captured by four different sets of phase registers. A delay element is present that delays the phases in order to adjust the TDC's resolution. A 12 b counter counts the oscillation periods. The output processing unit (OPU) provides the final TDC result. OPU can provide a parallel and serial TDC result.

In this way, each delay stage can start transitioning ahead of time and therefore, the delay per stage is significantly reduced. There are 9 delays stages in the ring which result in 9 phases with 18 possible states. An oscillation period is completed only after the signal passes through all the nine inverters twice. The use of three inputs limits the minimum number of delay stages to nine. Depending on the position of the inverters in the ring, the connection lengths can vary drastically. All possible permutations (362880 in total) were analyzed in Matlab and the configurations with most balanced connections between all 9 inverters in the ring were determined. In the end, out of 362880 possible combinations, only 174 have balanced lengths. The chosen combination is shown in Figure 3.6. The new layout topology presents seven equal phase connections with the smallest Q_8 phase connection (the phase connected to the counter).

The MGRO oscillates while an enable (EN) signal is high and enters a high-impedance state when the EN signal is low. The EN signal is formed using an SR latch from a START signal and one derived from the STOP as seen in Figure 3.4. While the ring is in high-impedance, nine reset transistors can be used to pull the nodes to a predefined value which represents the reset state.

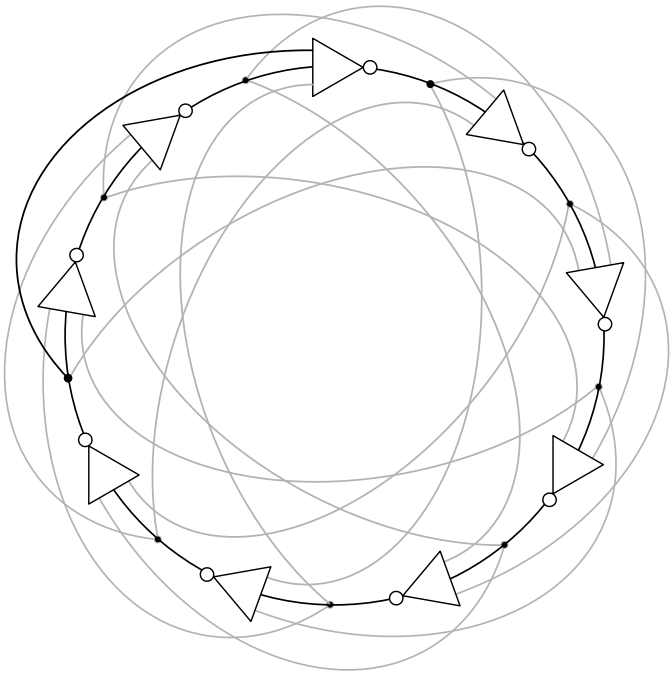
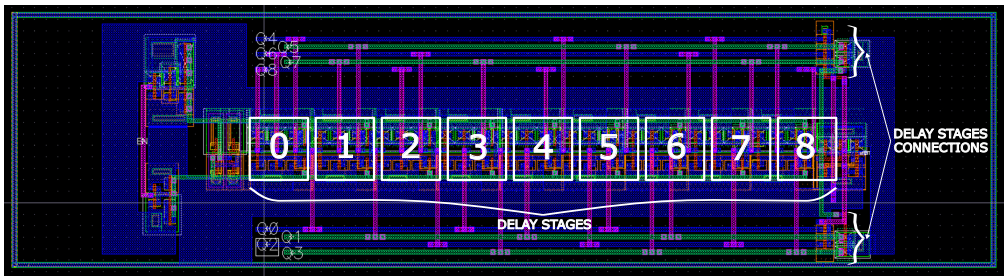
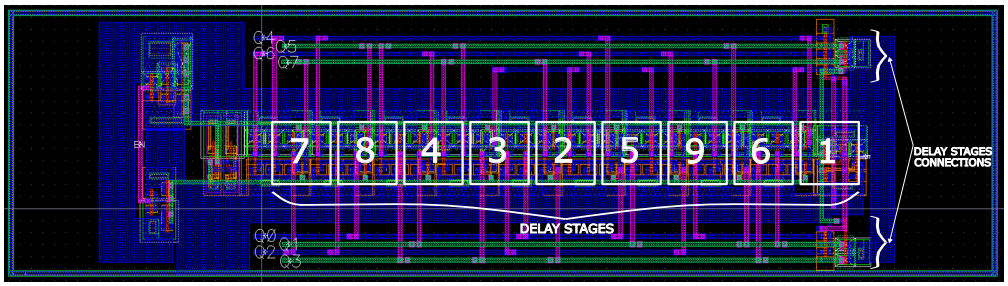


Figure 3.5: MGRO with 9 delay stages of 3 inputs each. Each delay stage has connections along different places in the ring.



(a)



(b)

Figure 3.6: a) First layout arrangement of the delays stages with unequal connections. b) Rearranged delay stages with almost equal metal connections.

Four banks of nine D-flip-flops act as phase detectors. Each bank is triggered by a rising edge on a signal derived from the STOP signal. Each flip-flop data input is connected to a corresponding inverter in the MGRO (k-th flip-flop from each bank to the k-th inverter). The four STOP signals connected to each phase detector bank originate from the main STOP signal delayed using a digitally controlled delay element cell with a quarter of a coarse LSB between each other. The digitally controlled delay element consists of two inverters connected in series with a bank of capacitors connected to the internal node as depicted in Figure 3.7.

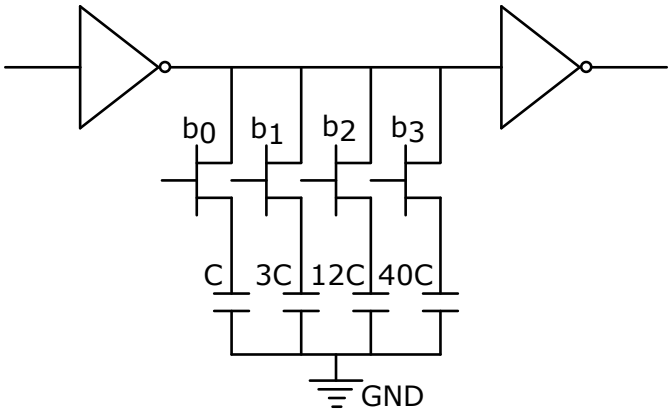


Figure 3.7: Digitally controlled delay element. The delay can be adjusted by controlling the transistors' gates.

The connection to each capacitor is gated by a transistor whose gate is digitally controlled from the outside. The capacitors are sized in such a way that the desired delay range between 350 ps and 450 ps can be obtained (with an increment step of 6.25 ps) in post-layout simulations.

The counter consists of 12 T-flip-flops with an externally controlled reset signal. The counter clock signal is formed by passing the ninth inverter's output (Q_8) through a latch referred to as CLK_TAP cell. The purpose of this cell is to bring the Q_8 phase to logic 0 or 1 when the ring is in high-impedance (the EN signal is low) in order to avoid racing conditions in the counter (uncontrolled oscillations of the T-flip-flops). The EN signal of the CLK_TAP cell is formed by using an SR latch connected to the START and the digitally controlled delayed STOP signal as presented in Figure 3.4.

The TDC's final code is determined by the output processing unit which acts as a decoder and as an error correction unit. The TDC's final output is given by the following equation:

$$TDC_{out} = A + B + C + D + 72 \times E \tag{3.1}$$

where A, B, C and D are the decoded output values of the phase detectors and E is the final counter value. The decoding of the phase detectors output consists of assigning a unique value between 0 and 17 to each state of the MGRO. The 72 factor is given by 4 × 18, where 18 is the number of distinct states of the MGRO and 4, the division factor of the coarse LSB.

The layout of the standalone TDC test chip and its photomicrograph are shown in Figure 3.8. The TDC stand-alone structure is identical with the TDCs present on the bottom tier of Blueberry. For testing purposes, the TDC has a multiplexed START signal and it can be started electrically or by a SPAD.

3.5 On-chip error correction

An on-chip error correction algorithm was implemented in the output processing unit in order to eliminate the bubbles resulting from glitches in the counter value.

The final count value E, is determined using the following equation:

$$E = counter_{value} + cor_{poz} - cor_{neg} \tag{3.2}$$

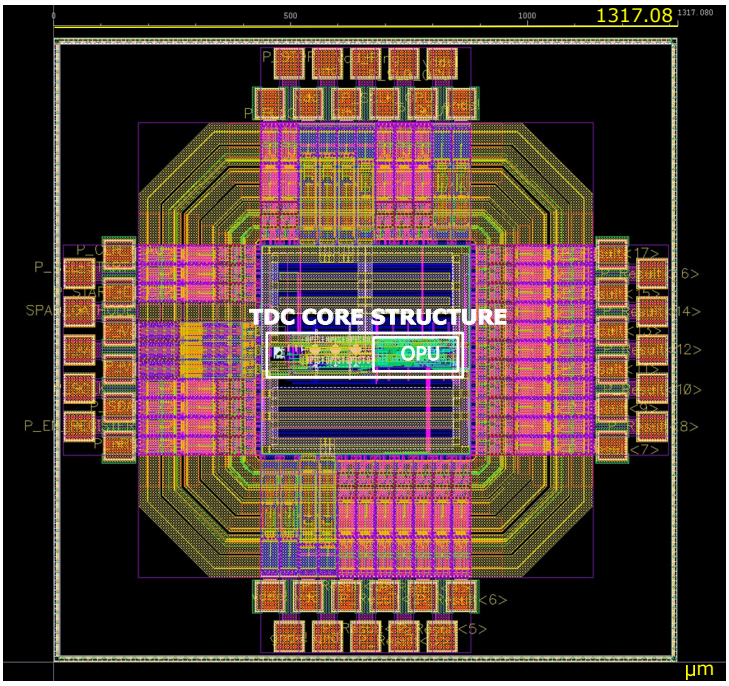
where $counter_{value}$ is the original counter value and cor_{poz} and cor_{neg} are two one bit error correction signals implemented as depicted in Figure 3.9.

There are two scenarios in which the value of the counter might be incorrect, both caused by the arrival of the STOP signal close to a clock transition:

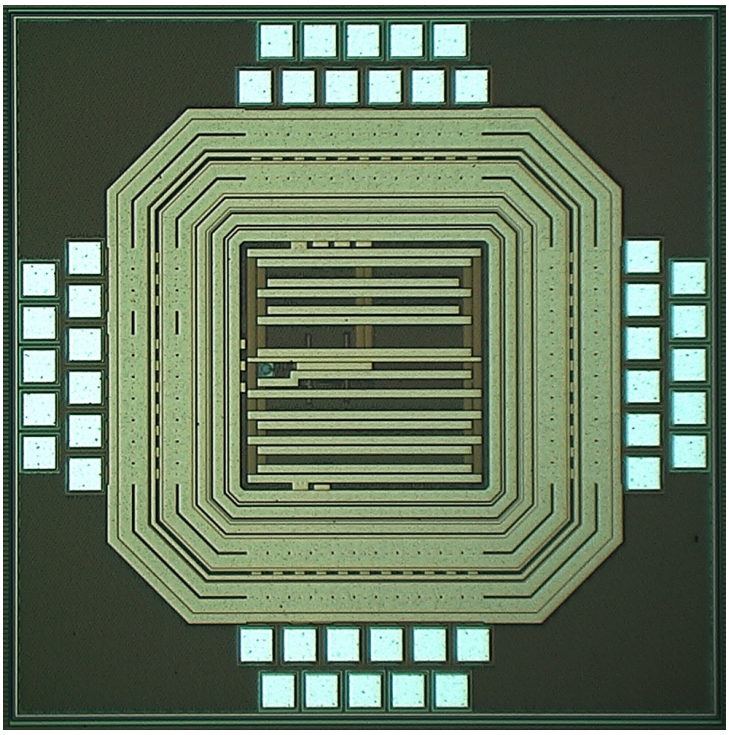
- When the rising edge of the counter clock signal coincides with the arrival of the STOP signal as depicted in Figure 3.10.
- When the falling edge of the counter clock signal coincides with the arrival of the STOP signal as depicted in Figure 3.11.

In the first circumstance, the CLK_TAP cell's feedback can cause its output to stop transitioning and fall back to logic '0'. This results in a short pulse that is not registered by the counter, and therefore the final result is one less than it should be. To compensate for this, the cor_{poz} signal goes high if such an event is detected.

The OPU determines the value of the cor_{poz} signal by monitoring the CLKcount signal and state of the nineth inverter of the MGRO as captured by the first bank of



(a)



(b)

Figure 3.8: Stand-alone TDC a) layout and b) photomicrograph. The TDC's core structure is composed of the MGRO with the phase registers, counter and delay elements and OPU for reading out the data.

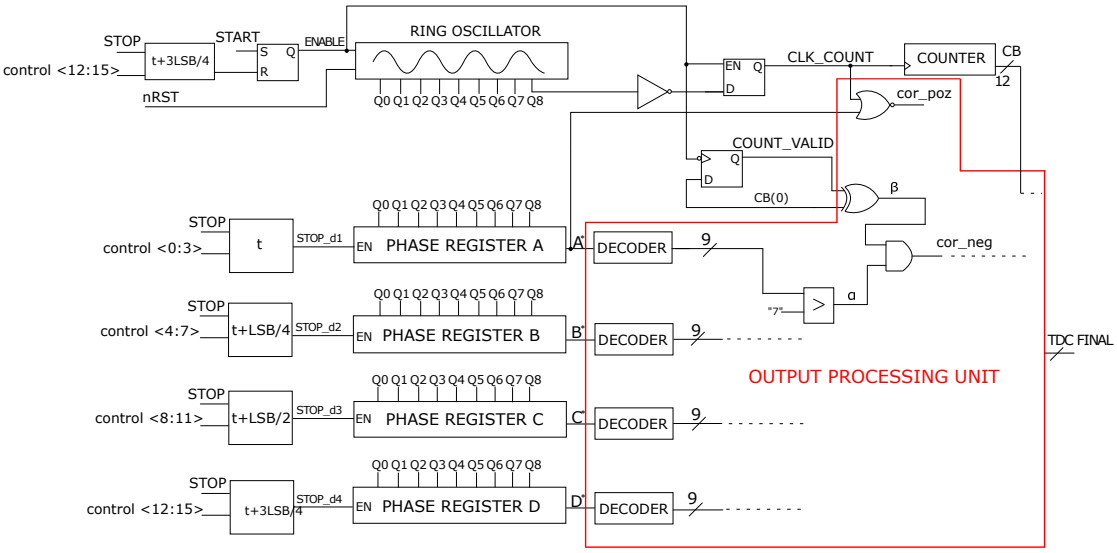


Figure 3.9: Output processing unit with detailed on-chip error correction implementation.

phase registers. In essence, the CLKcount signal has to be one when Q_8 is 0. The relationship between the signals is determined by the following equation:

$$cor_{poz} = Q_8 \; NOR \; CLKcount. \tag{3.3}$$

In the second case, the CLK_TAP cell's feedback can cause its output to stop transitioning and go back to logic '1'. In some cases, the resulting pulse is large enough to trigger an additional count in the counter. This is compensated for by setting cor_{neg} to '1'. Similar to the previous case, the OPU determines the value of the correction signal but this time by monitoring the counter value immediately after the arrival of the STOP signal, the settled value of the counter and the value captured by the first bank of phase detectors. In essence, the least significant bit of the counter should not settle to a value that is different from the one at the arrival of the STOP signal, unless the MGRO has completed a full oscillation cycle. In this case, the relationship between the signals is determined by the following equation:

$$cor_{neg} = (CB(0) \; XOR \; CountValid) \; AND \; (A > 7), \tag{3.4}$$

where CB(0) is the least significant bit of the counter (the settled value which is read out), CountValid is also CB(0) but captured by a flip-flop on the rising edge of the STOP signal and A is the decoded value of the first bank of phase registers.

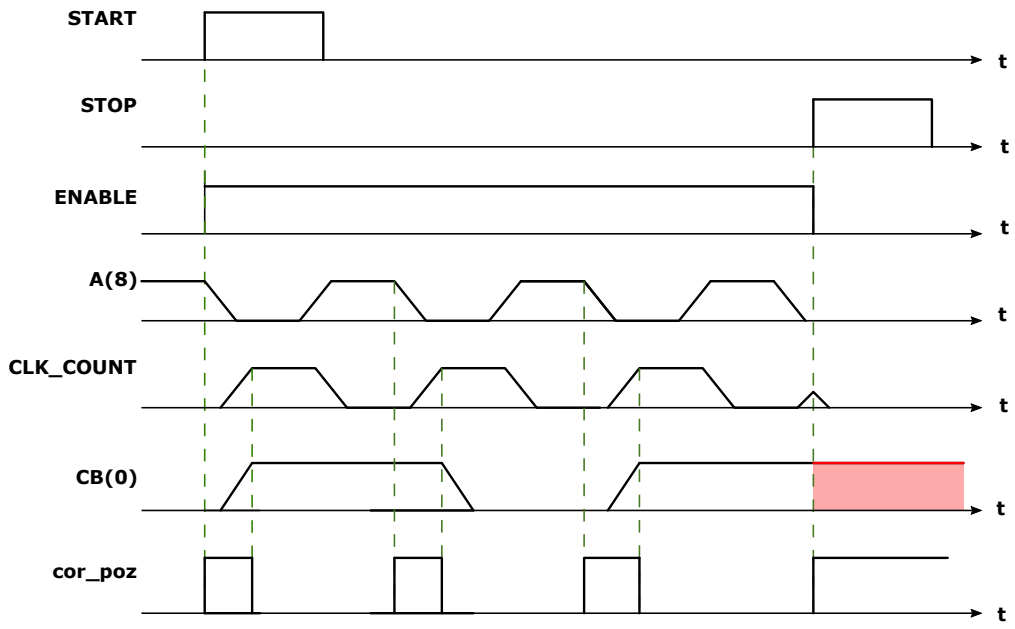


Figure 3.10: On-chip error correction: STOP signal arrives on the rising edge of the counter clock signal. The positive correction (cor_{poz}) signal is set to high.

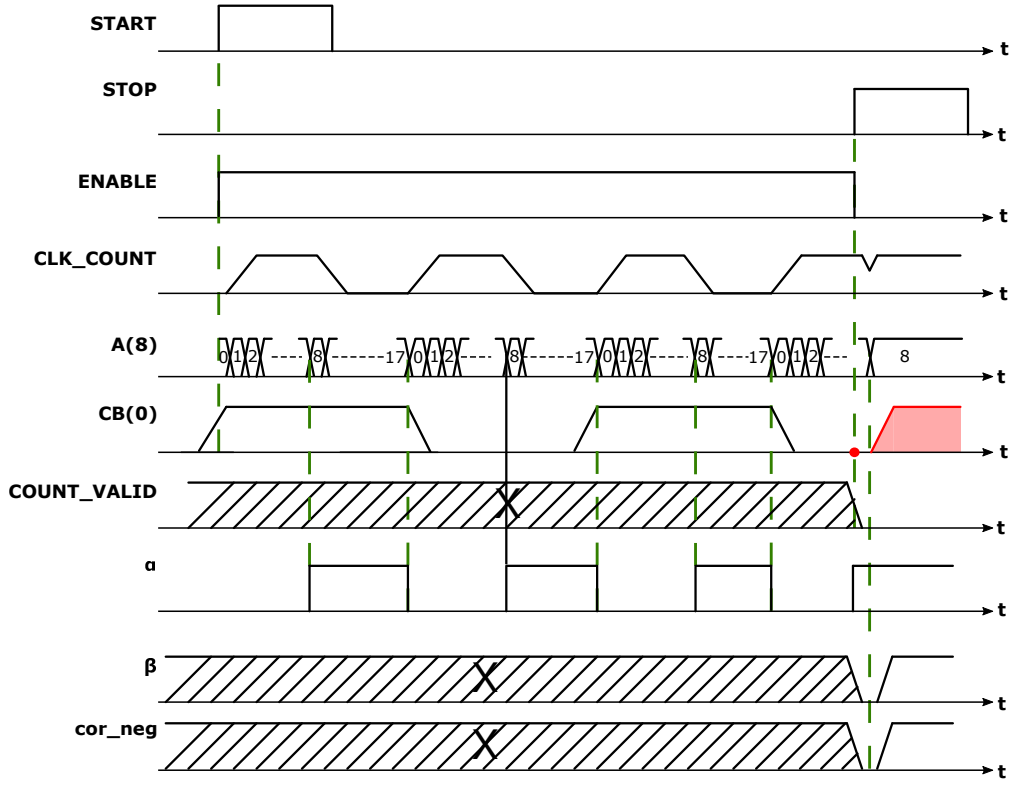


Figure 3.11: On-chip error correction: STOP signal arrives on the falling edge of the counter clock signal. The negative correction (cor_{neg}) signal is set to high.

3.6 TDC calibration

Each TDC comprises four sets of delay cells with four transistors and capacitors as presented in the previous section. In order to create the desired delay through each cell, 16 calibration bits need to be set for each TDC. The TDC's calibration system comprises two row and column shift registers with a serial interaface as depicted in Figure 3.12.

The calibration bits are sent serially for each TDC using a two wire serial interface (serial clock (SCK) and a the serial data input (SDI)), distributed using a binary tree. Due to process variations, a single set of calibration bits cannot suit all 64 TDCs in the array. In this way, each TDC can be accessed and calibrated independently.

3.7 Decision tree

A decision circuit has been implemented per cluster in order to determine the first firing pixel within a burst of events. The pixel outputs are connected to the decision circuit at the first level and a comparison is made between groups of 2 pixels at each level. The decision circuit is presented in Figure 3.13. When nRST signal is active ('0' logic), M2 and M5 become active. If I_0 switches before I_1, M3 is activated, discharging the gate of M5. In this way, I_1 is prevented from propagating through the cell. In the case in which, I_1 switches first, the gate of M2 is discharged and the propagation of I_0 is blocked.

The first arriving event is then propagated down to the next level and continues like this until the sixth level ($\log_2 64$, where 64 represents the total number of SPADs in a cluster) where all SPAD events are analyzed. At the end, a single output is provided which determines the winning event corresponding to that cluster. The winning event then triggers the TDC. A SPAD address tree has been implemented in each cluster as represented in Figure 3.14. At the end, a 6 bit encoder that takes as an input all the outputs of the decision circuit returns a 6-bit binary code providing the address of the pixel that fired first.

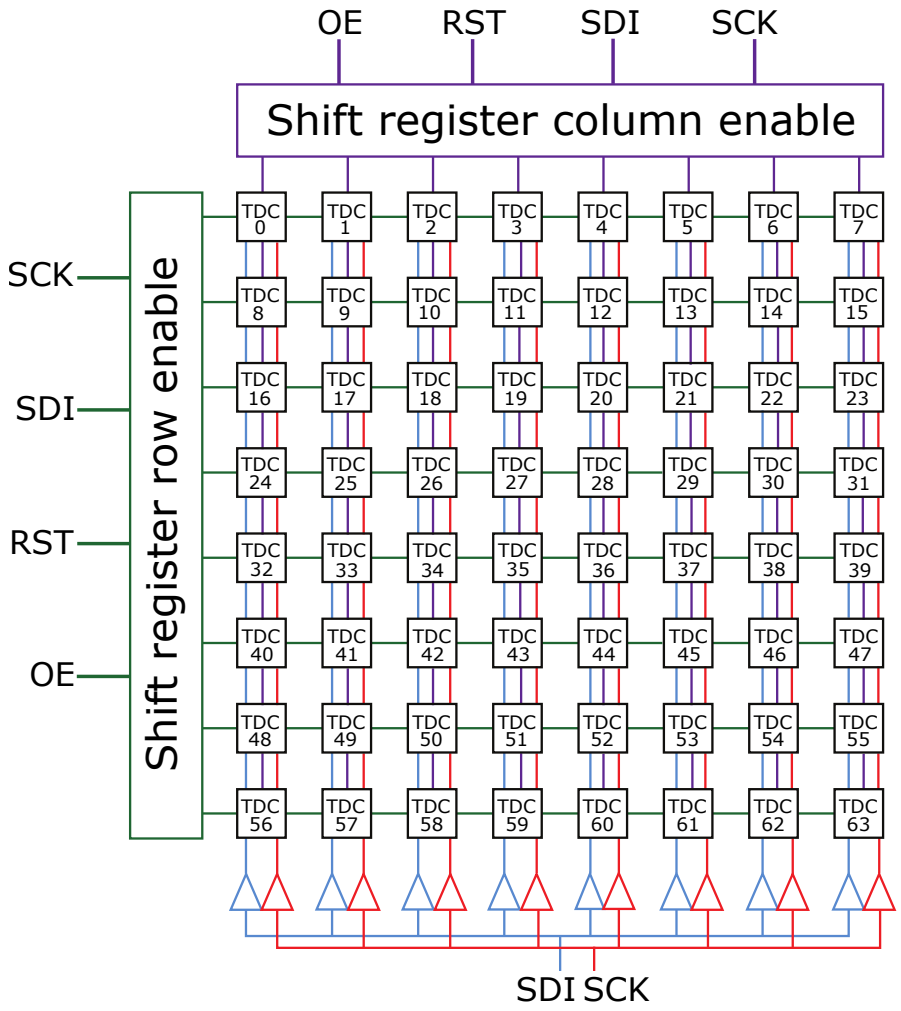


Figure 3.12: Conceptual diagram of the TDCs' calibration system. Two row and column shif registers are used to load the calibration bits into the TDCs. Two binary trees are used for serial communication: red tree - serial clock, blue tree - serial data. Serial interface: SCK - serial clock, SDI - serial bus for the data, OE - output enable to activate the parallel output of the shift registers, RST - reset signal [36].

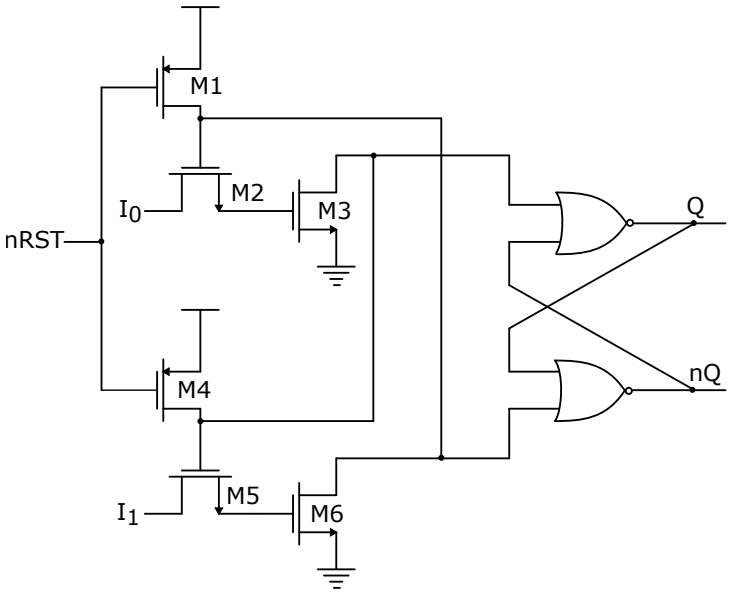


Figure 3.13: Decision circuit based on a WTA approach. The circuit determines which input fired first (I_0 or I_1) based on precharged logic.

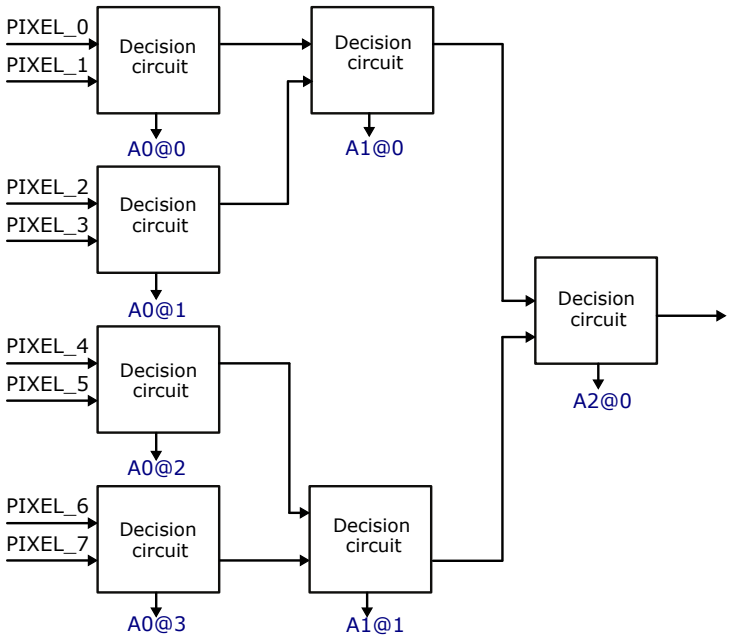


Figure 3.14: Decision tree concept presented on 8 pixels. The schematic can be extended to N pixels. AX@Y is the X-th address bit corresponding to decision element Y.

3.8 Characterization results

3.8.1 Time-to-digital converter

The TDC stand-alone structure was characterized by performing electrical and optical measurements.

First, the transfer function of the TDC was measured by sending START and STOP signals with a FPGA (Opal Kelly XEM7360 with Xilinx Kintex-7). The rising edge of the START signal triggers the TDC, while the rising edge of the STOP signal ends the measurement cycle. Each time interval of the START and STOP signals was sampled by the TDC 81920 times over a range of 85 ns. The transfer function of the TDC is shown in Figure 3.15 and it indicates an LSB of 15 ps.

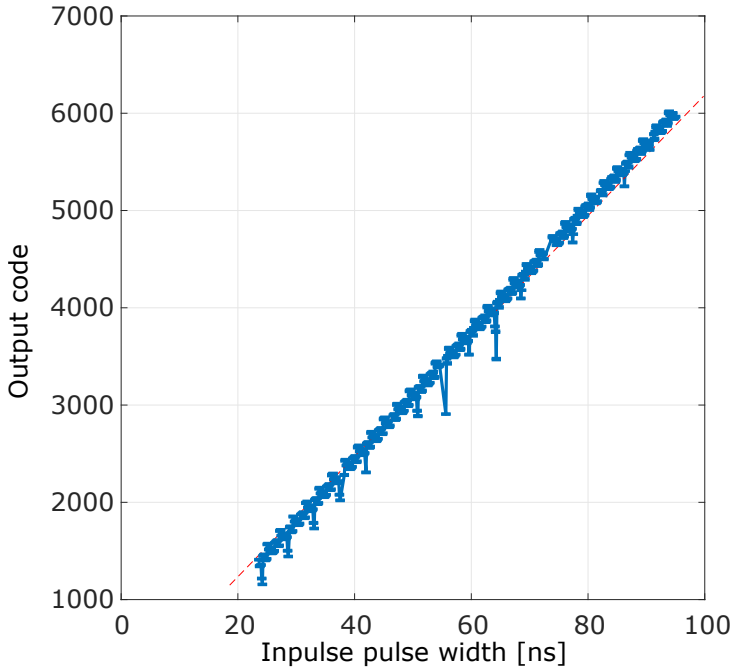
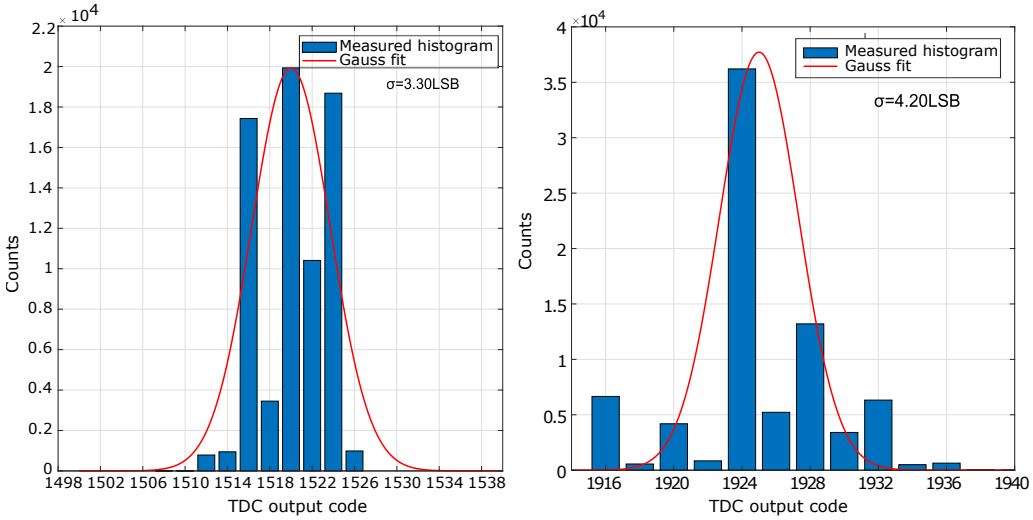


Figure 3.15: TDC transfer function.

Second, electrical single-shot precision results can be extracted from the TDC's transfer function. Because each time interval in the transfer function was sampled 81920 times, the standard deviation in the TDC's output code can be determined. Two different single-shot precision measurements are presented in Figure 3.16 and indicate a sigma of approximately 3 LSBs.

The large sigma can be reduced by proper calibration of the TDC (changing the calibration bits of the delay cells). An example is shown in Figure 3.17. In this case,



(a) (b)

Figure 3.16: a), b) Single-shot precision of different TDC output codes.

single-shot measurements were performed for an EN width of 52.8 ns where 16000 iterations were run for each point. A total of 100000 calibration code combinations were checked and the standard deviation was calculated for each of them. In this particular single-shot configuration, for an EN width of 52.8 ns, the sigma could be reduced from 28.9 LSBs to 3.8 LSBs.

For the measurements performed on the stand-alone structure a specific combination of the calibration bits was used. This combination was determined through different measurements and it proved to be the best choice from all the tested cases. However, this does not guarantee that this combination can be extended to the TDCs present on the bottom tier of Blueberry. In the case in which the TDCs present on the bottom tier are tested, each of them has to be calibrated individually. The calibration part is time consuming considering a total of 64 TDCs and has to be implemented carefully for each TDC in order to have a uniform TDC LSB for the entire array. Moreover, this technique proves to be cumbersome in the case in which the system is scaled up and a more robust method has to be considered. The code combinations which were tested for these measurements are present in Table 3.1

The TDC's non-linearities were measured through a code density test. The TDC's START signal is connected to a single SPAD which was illuminated with white light, thus the photons were randomly distributed in time. The TDC's timestamps were read out over a long period of time. In the ideal case scenario, the final histogram should be uniform, however, this is not possible due to different process and design variations. The DNL and INL results are shown in Figure 3.18 with 90% of the results presenting a

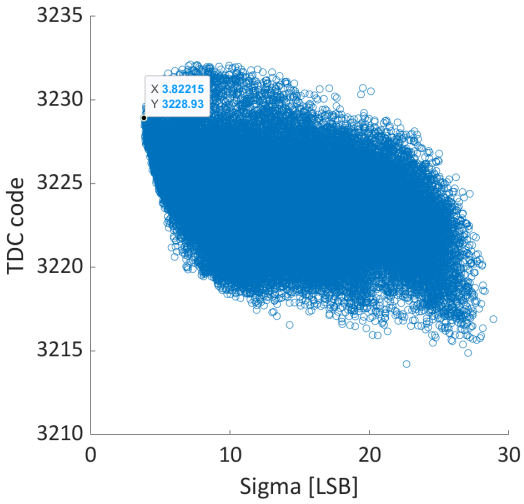


Figure 3.17: Pareto diagram of the TDC's calibration bits. Each blue circle represents a code combination of the delay cell. The standard deviation was measured on 16000 iteration points for each code combination for an EN width of 52.8 ns. The standard deviation can be reduced from 28.9 LSBs to 3.8 LSBs in this particular case.

Table 3.1: Calibration code combinations used during the characterization of the stand-alone TDC.

Register	CODE0	CODE1	CODE2	CODE3
CB	0000	0000	0000	0011
A	0000	0000	0111	1011
B	0000	0011	0111	1000
C	0000	0110	0111	0101
D	0000	1001	0111	0001

DNL of -1/2.45 LSB and an INL of -0.26/3.77 LSB. The INL was compensated for by using a LUT. These measurements were performed in a single control bits configuration (CB - 0000, A - 0000, B - 0011, C - 0110, D - 1001).

The average power consumption of the TDC stand-alone structure is 1.4 mW which translates into an estimated total power consumption of 89.6 mW for all 64 TDCs present on Blueberry's bottom tier.

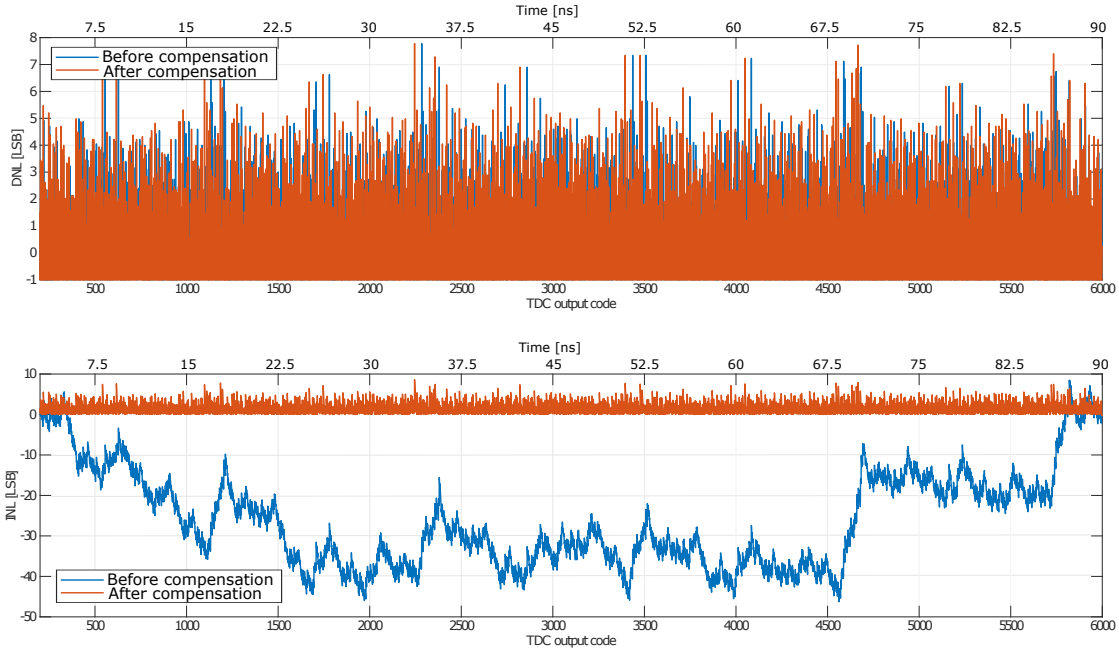


Figure 3.18: TDC's non-linearities - DNL and INL. Before compensation (blue) and after compensation (red).

3.9 Conclusions

This chapter presented the design concept and implementation of the first 3D-Stacked multi-Digital SiPM implemented in a FSI fashion. Considering the FSI implementation, Blueberry is a suitable photodetector for PET due to its sensitivity in the blue spectrum. The 3D integration facilitated the use of a separate top tier for the photosensitive area and a dedicated bottom tier which comprises solely the readout electronics. Different limitations were set by the manufacturing process which limited the detector to 7.5 mm × 4.2 mm. The SPADs present a fill factor of approximately 67%. An architectural overview of the entire sensor was presented in this chapter with a focus on the electronic blocks designed by the author. The TDC exhibits an LSB of 15 ps and a DNL of -1/2.45 LSB and an INL of -0.26/3.77 LSB for more than 90% of the results. An on-chip error correction algorithm was implemented for the TDC readout which was explained in detail. The TDC's timing resolution was obtained through a reconfigurable delay cell whose delay is controlled through different calibration code combinations. A Pareto diagram example of how the calibration bits influence the standard deviation of the TDC was presented. The TDC's timing resolution is an important parameter which contributes to the overall timing resolution of a PET detector module. One of the goals for the author was to design a TDC which pushes the limits in terms of timing resolution considering that the TDC is an important readout element

for a PET photodetector module. The SPAD address tree circuit implemented in each cluster was described in detail. The chapter included all the design, analysis and measurements performed by the author, additional characterization results related to the SPAD design and performance are presented in [36]. A comparison between Blueberry sensor and other SPAD-based sensors designed in similar technology nodes is presented in Table 3.2.

Compared to the other works presented in the table, Blueberry presents a high fill factor of 67% due to the nature of the 3D integration. The TDC achieves the best LSB of 15 ps, however with larger integral nonlinearities compared to other works. This is probably due to the nature of this topology, which is based on delay stages with three inputs, four sets of phase registers and delay calibration cells. The Blueberry TDC DNL of -1/2.45 LSB (LSB = 128 ps) is smaller than the ones reported by the other works. The DNL and INL can in principle be improved by using the delay calibration cells. Due to the complexity of this TDC, a larger power consumption was expected. This is visible in the comparison table where [130] and [51] have similar total power consumptions but more TDCs.

Table 3.2: Blueberry comparison with SPAD based detectors.

Parameters	Blueberry	[130]	[131]	[51]	[132]	[133]
Technology [nm]	180 CMOS [a]	180 CMOS	180 CMOS	130 CMOS	150 CMOS	180 CMOS
SiPM type	D-SiPM	D-SiPM	D-SiPM	D-SiPM	D-SiPM	D-SiPM
Microcell size [μm^2]	50× 50	28.5 × 28.5	28.5 × 28.5	25.4 × 25.4 [b]	25 × 25	55.66 × 64[f]
Microcells #	4096	1024	36288	92160	3840	3200/6400
Fill factor [%]	67	28	28	35.7	32.1	77.7[f]
PDP	55% @ 480 nm Vex = 6 V	47.8% @ 520 nm Vex = 5 V	47.8% @ 520 nm Vex = 5 V	45% @ 450 nm Vex = 1.5 V	NA	38.9% @ 450 nm Vex = 3.3 V
TDC type	antiphased MGRO	GRO	GRO	GRO	GRO	NA
Total # TDCs	64	128	1728	128	128	2[f]
TDC LSB [ps]	15	48.8	48.8	64.5	80	23.5[f]
Dynamic range [ns]	3932	204	204	261.59	81.84	12
DNL, INL [LSB]	-1/2.45 -0.26/3.77	-0.07/0.08 -0.38/0.75 [c]	-0.48/0.48 0.89/-1.67	-0.24/0.28 -3.9/2.3 [d]	-0.19/0.20 -2.40/0.35 [d]	< 0.5[f]
TDC area [mm^2]	0.0191	0.004	0.004	NA	NA	NA
Power/TDC [mW]	1.4	0.73 [e]	0.3	940 μ	NA	NA

[a] 3D-FSI.
[b] hexagonal arrangement.
[c] measured over 25% of DR.
[d] measured over 61% of DR.
[e] assuming 35.5 Mevents/s [134].
[f] for DPC3200-22-44 [133].

4 Smarty: An on-chip neural network

4.1 Motivation

In this chapter, we discuss Smarty, an on-chip neural network. This sensor has been designed as a proof of concept for a digital front-end that contains both timestamping circuitry and a neural network. Ten independent channels are directly connected to a reconfigurable on-chip neural network which can be trained for the desired task. In this case, the focus is kept on reconstructing the position of a radioactive source between two photodetectors facing each other. The data from the ten input channels is reduced to a single word per frame and special cases where not all the channels have fired are automatically discarded by the neural network through inference. Compared to other neural networks used in PET, which are mostly implemented in FPGAs or GPUs [89], [96], [135], the particularity of this chip is based on its full reconfigurability and on-chip design. The latter imposes more constraints in the design phase, such as area, architecture, as well as reconfigurability. Most of the neural network approaches are implemented either in FPGAs or GPUs, while Smarty is a fully integrated neural network. Besides this, the neural network can be accessed independently so that it could be trained for different purposes desired by the user.

A potential issue of the PET systems is the large amount of raw data generated during acquisition. A typical system can consist of tens of thousands of photodetectors [136], each one representing an independent channel whose data needs to be processed simultaneously with the rest. In order to implement the image reconstruction algorithm, the processing electronics must be capable of handling the large input data throughput. In practice, pre-processing of the sensor data is required in order to mitigate the effects of random coincidences, scattered and attenuated gamma photons so that the image quality is improved.

Neural networks are attractive candidates for image reconstruction [89], [135], as well as for data acquisition and pre-processing. Neural networks with thousands of inputs have already been demonstrated [137], [138] in the field of image processing which makes them very attractive candidates for the high throughput and large number of channels in a PET system. In addition, because of the nature of the training procedure, the pre-processing steps such as TDC gain correction or elimination of invalid frames can be inferred based on the desired training output. In this way, focus is kept on the desired task, such as coincidence detection. Moreover, because the pre-processing circuits are integrated, and only the distilled data needs to be forwarded to the rest of the processing electronics, the throughput is significantly reduced.

4.2 Neural network modelling

The first step towards the implementation of the neural network consisted in creating a model based on a small fully-reconfigurable neural network. In order to implement the on-chip neural network, a mathematical description of it is necessary. An example NN was used to create this model and it can be extended to a larger feed-forward artificial neural network topology.

In this section the mathematical model of a NN example describing the architecture of Smarty's neural network is formalized so that it serves as a basis for the on-chip implementation. Smarty's neural network is highly reconfigurable; the weights, biases, and even the network's topology can be modified on demand. By adjusting the weights and biases, learning is improved, therefore the NN will generate a more accurate result. Configuring Smarty is done by simply updating a topology file in which information regarding the number of input neurons, number of hidden layers and hidden neurons as well as output neurons is stored. As depicted in Figure 4.1, a fully-connected artificial neural network example followed by its mathematical model and the required variables which describe the architecture is depicted. This approach can be extended to other topologies of feed-forward NNs.

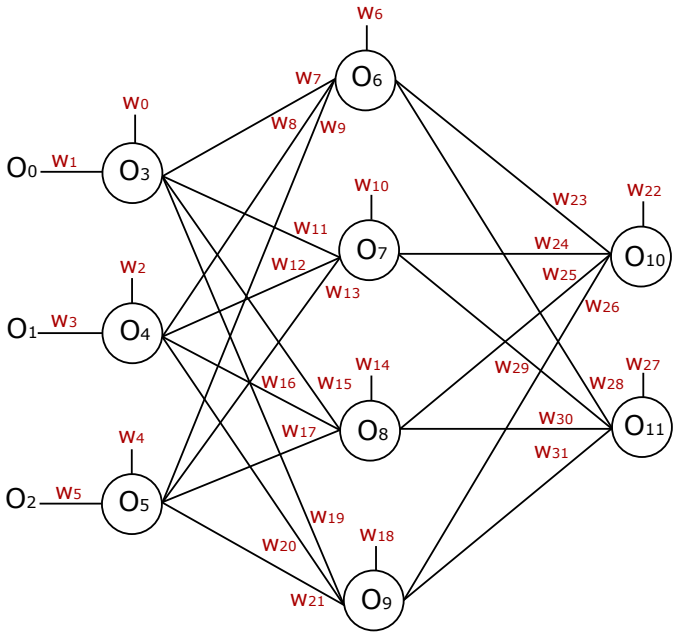


Figure 4.1: Example of a fully-connected neural network with three neurons in the input layer: (O_3, O_4, O_5), one hidden layer of four neurons: (O_6, O_7, O_8, O_9) and two output neurons (O_{10}, O_{11}). For each connection, the weights and biases are depicted.

The NN is mathematically described as follows:

$$
\begin{aligned}
O_3 &= w_0 + \sum_{i=1}^{1} w_i \times O_{i-1} = w_0 + w_1 \times O_0 \\
O_4 &= w_2 + \sum_{i=3}^{3} w_i \times O_{i-2} = w_2 + w_3 \times O_1 \\
O_5 &= w_4 + \sum_{i=5}^{5} w_i \times O_{i-3} = w_4 + w_5 \times O_2 \\
O_6 &= w_6 + \sum_{i=7}^{9} w_i \times O_{i-4} = w_6 + w_7 \times O_3 + w_8 \times O_4 + w_9 \times O_5 \\
O_7 &= w_{10} + \sum_{i=11}^{13} w_i \times O_{i-8} = w_{10} + w_{11} \times O_3 + w_{12} \times O_4 + w_{13} \times O_5 \\
O_8 &= w_{14} + \sum_{i=15}^{17} w_i \times O_{i-12} = w_{14} + w_{15} \times O_3 + w_{16} \times O_4 + w_{17} \times O_5 \\
O_9 &= w_{18} + \sum_{i=19}^{21} w_i \times O_{i-16} = w_{18} + w_{19} \times O_3 + w_{20} \times O_4 + w_{21} \times O_5 \\
O_{10} &= w_{22} + \sum_{i=23}^{26} w_i \times O_{i-17} = w_{22} + w_{23} \times O_6 + w_{24} \times O_7 + w_{25} \times O_8 + w_{26} \times O_9 \\
O_{11} &= w_{27} + \sum_{i=28}^{31} w_i \times O_{i-22} = w_{27} + w_{28} \times O_6 + w_{29} \times O_7 + w_{30} \times O_8 + w_{31} \times O_9
\end{aligned} \tag{4.1}
$$

Any fully-connected NN of any dimension can be described mathematically as in the above model. Additionally, the layer parameters which describe its architecture can be derived from the model as such:

- S_{ant}: the index of the first neuron of the layer in front of the current layer, represented as 7 bit unsigned integer.
- N_{ant}: the number of neurons of the layer in front of the current layer, represented as 7 bit unsigned integer.
- S_{act}: the index of the first neuron of the current layer, represented as 7 bit unsigned integer.
- N_{act}: the number of neurons of the current layer, represented as 7 bit unsigned integer.
- S_w: the index of the first weight or bias from the current layer, represented as 10 bit unsigned integer.

An example of the parameters describing the third layer of the neural network presented in Figure 4.1 are shown in Table 4.1. The topology file describes each layer of the neural network.

Table 4.1: Parameters describing the third layer of the NN presented in Figure 4.1.

S_{ant}	N_{ant}	S_{act}	N_{act}	S_w
6	4	10	2	22

All layers parameters are included in the topology file and determine the NN's architecture. Parameters which describe the NN topology are stored in a 624 bit memory, and the weights and biases are stored in a 10.24k bit memory.

A Matlab floating point model of the NN is used as a reference. The outputs of this model for a specific set of inputs are considered to be correct and will be referred to from now on as golden outputs. The NN's golden outputs are obtained through the Matlab code (it can describe any feed-forward NN topology and it uses floating point), which allows for simulation of the system's performance when floating point precision is used. Floating point precision is important to quantify the NN's capabilities. However, due to its large resource requirements, its on-chip integration is discouraged.

High-level synthesis (HLS) is an automated tool that translates a high-level programming language (such as C) into a register transfer level (RTL) hardware description.

Smarty's NN has been firstly described in C and HLS was used for its RTL implementation. The performance of the NN in Matlab is then compared to the C implementation which is bounded by the use of fixed point precision. The modelling code flow is depicted in Figure 4.2.

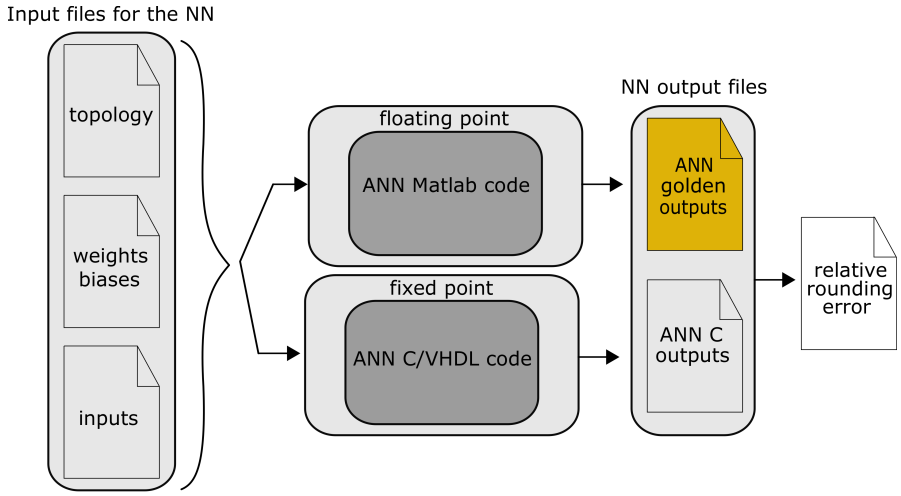


Figure 4.2: Schematic of Smarty neural network modelling procedure. The performance of the NN in Matlab (floating point) is compared to that of the HLS-inferred system (fixed point bounded). At the end, the obtained results are compared.

In order to analyze the NN performance in Matlab, three sets of parameters are needed: the NN's layers configuration (the previously described topology file), the input values, i.e. TDC codes, which are determined by taking into account the TDC range, and a set of weights, which were randomly chosen in Matlab considering a range of [-1, 1]. At the end, a file with the output values of each neuron in the NN is obtained.

The number of fractional bits required for the fixed point implementation was determined by analyzing the rounding error present at each neuron of the neural network when all the weights are random numbers between [-1, 1], sampled from an uniform distribution. A large NN considering the limits of this design, of 10 input neurons, 4 hidden layers of 8 neurons and a 6 neuron output layer was analyzed. The 20-bits TDCs' values were simulated as random number sampled from an uniform distribution across four different ranges: [0, 300000], [300000, 600000], [600000, 900000], [900000, 1000000]. The same files for the layer configuration, inputs and weights are used for the HLS test bench. The two output files, the Matlab-generated golden outputs and C are compared by checking the relative error of the outputs of the NN (Matlab - floating point, C - fixed point) as depicted in Figure 4.3.

The resulting relative rounding error is less than 0.03 % across all TDC input ranges, when an 8 fractional bit representation is used. In conclusion, considering the small

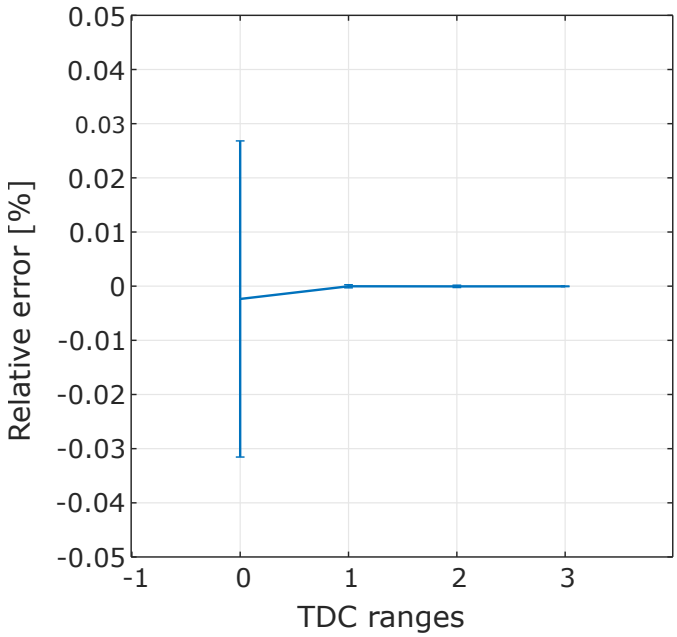


Figure 4.3: Smarty NN's outputs relative rounding error obtained by comparing the golden outputs in floating point and the fixed point outputs. TDC ranges: 0 - [0, 300000], 1 - [300000, 600000], 2 - [600000, 900000], 3 - [900000, 1000000].

error, the on-chip design was implemented with 8 fractional bit representation for the neurons and coefficients values.

4.3 System architecture

Smarty is part of a SoC developed in TSMC 16 nm FinFET process and comprises ten TDCs connected with an on-chip fully-connected reconfigurable NN, along with a stand-alone reference oscillator as presented in Figure 4.4. The photomicrograph of the entire SoC is illustrated in Figure 4.5.

In the following section, the architecture of the main electronic blocks such as: TDC, NN and the entire system will be described in detail, along with important considerations that needed to be made in order to meet the area constraints.

4.3.1 Time-to-digital converter

Each of Smarty's 10 TDCs comprises three main structures: a VCO, an asynchronous ripple-counter and a thermometer decoder connected as depicted in Figure 4.6.

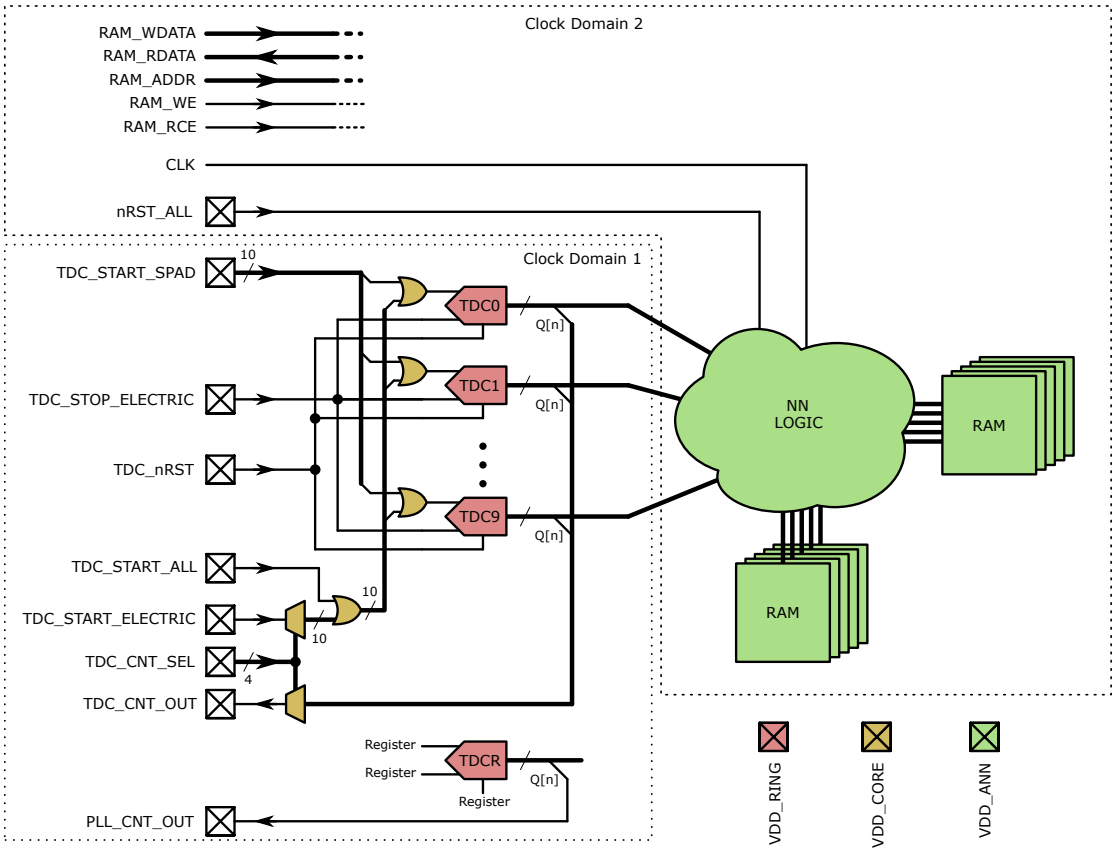


Figure 4.4: Smarty block diagram. The outputs of the 10 TDCs represent the inputs to the neural network. A stand-alone reference oscillator is used for calibration. Two dual port memories are used for storing the weights and biases and the NN outputs. The TDCs can be bypassed and the neural network can be used as a stand-alone structure.

The 20-bit counter increments upon each completion of a ring cycle and returns the most significant bits, while the least significant bits are provided by the four intermediate VCO's outputs ($Q<0:3>$) through the thermometer decoder. The VCO consists of four delay stages as illustrated in Figure 4.7 based on three types of standard cells: buffers, inverters and NAND gates. The VCO starts oscillating on the rising edge of the EN signal, which is formed by the SR latch as depicted in Figure 4.8. At the falling edge of the EN signal, the oscillation stops in its current state which can be read out using the four outputs $Q<0:3>$.

In order to reduce the power consumption, the VCO's outputs are buffered and are only available when the EN_read signal is asserted. Signal $Q<3>$ is an exception because it acts as a clock signal for the counter, and therefore always needs to be enabled. An always-on dummy buffer is present to balance the load across the four stages of the oscillator. The least significant bits are then interpreted by the thermometer decoder

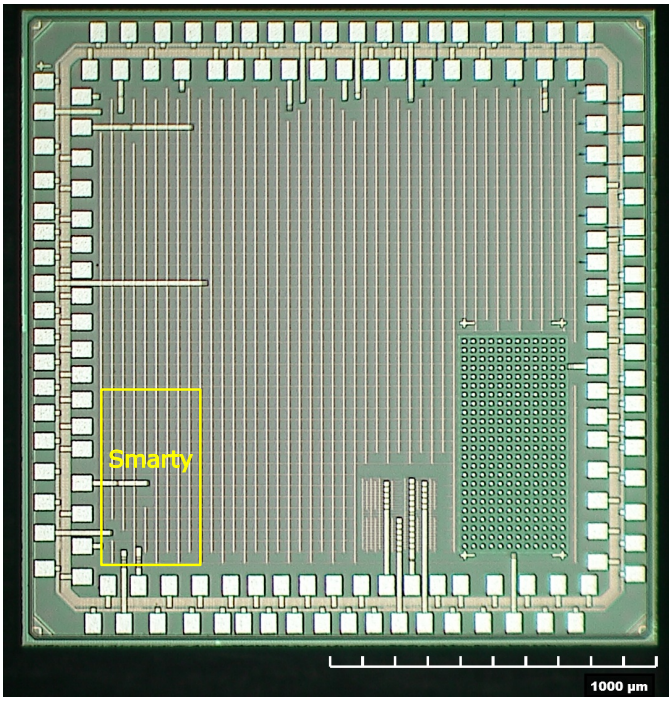


Figure 4.5: Overall SoC photomicrograph. Approximate representation of Smarty location is shown in the yellow rectangle.

and the two LSBs of the output code are generated.

The TDC thermometer decoder takes the four VCO bits, signals $Q < 0:3 >$, as an input and converts them into a 2-bit number. Additionally, the thermometer decoder checks the validity of each code by setting B2 bit to $'0'$ if the code is not valid, and to $'1'$ for a valid code. The logic functions of the thermometer decoder along with their implementation are illustrated in Figure 4.9.

The TDC layout is depicted in Figure 4.10.

The final result of the TDC is calculated as:

$$N_{result} = 4 \times N_{coarse} + N_{fine}, \tag{4.2}$$

where N_{coarse} is the counter value and N_{fine} is the decoded fine bits value.

Each of the ten TDCs can be read out independently by utilizing two different signals: TDC_START_ELECTRIC and TDC_STOP_ELECTRIC, both generated by a FPGA board (Opal Kelly XEM7360 with Xilinx Kintex-7). During electrical testing, a START signal is generated with the same frequency but an adjustable phase with respect to the STOP signal. This allows different impulse widths to be fed into the TDC, therefore sweeping

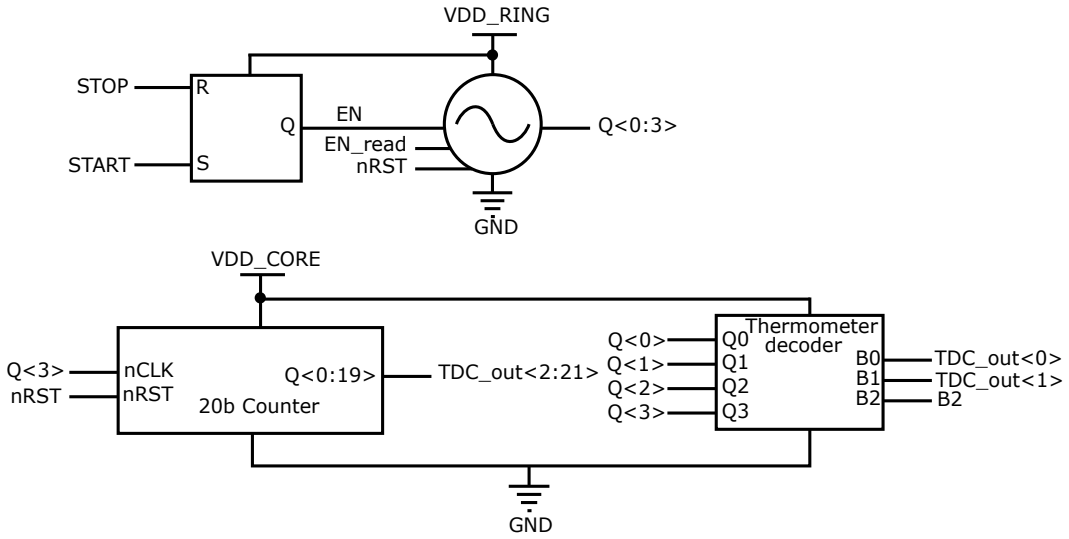


Figure 4.6: Smarty TDC architecture which comprises three main blocks: a VCO that returns the four phases of the TDC ($Q<0:3>$), a 20-bit asynchronous ripple counter that returns the MSBs of the TDC, and a thermometer decoder which returns the LSBs of the TDC code.

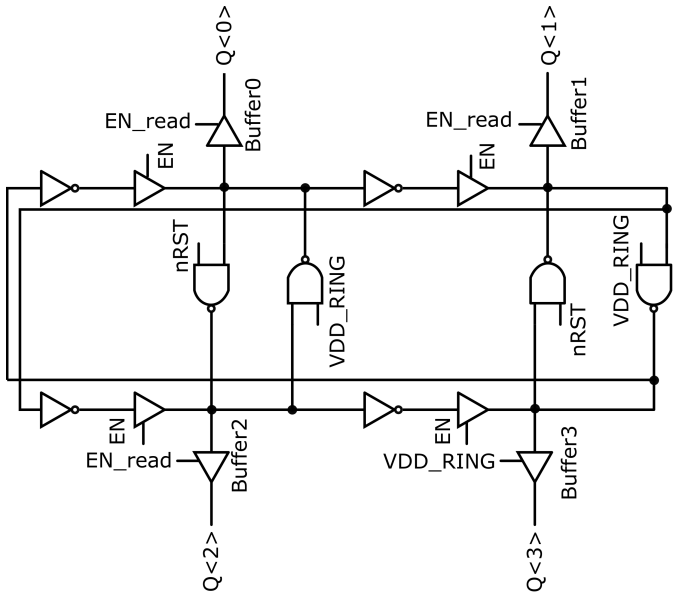


Figure 4.7: The ring oscillator structure implemented in Smarty TDC based on four delay stages [139].

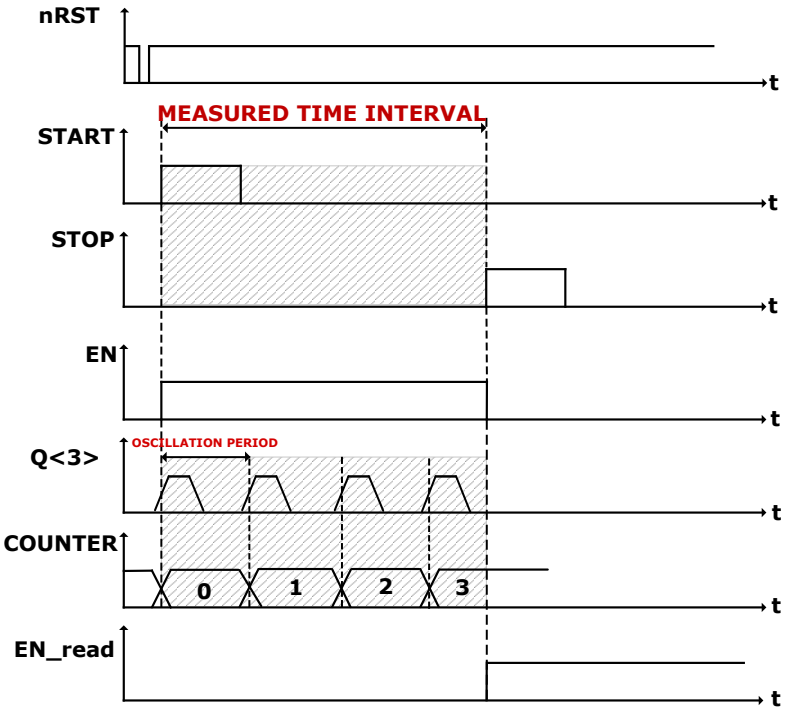
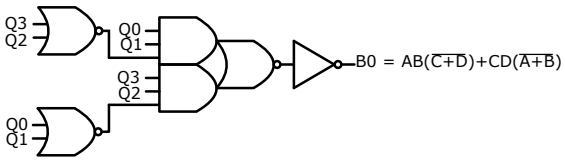
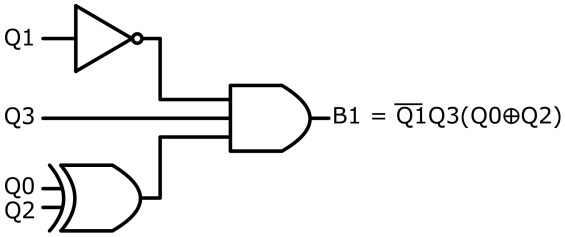


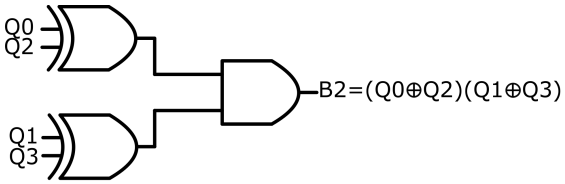
Figure 4.8: Smarty TDC operating principle. $Q<3>$ is the only transparent signal that is the counter CLK. The other signals, $Q<0:2>$ are not transparent after the buffers. The VCO oscillates only while the EN signal is set to high.



(a)



(b)



(c)

Figure 4.9: Thermometer decoder logic functions.

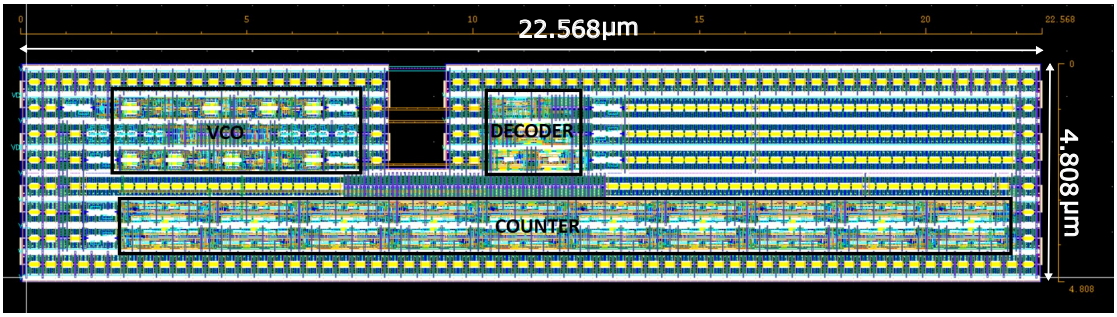


Figure 4.10: Smarty TDC layout. Only a portion of the decoupling capacitor bank is shown.

a larger TDC range. Alternatively, all the TDCs can be electrically measured altogether by using the TDC_START_ALL and TDC_STOP_ELECTRIC. In this way, the required measurement time is decreased and the same impulse width can be measured by all the TDCs so that the testing procedure can be sped up and simplified. In order to determine the oscillation period, the 7th counter bit of each TDC can be read out by the TDC_CNT_SEL, so that an analysis of the performance of all TDCs can be made.

The frequency of a reference stand-alone TDC (whose oscillator is identical with the VCOs of the other TDCs) that continuously oscillates is compared to the frequency of a reference signal. Considering that the supply voltage of the reference TDC (VDD_RING) is the same as for the other ten, one can assume that the frequency is the same for all of them. However, the influence on the substrate potential of the output load due to unequal routing led to body effect in the TDCs' VCOs. As a result, the reference TDC oscillates at a higher frequency compared to the other ten. However, this factor is deterministic and can be accounted for when using the reference TDC.

4.3.2 On-chip neural network

In order to facilitate an easier implementation, Smarty's neural network was written in C, and translated by a high-level synthesis tool. Due to its automated design process, the NN register-transfer level implementation was built based on an abstract behavioral specification of the digital system. Moreover, architectural optimizations such as Kernel optimizations (reduce area and device usage, reduce latency etc.), loop unrolling (creates multiple copies allowing iterations to occur in parallel), loop optimizations to reduce latency etc., pipeline are available so that the final design can be optimized accordingly. The optimizations were introduced in the HLS code that describes the NN's design.

The NN'S main blocks are depicted in Figure 4.11.

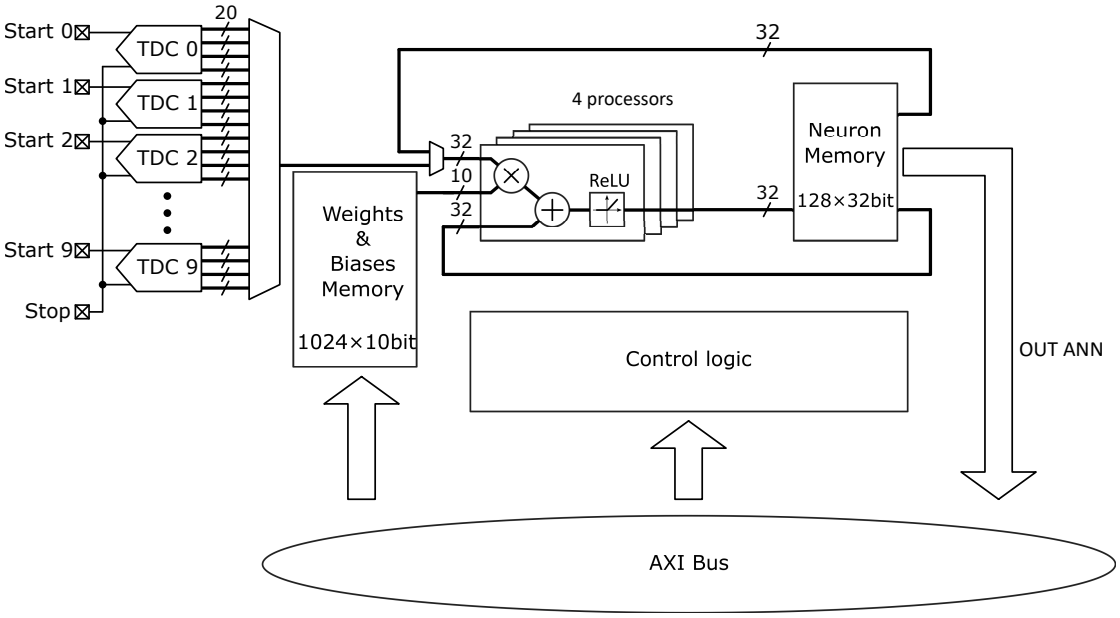


Figure 4.11: 10 TDCs connected to the NN. The NN's main blocks: weights and biases memory, neuron memory and the control logic unit. The communication with the NN is done through an AXI bus.

The NN comprises three main memory blocks. The neuron memory is a dual-port 4.096 k bit memory and contains the values for all the neurons from all the layers in the NN. The 10.24 k bit dual-port coefficient memory (weights and biases memory) comprises all the coefficients corresponding to all the connections in the NN. The coefficients are loaded into the memory after they have been determined by the user through extensive training. Additionally, a third small 624 bit memory is needed in order to store the topology of the NN as previously described. The implementation of an on-chip NN requires the usage of arithmetic and memory blocks which scale with the NN's topology. Due to the limited amount of area in this technology, constraints were imposed and, as a result, a total of maximum 128 neurons was chosen for the fully-connected NN.

The 4 processors are fully synthesized through HLS and represent hardware digital blocks meant to accelerate the required operations. The control logic unit implements all the necessary sequential steps that are described in the behavioral code. This unit is responsible to control the behavior of the NN and it is a finite-state machine (FSM) fully inferred from the HLS. Smarty's layout is illustrated in Figure 4.12. The ANN area is 89.79 μm × 182.16 μm and the TDC bank is 20 μm × 250 μm.

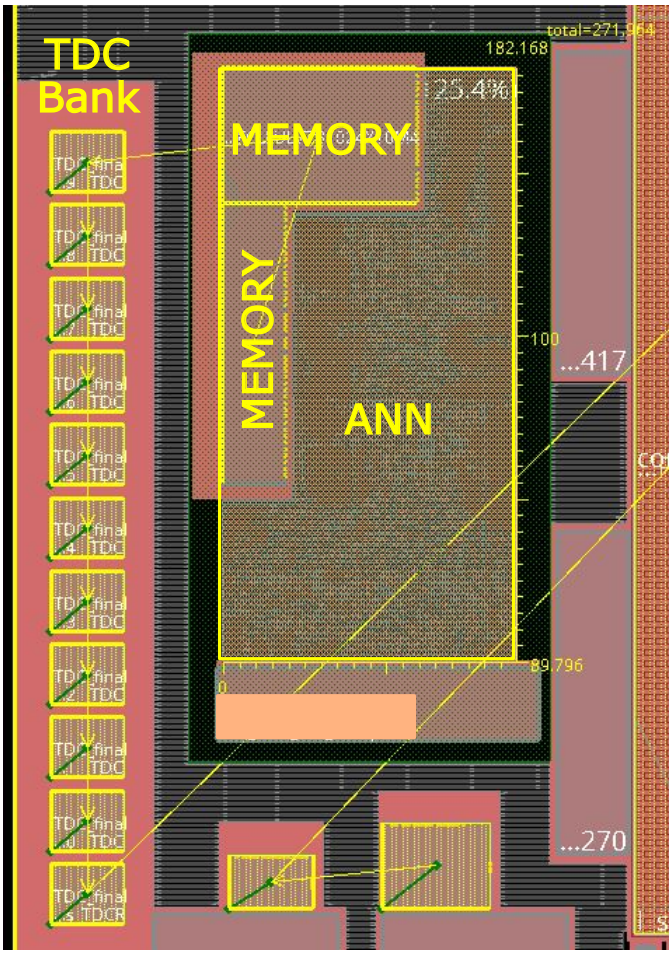


Figure 4.12: Smarty layout.

4.3.3 System

Smarty is part of a SoC design and interfaces with a RiscV processor through an AXI bus. The bus is used for configuration, control and read-out of the TDCs and NN. An address range of 1152 locations is assigned to the chip, which are then mapped to the neuron and weights RAM blocks and configuration registers. The TDC reference oscillator and debug signals have dedicated input-output pads.

The system clock is generated by the SoC PLL and can be configured through software. The NN was designed to operate at a nominal clock frequency of 500 MHz.

Smarty has its own isolated power domains:

- VDD_RING: one supply voltage assigned only to the VCOs, so that the oscillators'

frequency can be changed with limited influence to the rest of the electronics.

- VDD_ANN: a supply voltage dedicated solely to the NN. In this way, the NN can be tested independently.
- VDD_CORE: a dedicated supply voltage for all the rest of the circuits in Smarty.

4.4 Neural network performance characterization

4.4.1 Simulation setup

The performance of the NN was tested by using synthetic data generated with the Geant4 platform [140]. Geant4 is a platform that is used to simulate the particles passing through matter by using Monte Carlo simulations. It is heavily used in many different research areas such as space and medical applications, radiation effects in microelectronics, nuclear physics and PET [141]. The simulation model emulates the behavior of the gamma interaction inside a monolithic 20 mm × 4 mm × 4 mm LYSO scintillator. The ^{22}Na source is a sphere of 3 mm diameter with an intensity of 3.7 MBq. The source case is a disk of 25 mm diameter and 6 mm thickness. The gamma photons interact with the scintillator and a burst of visible photons is produced. In this case, the resulting detection times at the reading surface (the surface covered by the SiPMs) of the crystal are of main interest because they are recorded by the TDCs, and then further processed by the NN.

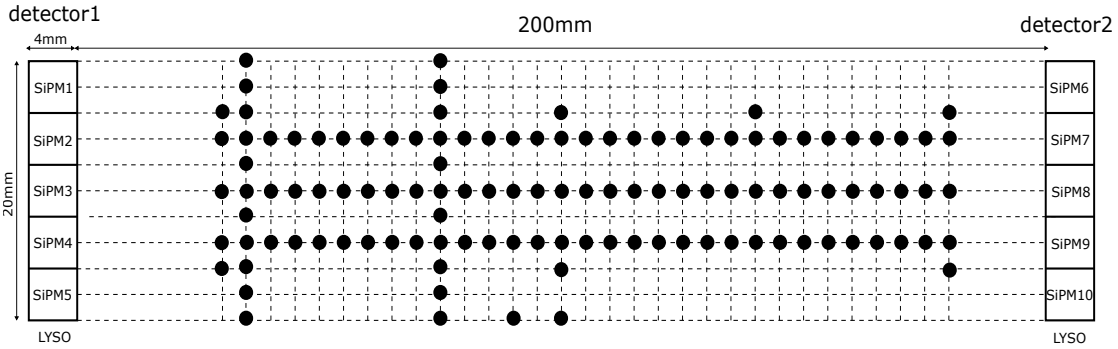


Figure 4.13: Simulation setup: two 20 mm × 4 mm × 4 mm LYSO crystals are placed in coincidence. Each detector comprises five SiPMs of 4 mm × 4 mm which are placed along the 20 mm surface. The distance between the two crystals is 200 mm. The black dots represent the radioactive source positions that are simulated in the Geant4 environment one-at-a-time.

The simulation setup is presented in Figure 4.13. A 200 mm distance between the two detectors was chosen for the Geant4 simulations in order to cover a larger area suitable

for a TDC's LSB value of 27 ps (measured reference TDC LSB). The LSB corresponds to a distance of 8.1 mm. The timestamps datasets that are used for training are only the ones that reach the crystal surface, everything else being discarded because it cannot be detected by the SiPMs. At the detection surface, the two monolithic LYSO scintillators are covered by a group of five A-SiPMs.

The timestamps collected by all the SiPMs for different source positions in space as presented in Figure 4.13 are recorded. Each source position has its own dataset which contains information about the space coordinates of the source (X_{source}, Y_{source}), the time of arrival of the photons at the reading output surface and the ID of the crystal on where these photons arrived (detector 1 or detector 2 as presented in 4.13). Considering that the two crystals are placed in coincidence, and that the source is placed in the plane of the detector pair, each detector has an identification ID of detector 1 or detector 2 so that the spatial information regarding on which detector the gamma arrived is known.

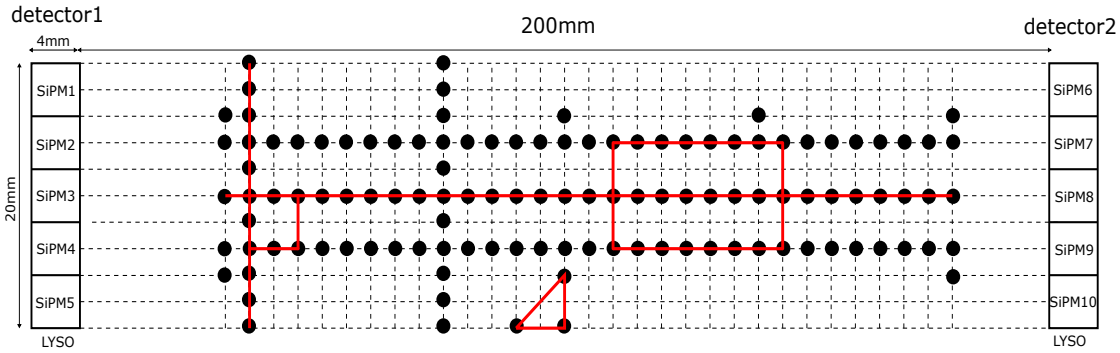


Figure 4.14: Simulation setup: Example of different source positions patterns that are used for the NN analysis.

The source position placements were chosen in such a way that different patterns such as triangles, rectangles, squares and lines can be used to test the training and evaluate the NN as presented in Figure 4.14. A large crystal of 4 mm × 4 mm × 20 mm was chosen for a greater light yield, as well as a larger field of view (FoV) coverage necessary in order to reconstruct the source position. This setup allows for easy source position reconstruction while conforming to the stringent requirements of the hardware implementation.

4.4.2 Neural network training

The NN training has been implemented in Python 3.6 by making use of the PyTorch open source machine learning framework. The raw data obtained from Geant4 simu-

lation framework was organized for training as follows:

- All the timestamps are sorted and all their corresponding source position coordinates, X_{source}, Y_{source} are kept.
- The data is organized in exposure frames of 100 ns in order to increase the number of frames available for the training process.
- A maximum of 10 timestamps are kept for each frame considering that the maximum number of A-SiPMs available in the configuration is 10. However, there might be frames with a fewer number of timestamps. The first timestamp for each SiPM in one frame is kept.
- The corresponding SiPM number from the system is assigned to each timestamp.
- The dataset is then organized in a training set (80% of the entire dataset) and a validation set (20% of the entire dataset).

The neural network training was performed with the aid of a genetic algorithm. Widely used for optimization and complex search spaces, genetic algorithms rely on processes related to evolution and natural selection. Originally formulated by Charles Darwin, genetic algorithms present different parameters, such as population, chromosomes, individuals, mutation, crossover, generation. Each of these parameters are described in the following paragraphs [142]–[144].

The GAs are non-mathematically guided algorithms and their optima evolves from one generation to another without mathematical formulation. As a result, they are used in many different areas such as: image processing, speech recognition, sensor networks, healthcare and machine learning [142], [145]–[148]. The main steps of a traditional genetic algorithm flow are depicted in Figure 4.15.

The process starts with a set of n individuals which are part of a population. The population size depends on the nature of the problem that needs to be solved and it is decided by the user. There are different studies which analyzed the impact of the population size on the genetic algorithm. Some argue that a very small population might lead to poor solutions, while a larger population size needs more computation time to find the desired solution [150]–[152]. However, there is no optimum number concerning the population size, all being based on the approach of the user in what concerns the problem solving.

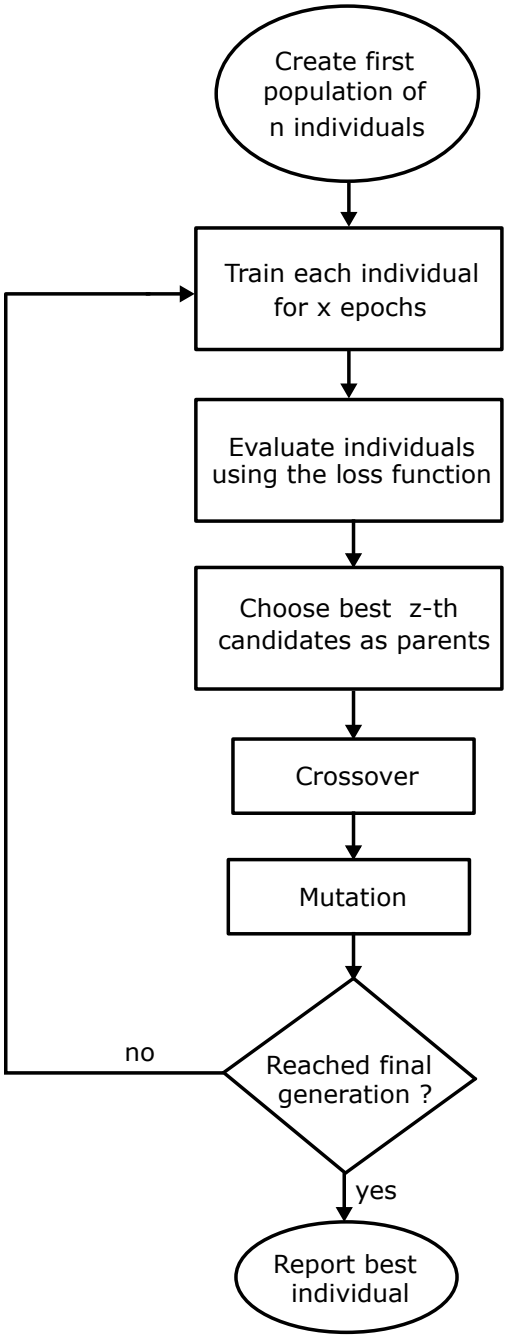


Figure 4.15: Main steps of a genetic algorithm. Adapted from [149].

The evolution usually starts from a randomly generated population with a certain number of individuals. Each solution from each individual is characterized by a set of variables known as genes which are part of the chromosomes and represent the solution to the problem that needs to be solved. Each individual in the population is characterized by a unique chromosome. Starting an iterative process, the loss functions of each individual in the present population are evaluated during the training and represent the performance quality of each individual in the population. Those individuals which present the best loss values (the parents) from the current population are selected and their chromosomes are recombined by using crossover and mutation parameters.

The main goal of the selection process is that the better an individual is, the higher are the chances of being a parent [142]. There are different strategies which can be used in order to determine the parents selection, which should be applied in line with the problem in question [153], [154]. The result of the mating process or the so-called reproduction process is a number of offsprings which serve as members of the next evolved generation. Two main important genetic operators that represent the fundamental basis of the GAs are the crossover and mutation. These parameters have a direct impact on the quality of the solution, so that extensive analysis is required in order to find the desired values.

The crossover represents a stochastic approach of recombining the chromosomes of two parents to generate a new offspring. The mostly used crossover types are illustrated in Figure 4.16, however, there are other evolving crossover techniques that can be implemented [142], [155].

One-point crossover is a type of crossover in which a random crossover point is selected along the parents' chromosomes. Starting from that point, the genes are exchanged between the parents in order to create two children for the next generation. The first parent transfers its genes to the second child, and the other way around for the other one. Two-point crossover is similar to the one-point crossover, the difference being that two crossover points are selected along the parents' chromosomes. The gene exchange takes places between the two crossover points for the production of the two children. The same exchange principle is applied as in the case of the one-point crossover. However, the exchange takes place this time only between the two crossover points. In the case of the uniform crossover, each gene is chosen from either parent with a certain probability in order to be transferred to one of the two children. As a result, one of the children will inherit more genetic information from one of the parents than the other.

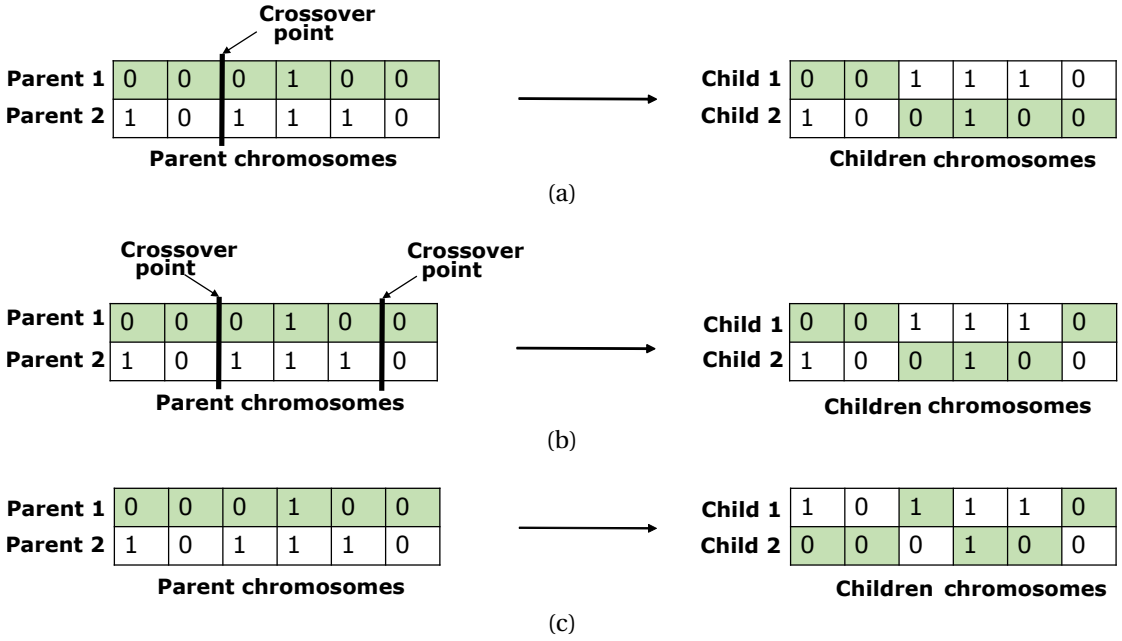


Figure 4.16: a) One-point crossover: on the right of the crossover point, the genes are interchanged between the two parent chromosomes resulting in two children who carry the genetic information from the parents. b) Two-point crossover: two crossover points are randomly selected for the parent chromosomes. The genes are interchanged between these two points. c) Uniform crossover: genes are changed randomly from the parents to their children.

Mutation is another genetic operator that, in general, takes place after the crossover occurred. The mutation represents a random change of one or more genes to produce a new offspring. The role of the mutation rate is to create new solutions so that the algorithm converges to better solutions by skipping the local optima. Solely using the crossover to produce offsprings, the GA can easily get stuck in the local optima, thus, mutation is essential in order to assure population diversity [142].

After the mutation has taken place, the GA can end by following different termination conditions which are chosen by the user, such as: a specific number of generations, the desired loss value was reached, or no improvement in the best loss value was found [142].

In the following, the GA with the chosen parameters that was implemented for this application will be described in detail.

Starting from the previously described training dataset which contains all the timestamps with their corresponding SiPM number, detector ID and source position, orga-

nized in frames the GA acts as follows:

- The GA starts with a population of 20 individuals. In this case, each individual is a neural network with a fixed topology that will be presented along with the simulation results.
- Each NN is trained by using the Adam optimizer [156], a certain number of epochs (it was changed across different trainings), and a variable learning rate which decreases from 0.01 in the first epoch to 0.001 in the last epoch.
- The algorithm is run for a certain number of generations decided by the user.
- From one generation to another, the genetic operators of crossover and mutation are applied. An explanatory picture of the previously described architecture can be seen in Figure 4.17.
- Each frame has a maximum of 10 timestamps that correspond to a specific SiPM in the system, a specific detector and source position in the plane as shown in Figure 4.13. Each individual (NN) is trained by using the aforementioned described method.
- The value of the loss is calculated for each individual for each frame in each generation. The loss function is the distance value between the estimated source position and the actual source position in space as represented in Figure 4.18 and it is calculated as follows:

$$loss = ||P_1P_2|| = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} \tag{4.3}$$

- The average loss value of each individual is calculated for each generation.
- At the end of the algorithm, the individual with the best average loss value from the last generation is reported.

In this case, the GA is used to train a given neural network with a fixed topology across all generations and individuals and only acts on the weights and biases. The crossover and mutation change the weights and biases in order to minimize the loss function. However, GAs can be given more freedom and allowed to modify the structure of the neural network as well, such as changing the number of layers or neurons in order to find the best topology for a given problem. This technique has not been applied in this work.

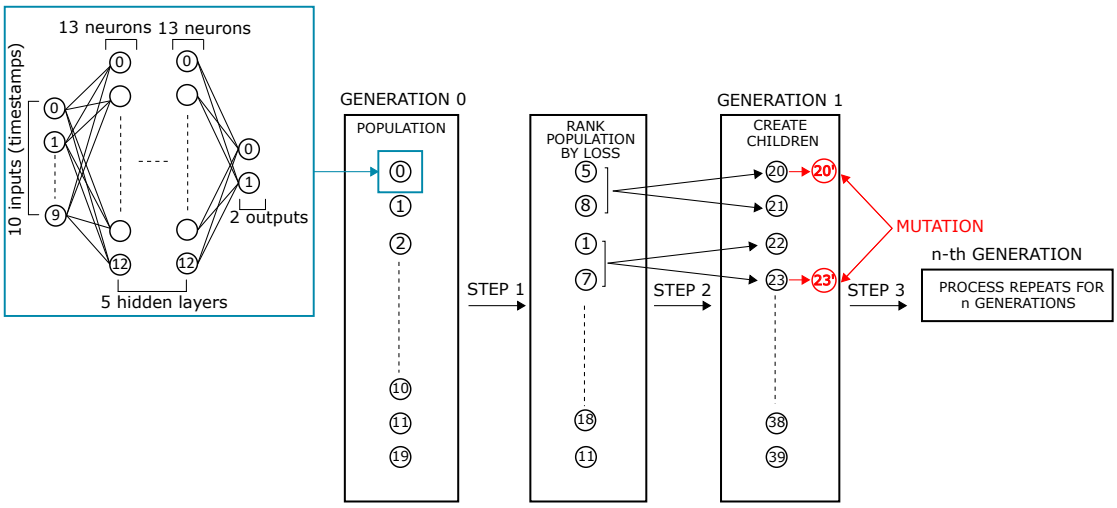


Figure 4.17: A NN example with 10 neurons in the input layer, 13 neurons/hidden layer, 5 hidden layers and 2 output neurons was chosen. Each circle in the generation is a NN as represented in the blue squares. Each generation has 20 individuals. In each generation, each NN is ranked by the loss value. The chosen parents recombine to create the children of the next generation. Randomly a mutation occurs in the chromosomes of some individuals in the population as represented by the red circles.

4.4.3 Training results with genetic algorithm

In order to confirm the NN training performance, initially, datasets obtained from Geant4 simulations were used. A certain number of source positions along with their corresponding timestamps were recorded and used for the training algorithm as previously described. In the following, an analysis based on different training scenarios is presented.

The NN topology can be tuned by the user by taking into consideration its upper limit constraints: 128 - maximum number of neurons, 1024 - maximum number of

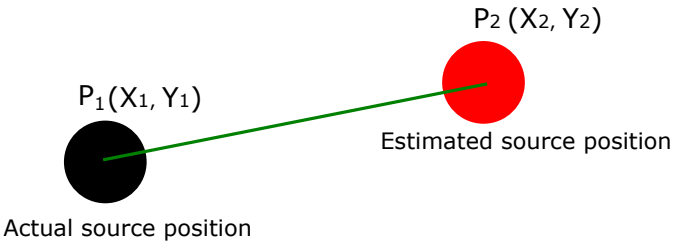


Figure 4.18: Measurement performed by the neural network. The black point corresponds to the actual source position, while the red point represents the estimate given by the NN. The loss value is the linear distance between the two points.

weights and biases, 16 - maximum number of hidden layers. This is one of the important features of the NN. As a consequence, the NN's performance was analyzed by changing its topology. Two main topologies of different sizes were analyzed: narrow-deep and wide-shallow as depicted in Figure 4.19a and Figure 4.19b respectively.

In the following, results from the training of the two NN topologies with different mutation rates are presented. The NNs were trained for 6 different source positions along the X axis at the -70, -50, -10, 40, 60 and 80 mm positions, with 0 mm denoting the midpoint between the two detectors. The GA was used with 30 generations with 20 individuals per generation and 30000 epochs each. The learning rate is variable, changing from 0.01 to 0.001 from the first to the last epoch respectively. The crossover rate is 50%. The mutation rate was changed during the training for both topologies and the final loss was reported. This represents the best loss from the final generation. The effects of the mutation rate are depicted for both NN topologies in Figure 4.20.

The NN's performance results are illustrated in Figure 4.21 and Figure 4.22. The narrow-deep topology has a much better performance compared to the wide-shallow approach, as can be observed from the absolute error on the X coordinate estimation of the NN with a blind validation set input.

For the remainder of this analysis, only narrow-deep NNs with mutation rates less than 2% will be considered.

4.4.4 Measurements and evaluation

TDCs' performance

The TDCs represent the interface between the photodetectors and the ANN, therefore, their functionality is crucial for the performance of the entire system. All ten TDCs were tested electrically and their transfer functions are presented in Figure 4.23.

In order to measure the TDC transfer functions, the TDC_STOP_ELECTRIC and TDC_START_ALL signals were sent by the FPGA board with different phases with respect to each other so that a larger range of impulse widths could be covered. In order to determine the LSB of each TDC, the oscillation period of the 7th counter bit of each TDC was measured with the oscilloscope. As it can be seen from Table 4.2, the middle TDCs (4 to 6) present a slightly different LSB due to the variations in the output load caused by unequal distances between the TDCs and the processing electronics.

The nonlinearities of TDC0 are depicted in Figure 4.24. TDC0 exhibits a DNL of -

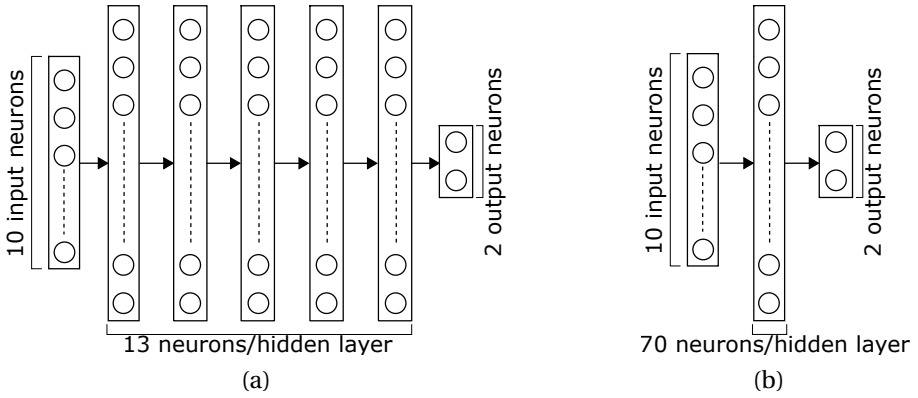


Figure 4.19: NN topologies: a) Narrow-deep fully-connected NN consisting of 10 input neurons, 5 hidden layers of 13 neurons each, and 2 output neurons; b) Wide-shallow fully-connected NN consisting of 10 input neurons, 1 hidden layer of 70 neurons and 2 output neurons.

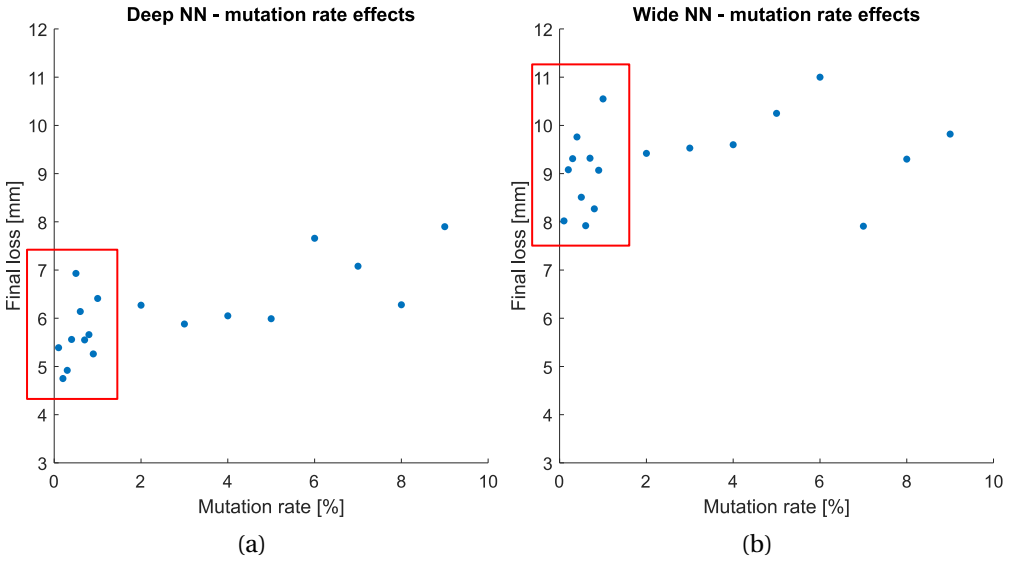


Figure 4.20: Mutation rate effects of a) Narrow-deep NN and b) Wide-shallow NN. Considering the final loss values, the narrow-deep NN topology clearly performs better. In both cases, a) and b) the best losses are obtained for mutation rates less than 2% as indicated in the red caption.

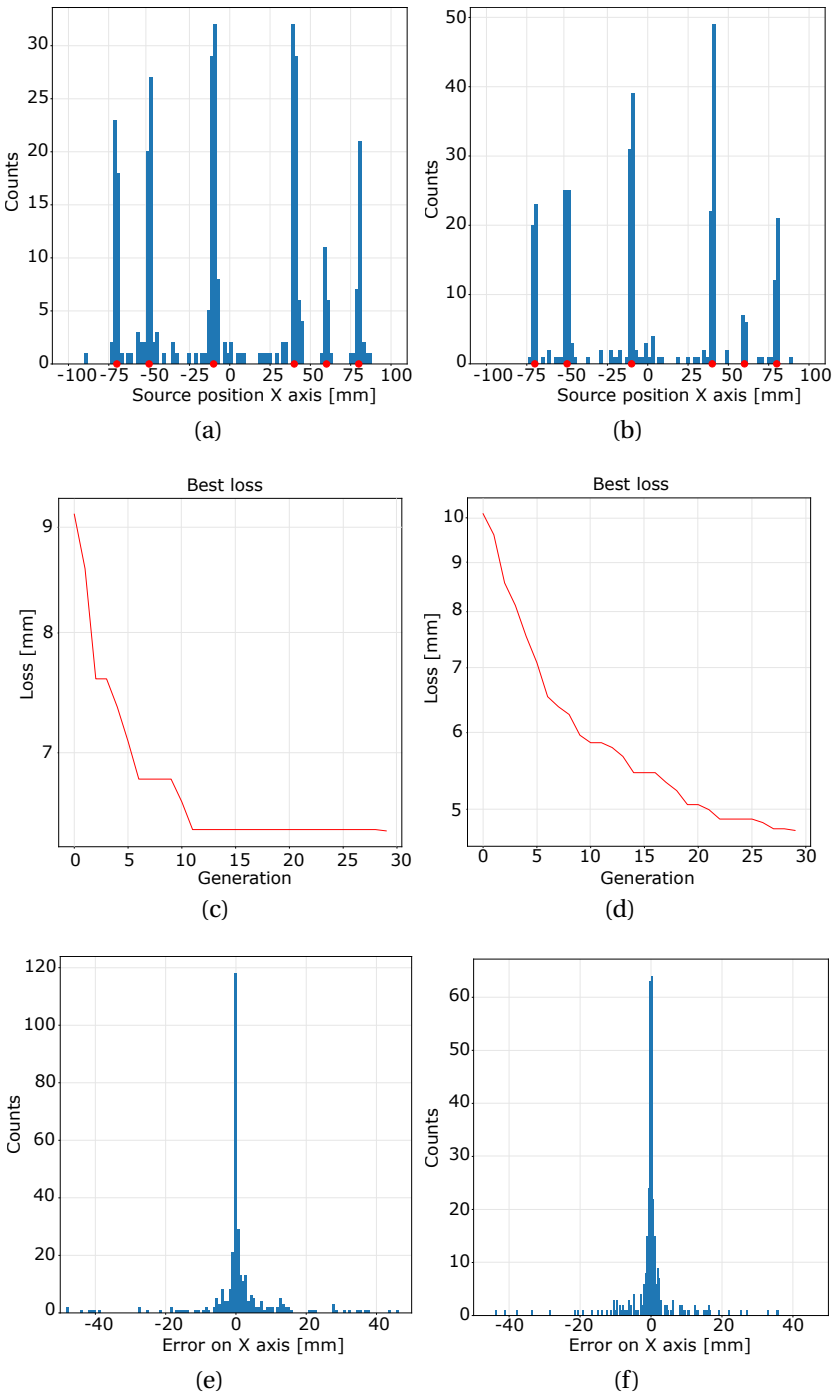


Figure 4.21: Narrow-deep fully-connected NNs. a), b) Histogram of the X coordinate estimation at the output of the neural network when presented with never-before-seen validation input frames for 6 radioactive source positions. Ground truth is shown with red dots. c), d) The average loss of the best performing individual in each generation of the GA. e), f) The absolute error of the X coordinate estimation at the output of the neural network when presented with the never-before-seen validation input frames. a), c), e) Mutation rate of 1 %, b), d), f) Mutation rate of 0.2 %.

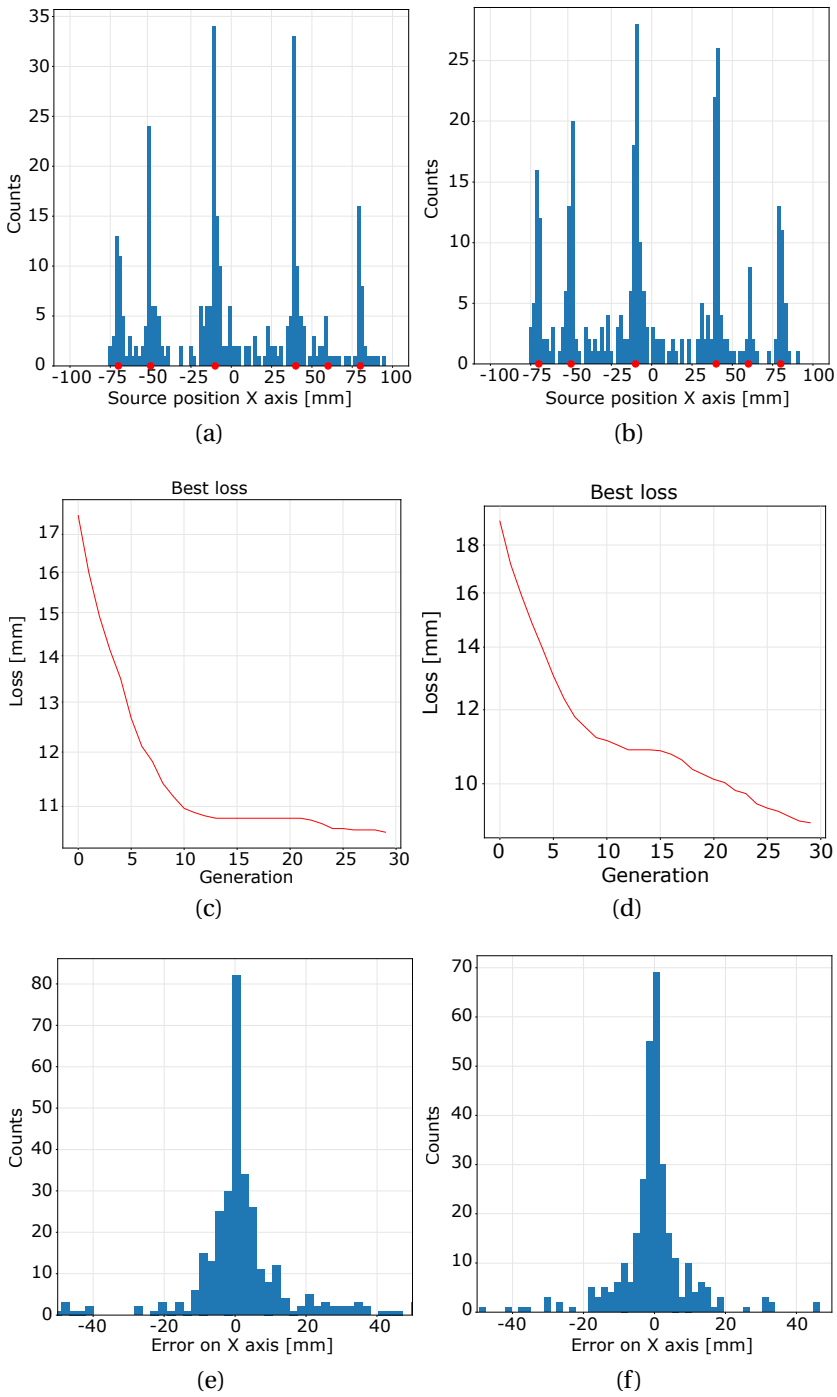


Figure 4.22: Wide-shallow fully-connected NNs. a), b) Histogram of the X coordinate estimation at the output of the neural network when presented with never-before-seen validation input frames for 6 radioactive source positions. Ground truth is shown with red dots. c), d) The average loss of the best performing individual in each generation of the GA. e), f) The absolute error of the X coordinate estimation at the output of the neural network when presented with the never-before-seen validation input frames. a), c), e) Mutation rate of 1 %, b), d), f) Mutation rate of 0.2 %.

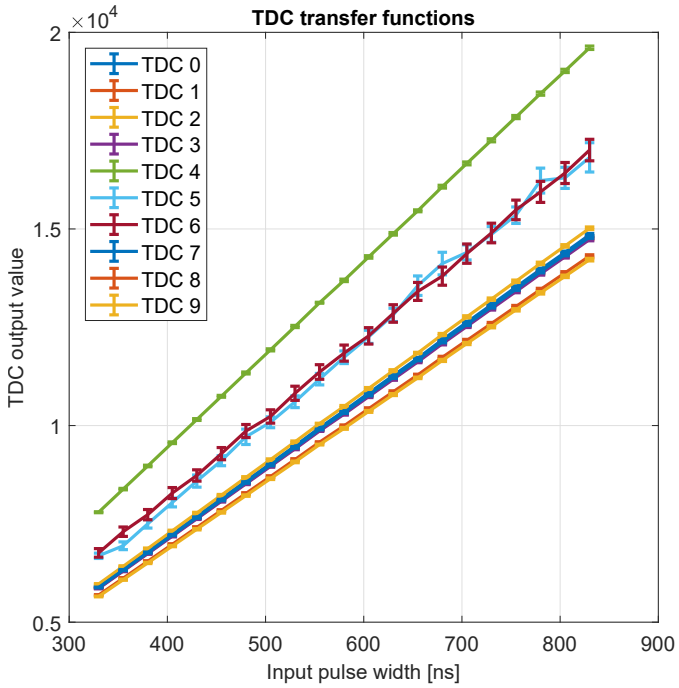


Figure 4.23: Transfer functions for each of Smarty's 10 TDCs, measured over a range of 800 ns. Results obtained through electrical measurements.

0.19/0.15 LSB and an INL of -0.77/0.9 LSB. The DNL and INL values of all ten TDCs are shown in Table 4.3.

On-chip neural network performance

In order to test the functionality of the entire system, optical measurements were performed with the Smarty board. First, a dedicated support was 3D-printed for Smarty so that it could be attached to an optical table and kept in a stable position. In addition, a scintillator could be attached to the analog silicon photomultipliers as presented in Figure 4.25. The first optical measurement setup is illustrated in Figure 4.26.

In order to test the entire chain's functionality (A-SiPMs' board, Smarty board, control board and FPGA) an optical measurement was performed by illuminating all the five A-SiPMs with a 375 nm PiL037-FC laser and performing single-shot measurements. Each A-SiPM board contains five independent channels which comprises the S14160/S14161 series Hamamatsu A-SiPMs that feature high detection efficiency and low operation voltage for PET photodetectors, a 100 × amplification stage and a fast comparator with adjustable threshold voltage. This Hamamatsu series is available in

Table 4.2: TDC LSBs measured at nominal power supply of 0.8 V.

TDC	LSB [ps]
TDC 0	55.9
TDC 1	56.2
TDC 2	55.3
TDC 3	56.4
TDC 4	42.3
TDC 5	47.6
TDC 6	48.9
TDC 7	56.1
TDC 8	58
TDC 9	58.4

Table 4.3: DNL and INL values of all ten TDCs present in Smarty, @ 0.8 V.

TDC	DNL [LSB]	INL [LSB]
TDC 0	-0.19/0.15	-0.77/0.90
TDC 1	-0.11/0.16	-0.15/1.18
TDC 2	-0.29/0.29	-1.02/2.17
TDC 3	-0.12/0.13	-1.13/0.37
TDC 4	-0.45/0.55	-1.99/0.87
TDC 5	-0.34/0.36	-1.04/1.66
TDC 6	-0.43/0.39	-1.42/1.14
TDC 7	-0.14/0.15	-0.52/0.90
TDC 8	-0.22/0.23	-0.51/0.46
TDC 9	-0.13/0.12	-0.63/0.25

different sizes, the ones chosen for this design have a 4 mm × 4 mm photosensitive area and a single channel. The pixel pitch is 50 μm with a total number of 6331 pixels/channel. The A-SiPM features 50% PDE @ 450 nm for an excess bias of 2.7 V [157]. In all the experiments, the A-SiPMs were operated at the recommended operating voltage of 40.7 V (@ 2.7 V excess bias).

The single shot measurements consisted in using a START signal generated by the FPGA as a trigger for the laser controller and measuring the arrival of the SiPM output pulse with respect to a STOP signal also generated by the FPGA as illustrated in Figure 4.27. The delay between the START and STOP signals was adjusted by the user in order to test multiple points from the TDC range. The FPGA acted as a master and the laser controller was the slave. All 5 TDC output codes, each corresponding to one SiPM channel were read by Smarty.

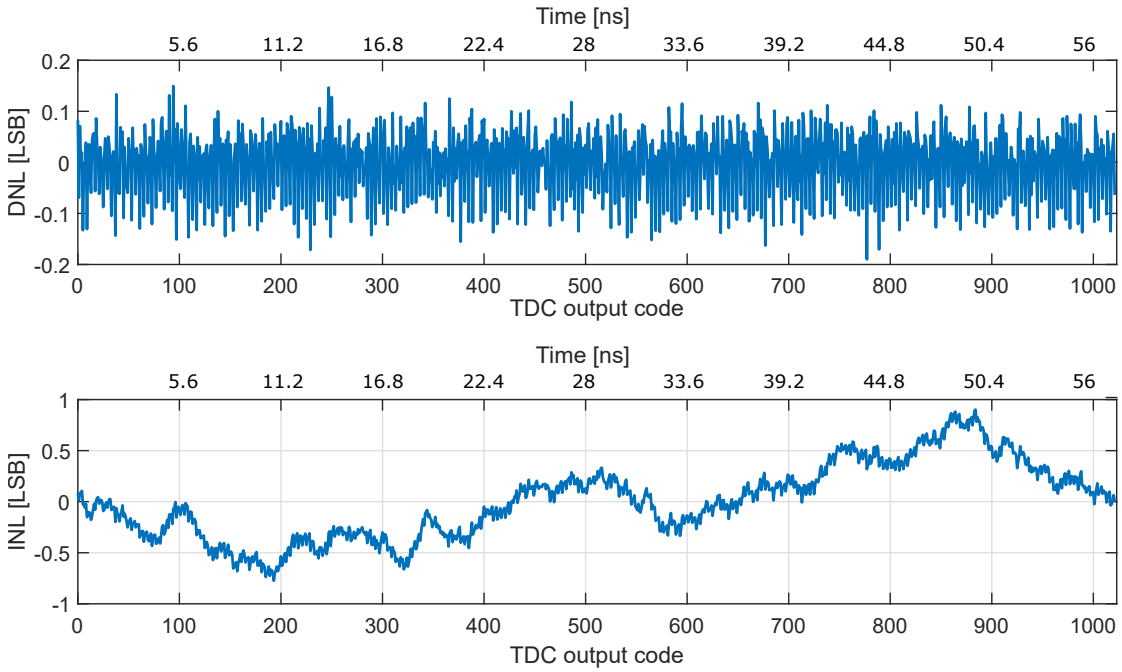
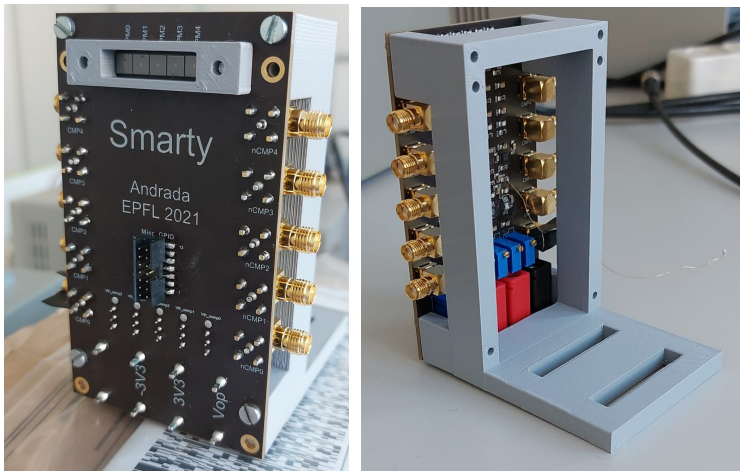


Figure 4.24: DNL and INL of TDC0.

This measurement was performed so that the ANN could be trained to identify the distance between the START (laser trigger) and STOP signals based on the TDCs' values read out from all the A-SiPMs during the single-shot measurement. In order to gather sufficient data for the training, 10000 frames were accumulated for each single-shot measurement. All the data was then transferred to the aforementioned Python training flow, this time, the difference being that the ANN was for the first time trained with measured, rather than simulated data. All 5 TDC output values were used as inputs for the ANN.

Considering the analysis presented before, the chosen topology for the ANN for training was narrow-deep. An ANN with 5 inputs, 5 hidden layers with 13 neurons each and one output was selected for training. The genetic algorithm was used again with the following parameters: 10 generations, 30 individuals per generation, uniform crossover, 0.2 % mutation rate and 10000 epochs. The inputs of the ANN are given by the measured TDCs' values while the output of the ANN is given by the distance between the START and STOP signal (called enable pulse width *EN width*) which is set from the GUI interface as presented in Figure 4.28. During measurements, the user can set the EN width from the GUI interface in steps of 5 ns clock cycles.

The results obtained during the training session are presented in Figure 4.29. As it can be observed, the ANN has been able to identify with very good precision the EN width



(a) (b)

Figure 4.25: Smarty 3D printed supports. a) Scintillator support that frames the 5 analog silicon photomultipliers. b) Entire 3D-printed frame.

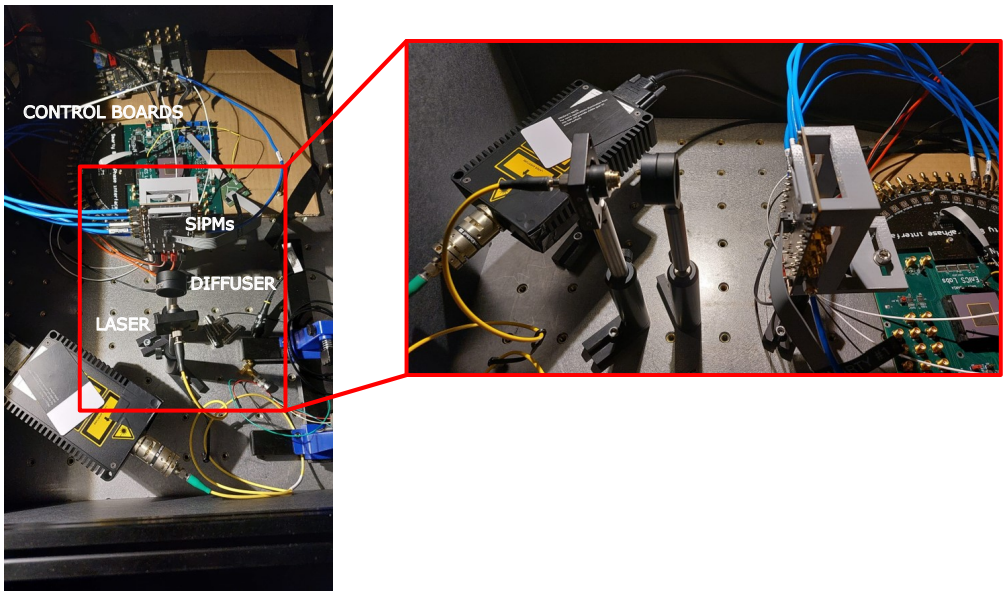


Figure 4.26: Optical measurement setup. Five A-SiPMs are illuminated with a 375 nm picosecond laser. A diffuser is placed in front of the laser.

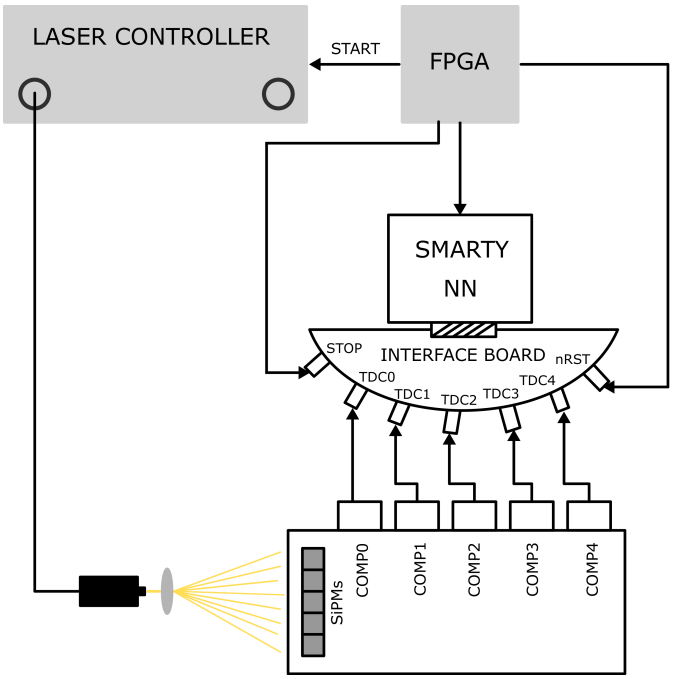


Figure 4.27: Optical setup for single-shot measurements. START signal acts as the laser trigger. START and STOP signals are generated by the FPGA. nRST signal for the TDCs is generated by the FPGA. Five A-SiPMs are illuminated with a 375 nm picosecond laser. A DG20-220-MD ThorLabs diffuser is placed in front of the laser.

values set by the user by utilizing solely the TDCs' output codes. In a conventional single-shot post-processing analysis, the results are plotted using Matlab and the histogram's peak indicates the desired value. In this case, this analysis has been done solely by the ANN without any pre-processing.

The weights and biases have been obtained by training the neural network in Python. The goal here is to demonstrate that the ANN implemented on Smarty is working on-chip, therefore, the coefficients were saved and converted to fixed point values and uploaded to Smarty. The training has been performed by using weights and biases in floating point, however, as previously described, floating point cannot be used on-chip and the conversion to fixed point is necessary, which results in loss of precision. The coefficients on Smarty are implemented with 2 signed integer and 8 fractional bits. A naive quantization of the coefficients to fixed point obtained solely by multiplying them with 2^8 and converting to integer is not sufficient. As depicted in Figure 4.30 the naive conversion is far from the desired values (yellow line). We can also observe that three additional never before seen points have been introduced, 28, 33, 47 and successfully interpolated by the neural network.

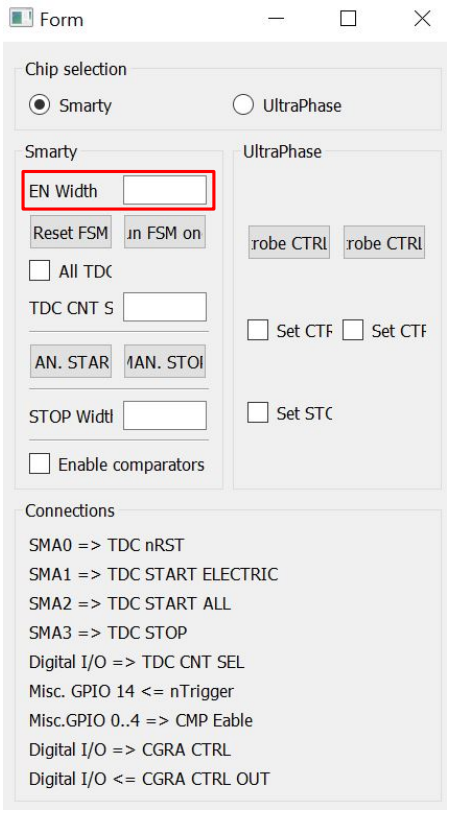


Figure 4.28: GUI for the Smarty board. The distance between the START and STOP signals in terms of clock cycles can be set in the EN width block.

A different quantization method that scales the weights and biases in order to mitigate the effects of truncation while preserving the output proves better. The motivation behind it comes from the range of the initial coefficients. The fixed point representation is limited to 2 integer bits and in some cases, the weights exceed this range. The new quantization method guarantees that the weights and biases stay in the fixed point range. The method multiplies all the weights of layer i with α_i and all the biases with β_i. As long as α_i and β_i follow:

$$\begin{aligned} \beta_1 &= \alpha_1 \\ \beta_i &= \alpha_i \times \beta_{i-1}, \qquad i > 1 \end{aligned} \tag{4.4}$$

the output will be scaled by a factor γ:

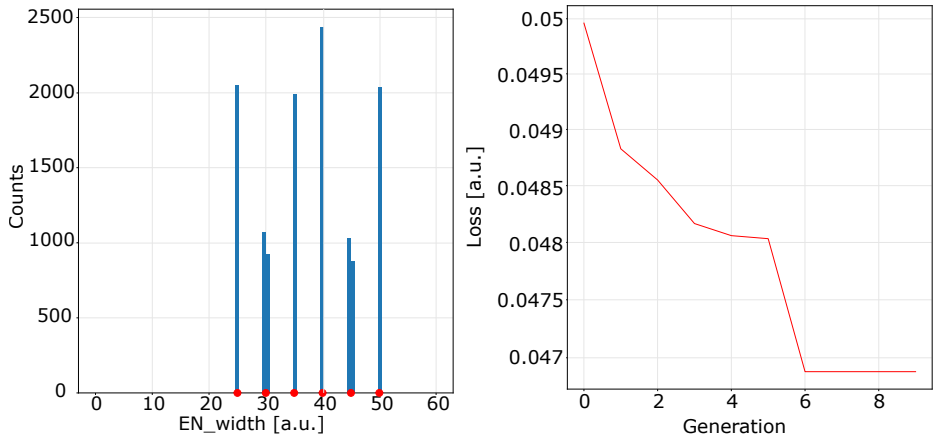


Figure 4.29: a) Histogram of the EN width estimation at the output of the neural network when presented with blind validation input frames for 6 different values of EN width from the single-shot optical measurements. Ground truth is shown with red dots. b) The average loss of the best performing individual in each generation of the GA. The EN width is presented in arbitrary units and it represents the number of clock cycles.

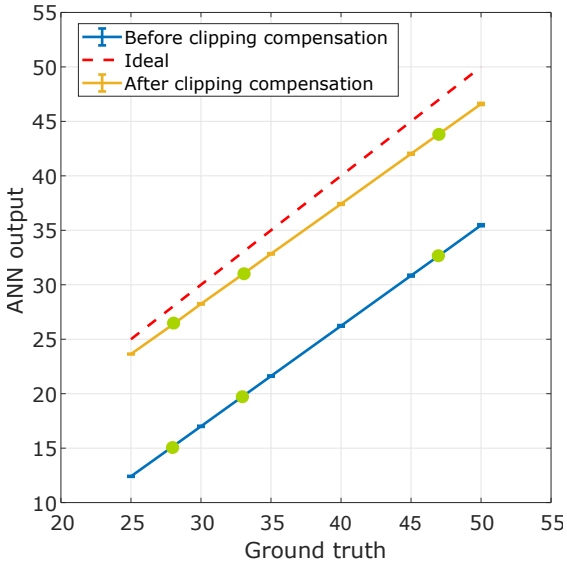


Figure 4.30: Coefficient quantization effect on the ANN's output. Two different quantization methods were used: a naive quantization method which multiplies the coefficients by 2^8 (blue) and a clipping method in which the coefficients are clipped within the desired range (yellow). (28, 33, 47) are never- before-seen points by the neural network.

$$\frac{1}{\gamma} = \prod_{i=1}^{n} \alpha_i, \tag{4.5}$$

where n is the number of layers.

Execution time

The execution time of a neural network with 10 input neurons, 5 hidden layers with 13 neurons per hidden layer and 3 output neurons is 22.44 μs when running at a 105 MHz clock. Considering a total of 1710 operations, the NN executes 76.15 MOPS. The NN was designed to run at a maximum frequency of 500 MHz which results in a 363 MOPS maximum performance.

Power consumption and area

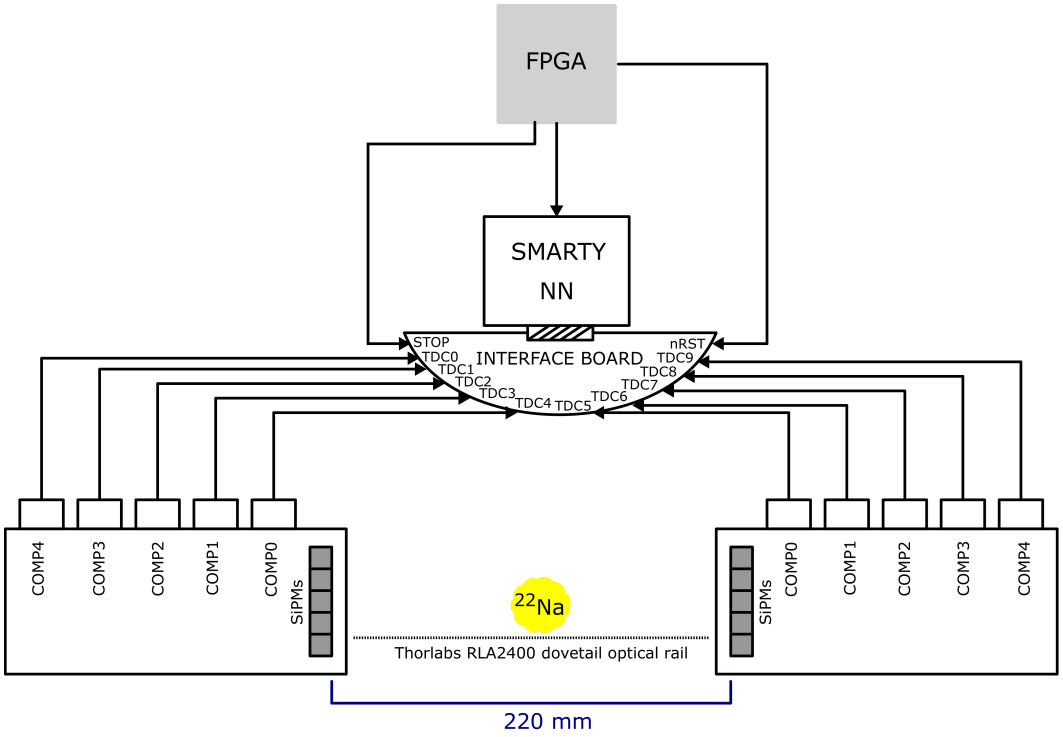
The neural network itself consumes 0.4 mW @ 100 MHz, which is equivalent to 190 GOPS/W.

4.4.5 Source position reconstruction

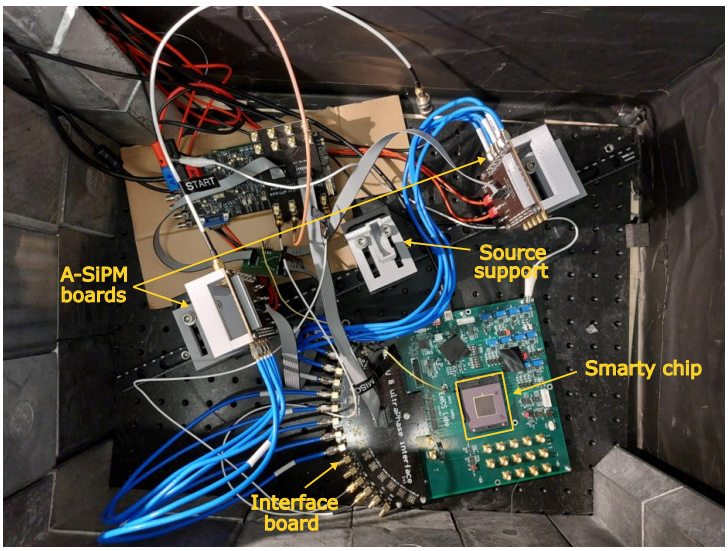
The measurement setup is shown in Figure 4.31. It comprises two A-SiPM boards as presented in Figure 4.25 placed in coincidence at a distance of 220 mm from each other with a ^{22}Na radioactive source in between. The 10 comparator outputs from the two boards are directly connected to the Smarty board via equal length coaxial cables and serve as inputs for the 10 TDCs from the chip. The ^{22}Na source is attached to a Thorlabs RLA2400 dovetail optical rail and can be moved along the axis between the two SiPM boards to precise positions.

During normal operation, a FPGA is used to create a 600 ns exposure window that begins with a reset of the TDCs and ends with a STOP signal. The arrival of the first photon during the exposure window on any of the channels will trigger the corresponding TDC. With the arrival of the STOP signal, the acquired timestamps were processed by the ANN implemented in Smarty and the results are read out, at which point a new exposure cycle begins.

In order to perform the training of the ANN, thousands of frames are accumulated with the TDCs at each source position along the X axis (same principle presented in the *Simulation setup* section). In the end, 14000 frames are kept for each source posi-



(a)



(b)

Figure 4.31: Smarty measurement setup a) diagram and b) photo. Two A-SiPM boards are placed in coincidence at a distance of 220 mm from each other. Each board is coupled with a LYSO scintillator of 4 mm × 4 mm × 20 mm. A ^{22}Na source can be moved along the X axis between the two sensor boards on a dovetail optical rail. An interface board is used between the Smarty and the sensor boards. The comparator outputs from all 10 A-SiPMs represent the TDC inputs and are connected to Smarty via equal length coaxial cables.

tion in order to have an equal number of frames for each point and used as training, validation and test sets for the NN. The ANN training is performed in Python and the weights and biases are extracted at the end of the training. The same previously described methods were used for the training, including the genetic algorithm, however, this time, a classification approach was used as it proved to have better performance when quantized. The output of the NN was divided into a number of classes according to how many source positions were trained. In this way, the NN did not report the absolute value of the source position in mm, but the class it corresponds to.

The NN classification results of the two source positions at -57 mm and 65 mm along the X axis are presented in Figure 4.33 for both floating point and quantized models in the form of confusion matrices. The average accuracy and precision of the floating point model is 83.59% and 68.69%. The quantized model has an accuracy of 83.48% and a precision of 68.59%. The accuracy is defined as the ratio between the correctly predicted observations to the total observations, while the precision is defined as the ratio between the correctly predicted positive observations to the total predicted positive observations. In this case, a neural network with 10 input neurons (values provided by the TDCs), 5 hidden layers with 13 neurons each and 3 outputs was used. For the training process, frames in which only the TDCs corresponding to the left detector or right detector fired were considered non-valid and marked with -120 mm, a position beyond the left detector, therefore, nonsensical. A conceptual example of valid and non-valid frames is shown in Figure 4.32.

Frame	Left detector 1					Left detector 2				
0	0	0	0	0	0	125	122	140	130	120
1	0	0	125	122	0	0	0	0	0	0
2	125	122	140	130	120	170	123	272	129	111
3	0	0	125	122	0	0	0	0	229	420

Figure 4.32: Conceptual representation of valid and non-valid frames. Four different frames with TDC data are shown. The non-valid frames (frame 0 and frame 1) in which only the left detector or right detector fired are presented in red and marked with -120 for the neural network training. The last two frames present TDC data on both detectors, therefore, are considered as valid frames (frame 2 and frame 3).

The NN clearly distinguished between valid and non-valid frames with high degree of certainty. The floating point model was able to distinguish between -57 mm and 65 mm positions, however, it favors the former, most likely due to unequal number of frames in the input dataset and bias in differences between the two detectors that lead to a slight increase in the count rate on the SiPMs on the left. The effect is exacerbated when quantizing the model and results in less certainty.

The analysis was repeated for three source positions, namely -120 mm, -57 mm, 0 mm and 65 mm and the results are presented in Figure 4.34. The average accuracy and precision of the floating point model is 81.26% and 53.70%. The quantized model has an accuracy of 81.34% and a precision of 53.88%. Again, the NN was able to successfully distinguish between valid and non-valid frames with high degree of certainty. The quantized implementation performs similarly to the floating point model.

4.5 Conclusions

Smarty is a small, feed-forward, fully integrated and reconfigurable artificial neural network with 10 input TDCs designed in 16 nm FinFET technology node. This design aimed to reduce the system complexity and output data throughput when reconstructing radioactive source positions between two photodetectors placed in coincidence.

The design was validated experimentally by successfully reconstructing 3 distinct source positions along the axis between the detectors and up to 6 different source positions in the simulations. Apart from this, the neural network successfully distinguished between non-valid and valid frames during all of the experiments. This is important as it has the potential to eliminate the need for a separate coincidence window preprocessing step.

Reconfigurability is one of the features which should be considered for the on-chip implementations which do not benefit from the same flexibility as the NNs developed on FPGAs or GPUs. In the case of Smarty, the NN's topology can be changed within certain limits in order to achieve the best performance for the current problem. Various feed-forward configurations were explored both experimentally and in simulations for different neural network depths, widths, number of neurons, number of inputs and with or without classification.

The chip can execute 363 MOPS with a maximum power consumption of 190 GOPS/W. A comparison with current silicon for neural network implementations is difficult. Most modern machine learning implementations target image processing which requires the use of other types of neural networks, such as convolutional neural networks that process high resolution images. As a result, the architectures of such systems are not always scalable and are designed with high computation power in mind, in the order of TOPS [158]–[160]. The goal of Smarty was to bring a level of preprocessing as close as possible to the detector in order to reduce the amount of data that needs to be read out and the overall system complexity. The different nature of input data,

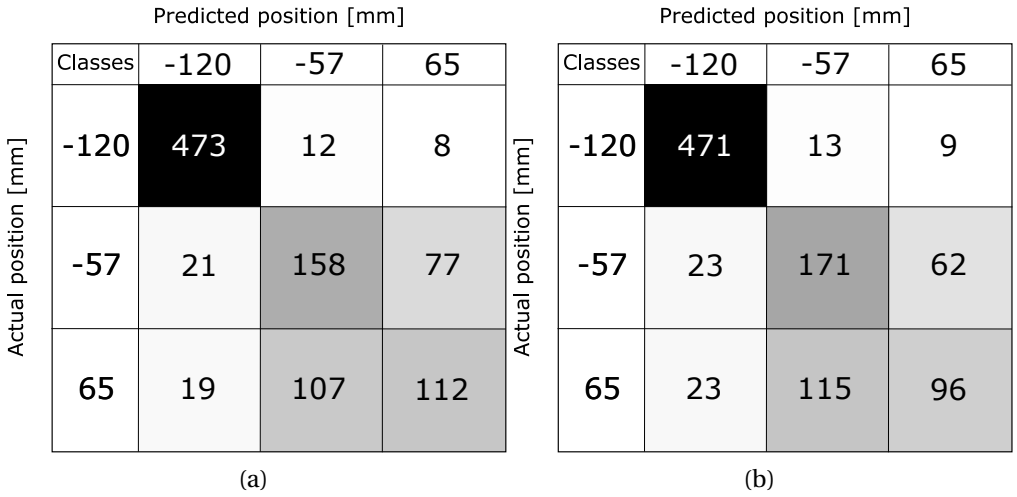


Figure 4.33: Confusion matrices of the neural network classification results for 2 different radioactive source positions in a) floating point and b) quantized implementations.

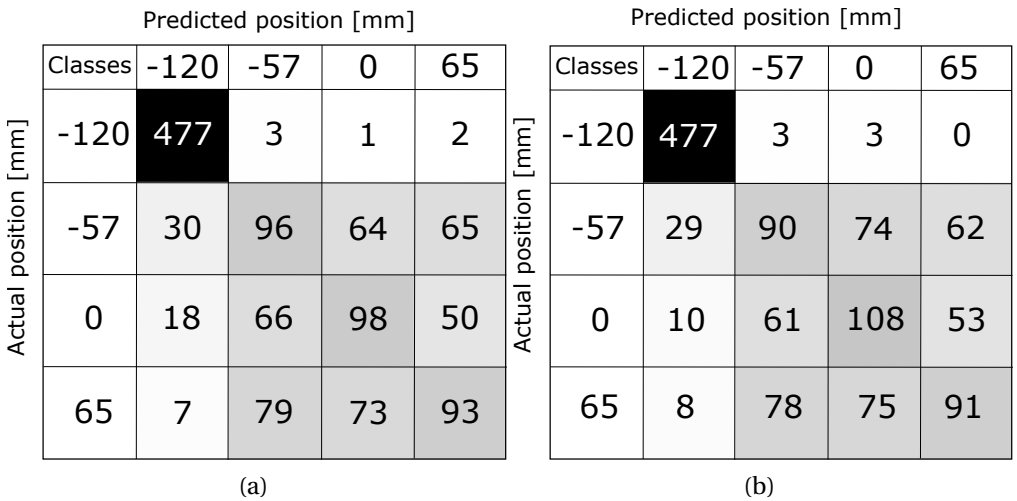


Figure 4.34: Confusion matrices of the neural network classification results for 3 different radioactive source positions in a) floating point and b) quantized implementations.

timestamps as opposed to 2D images, makes convolutional neural networks ill suited for this type of application, therefore, Smarty targeted feed-forward NNs.

5 Conclusions and future work

5.1 Conclusions

The main goal of this thesis was to develop and characterize different readout and pre-processing techniques for sensors used in time-of-flight PET. As a result, three different designs were implemented within the scope of this thesis with the main focus on full integration.

The first part of the thesis presents the design and characterization of Blumino which is the first fully-integrated analog silicon photomultiplier with on-chip discrimination and time conversion. Blumino was designed by combining a CMOS SPAD-dedicated process with standard CMOS in a 350 nm technology node. The main goal of this sensor was to create a simple, compact and fully functional design suitable as a PET photodetector, as well as to study the effects of integrating standard and custom CMOS processes together. The A-SiPM has a sensitive area of 3 mm × 3 mm and a fill factor of 75% with a PDP of 46% at 425 nm for an excess bias of 2.5 V. The A-SiPM has an unique topology which comprises a standard and a fast terminal. The standard terminal is dedicated to energy measurements, while the fast terminal is suitable for timing measurements. The fast terminal has been integrated with a discriminator and TDC, while the standard terminal was connected to external electronics. This implementation preserves the backward compatibility of the sensor. Blumino was characterized extensively through electrical and optical measurements as presented in Chapter 2 and proved to be fully functional. Two different measuring platforms were implemented, with the last implementation being a dedicated small readout board of 10 mm × 70 mm designed to improve the timing performance by bringing the amplification and comparison in close proximity to the standard terminal. The particular compact design is suitable for tiling. CTR measurements performed using the standard terminal and the specially designed small measuring platform brought

23.17% improvement to the CTR compared to the previous implementation when using $2.5 \times 2.5 \times 20$ mm^3 LYSO scintillators.

The second part of this thesis focuses on the design and implementation of Blueberry, a 3D-Stacked FSI Multichannel D-SiPM implemented in a 180 nm CMOS technology node for both top and bottom tiers. The FSI implementation optimizes the sensor for the blue part of the light spectrum which is of interest in PET applications. The top tier contains solely SPADs, while the bottom tier includes all the electronic circuits. The goal of this design was to explore the benefits of 3D integration as opposed to the 2D approach, as well as to push the limits in terms of timing resolution. Due to the 3D implementation, different functionalities could be included on the bottom tier while preserving the fill factor. The sensor is divided into two independent cores, each of them containing 64 clusters. Each cluster is made of an array of 64 SPADs with their corresponding pixel circuits, a TDC that timestamps the first detected photon, a SPAD address tree that identifies the address of the first firing SPAD, and a photon counting system, which counts the number of detected photons. The TDC is based on an antiphased multi-path gated ring oscillator architecture and it includes an on-chip error correction algorithm that reduces miscounts in the coarse TDC bits. The TDC exhibits an LSB of 15 ps. The oscillator's topology was derived from the TDC design presented in Chapter 2 and, due to the antiphased approach, it can be concluded that this implementation advanced the previous design in terms of timing resolution.

The final part of the thesis is dedicated to Smarty, a fully reconfigurable feed-forward on-chip neural network designed in 16 nm FinFET technology. The goal of this design was to provide an efficient means of on-chip data pre-processing that could serve PET applications handling large amounts of data. Smarty can accommodate feed-forward neural network topologies of up to 128 neurons with a maximum of 1024 weights and biases. The desired neural network topology can be chosen by the user based on the researched problem within these limits. The chip can execute 363 MOPS with a maximum power consumption of 190 GOPS/W. Ten TDCs were designed to be used as the neural network's inputs. The system was tested in a coincidence experimental setup and it successfully distinguished different radioactive source positions along the X axis by using only TDC timestamps. The floating point models showed its ability to distinguish up to 6 different positions, while the quantized implementation reconstructed 3 different source positions. Smarty, a simple implementation and small on-chip reconfigurable artificial feed-forward neural network, showed the viability of emission reconstruction.

5.2 Future work

In terms of future research topics related to the work presented in this thesis, one can focus on improving the overall performance of the sensors.

With updated transistor models, future iterations of Blumino could exhibit a better timing resolution. In addition, the comparator could be modified to become sensitive to single photon levels which will improve the overall timing performance of the system. In what concerns the CTR measurements performed with the integrated fast terminal and standard terminal, the small sensor readout board needs to be redesigned to function in an event-driven configuration. In this way, the system efficiency would be drastically improved.

At the time of writing this thesis, the full characterization of Blueberry is not complete. Therefore, more effort needs to be dedicated towards this task. The TDC could benefit of fine tuning of the delay cell design in order to mitigate the effect of process variations and improve the linearity.

A significant challenge in the latter stages of the thesis came from choosing the right quantization methods for the neural network model for Smarty. This can be a standalone research subject of a new thesis. Current tools are tailored towards computer science applications and are not flexible enough to be used for hardware implementations where resources are limited. Therefore, it would be very useful to develop quantized model libraries with custom data formats that can be efficiently used in training algorithms running on accelerator platforms such as GPUs.

Smarty has paved the way for radioactive source position reconstructions using a small neural network. Further research into the advantages of using multiple channels placed in various positions or the addition of energy information as an input to the neural network should be carried out.

Bibliography

[1] G. Miguel, L. Miriam Mikhail, A.-W. May, G. Francesco, P. Olivier, and P. Diana, "Addressing Global Inequities in Positron Emission Tomography-Computed Tomography (PET-CT) for Cancer Management: A Statistical Model to Guide Strategic Planning", *Medical Science Monitor: international medical journal of experimental and clinical research*, vol. 26, 2020. DOI: doi.org/10.12659/MSM.926544. [Online]. Available: https://doi.org/10.12659/MSM.926544.

[2] H. C. Verduzco-Aguirre, G. Lopes, and E. Soto-Perez-De-Celis, "Implementation of diagnostic resources for cancer in developing countries: a focus on PET/CT", *Ecancermedicalscience*, vol. 13, 2019. DOI: 10.3332/ecancer.2019.ed87. [Online]. Available: https://doi.org/10.3332/ecancer.2019.ed87.

[3] O. Gerke, R. Hermansson, S. Hess, S. Schifter, W. Vach, and P. F. Høilund-Carlsen, "Cost-effectiveness of PET and PET/computed tomography: a systematic review", *PET clinics*, vol. 10, pp. 105–124, 2015. DOI: 10.1016/j.cpet.2014.09.008. [Online]. Available: https://doi.org/10.1016/j.cpet.2014.09.008.

[4] E. A. Perini, M. Skopchenko, T. T. Hong, *et al.*, "Pre-feasibility Study for Establishing Radioisotope and Radiopharmaceutical Production Facilities in Developing Countries ", *Curr Radiopharm.*, vol. 12, pp. 187–200, 2019. DOI: 10.2174/1874471012666190328164253. [Online]. Available: https://doi.org/10.1016/j.cpet.2014.09.008.

[5] G. V. Hirsch, C. M. Bauer, and L. B. Merabet, "Using structural and functional brain imaging to uncover how the brain adapts to blindness", *Annals of Neuroscience Psychology*, pp. 2–5, 2015.

[6] T. F. Massoud and S. S. Gambhir, "Molecular imaging in living subjects: seeing fundamental biological processes in a new light", *Genesis and Development*, vol. 17, pp. 545–580, 2003. DOI: 10.1101/gad.1047403.

[7] S. Baillet, J. Mosher, and R. Leahy, "Electromagnetic brain mapping", *IEEE Signal Processing Magazine*, vol. 18, no. 6, pp. 14–30, 2001. DOI: 10.1109/79.962275.

[8] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal", *Nature*, vol. 412, pp. 150–157, 6843 2001. DOI: 10.1038/35084005.

[9] A. Ahmad, P. Nikolaos, and B. Jamshed, "18F-FDG PET/CT Imaging In Oncology", *Annals of Saudi Medicine*, 201. DOI: 10.4103/0256-4947.75771. [Online]. Available: https://doi.org/10.4103/0256-4947.75771.

[10] D. L. Bailey, D. W. Townsend, and P. E. Valk, *Positron Emission Tomography*. Springer London, 2005. DOI: https://doi.org/10.1007/b136169.

[11] M. E. Phelps, *PET Physics, Instrumentation and Scanners*. Springer Science+Business Media, LLC, 2006.

[12] D. R. Schaart, "Physics and technology of time-of-flight PET detectors", *Physics in Medicine and Biology*, vol. 66, 9 Apr. 2021. DOI: 10.1088/1361-6560/abee56.

[13] R. Mao, L. Zhang, and R.-Y. Zhu, "Optical and Scintillation Properties of Inorganic Scintillators in High Energy Physics", *IEEE Transactions on Nuclear Science*, vol. 55, no. 4, pp. 2425–2431, 2008. DOI: 10.1109/TNS.2008.2000776.

[14] Philips, *Vereos Digital PET/CT - Media Gallery*. [Online]. Available: https://www.usa.philips.com/healthcare/product/HC882446/vereos-digital-petct-proven-accuracy-inspires-confidence#galleryTab=CLI.

[15] S. Seifert, H. T. van Dam, R. Vinke, *et al.*, "A Comprehensive Model to Predict the Timing Resolution of SiPM-Based Scintillation Detectors: Theory and Experimental Validation", *IEEE Transactions on Nuclear Science*, vol. 59, no. 1, pp. 190–204, 2012. DOI: 10.1109/TNS.2011.2179314.

[16] J. P. Schmall, S. Surti, P. Dokhale, *et al.*, "Investigating CeBr3 for ultra-fast TOF-PET detector designs", in *2016 IEEE Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop (NSS/MIC/RTSD)*, 2016, pp. 1–4. DOI: 10.1109/NSSMIC.2016.8069517.

[17] D. Schaart, S. Seifert, R. Vinke, *et al.*, "LaBr(3):Ce and SiPMs for time-of-flight PET: achieving 100 ps coincidence resolving time", vol. 55, pp. 179–189, 2010. DOI: 10.1088/0031-9155/55/7/N02.

[18] M. V. Nemallapudi, S. Gundacker, P. Lecoq, *et al.*, "Sub-100 ps coincidence time resolution for positron emission tomography with LSO:Ce codoped with Ca", *Physics in Medicine and Biology*, vol. 60, no. 12, pp. 4635–4649, 2015. DOI: 10.1088/0031-9155/60/12/4635. [Online]. Available: https://doi.org/10.1088/0031-9155/60/12/4635.

[19] J. W. Cates and C. S. Levin, "Advances in coincidence time resolution for PET", *Physics in Medicine and Biology*, vol. 61, no. 6, pp. 2255–2264, 2016. DOI: 10.1088/0031-9155/61/6/2255. [Online]. Available: https://doi.org/10.1088/0031-9155/61/6/2255.

[20] S. Gundacker, R. M. Turtos, E. Auffray, M. Paganoni, and P. Lecoq, "High-frequency SiPM readout advances measured coincidence time resolution limits in TOF-PET", *Physics in Medicine and Biology*, vol. 64, no. 5, p. 055 012, 2019. DOI: 10.1088/1361-6560/aafd52. [Online]. Available: https://doi.org/10.1088/1361-6560/aafd52.

[21] R. I. Wiener, M. Kaul, S. Surti, and J. S. Karp, "Signal analysis for improved timing resolution with scintillation detectors for TOF PET imaging", in *IEEE Nuclear Science Symposuim and Medical Imaging Conference*, 2010, pp. 1991–1995. DOI: 10.1109/NSSMIC.2010.5874124.

[22] S. Pourashraf, A. Gonzalez-Montoro, J. Y. Won, *et al.*, "Scalable electronic readout design for a 100 ps coincidence time resolution TOF-PET system", *Physics in Medicine and Biology*, vol. 66, no. 8, p. 085 005, 2021. DOI: 10.1088/1361-6560/abf1bc. [Online]. Available: https://doi.org/10.1088/1361-6560/abf1bc.

[23] S. Gundacker, R. Turtos Martinez, N. Kratochwil, *et al.*, "Experimental time resolution limits of modern SiPMs and TOF-PET detectors exploring different scintillators and Cherenkov emission", *Physics in Medicine and Biology*, vol. 65, no. 2, 2020. DOI: doi.org/10.1088/1361-6560/ab63b4.

[24] C. S. Levin and E. J. Hoffman, "Calculation of positron range and its effect on the fundamental limit of positron emission tomography system spatial resolution", *Physics in medicine and biology*, vol. 44, no. 3, pp. 781–799, 1999. DOI: 10.1088/0031-9155/44/3/019.

[25] S. Vandenberghe, P. Moskal, and J. S. Karp, "State of the art in total body PET", *EJNMMI physics*, vol. 7, no. 1, 2020. DOI: 10.1186/s40658-020-00290-2.

[26] S. R. Cherry and M. Dahlbom, *The PET: Physics, Instrumentation, and Scanners Book*. Springer, 2006. DOI: 10.1007/0-387-34946-4_1.

[27] A. K. Shukla and U. Kumar, "Positron emission tomography: An overview", *Journal of Medical Physics*, vol. 1, pp. 13–21, Jan. 2006. DOI: 10.4103/0971-6203.25665.

[28] R. I. Wiener, M. Kaul, S. Surti, and J. S. Karp, "Signal analysis for improved timing resolution with scintillation detectors for TOF PET imaging", in *IEEE Nuclear Science Symposuim and Medical Imaging Conference*, 2010, pp. 1991–1995. DOI: 10.1109/NSSMIC.2010.5874124.

[29] S. Strother, M. Casey, and E. Hoffman, "Measuring PET scanner sensitivity: relating countrates to image signal-to-noise ratios using noise equivalents counts", *IEEE Transactions on Nuclear Science*, vol. 37, no. 2, pp. 783–788, 1990. DOI: 10.1109/23.106715.

[30] M. Conti, "Effect of randoms on signal-to-noise ratio in TOF PET", *IEEE Transactions on Nuclear Science*, vol. 53, no. 3, pp. 1188–1193, 2006. DOI: 10.1109/TNS.2006.875066.

[31] P. Lecoq, "Development of new scintillators for medical applications", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 809, pp. 130–139, 2016, Advances in detectors and applications for medicine, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2015.08.041. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900215009754.

[32] M. Daube-Witherspoon, S Surti, A Perkins, *et al.*, "The imaging performance of a LaBr3-based PET scanner", *Physics in Medicine and Biology*, vol. 55, pp. 45–64, 2010. DOI: 10.1088/0031-9155/55/1/004.

[33] S. Blahuta, A. Bessière, B. Viana, P. Dorenbos, and V. Ouspenski, "Evidence and Consequences of Ce4+ in LYSO:Ce, Ca and LYSO:Ce,Mg single crystals for medical imaging applcations", *IEEE Transactions on Nuclear Science*, vol. 60, no. 4, pp. 3134–3141, 2013. DOI: 10.1109/TNS.2013.2269700.

[34] S. Gundacker, E. Auffray, K. Pauwels, and P. Lecoq, "Measurement of intrinsic rise times for various L(Y)SO and LuAG scintillators with a general study of prompt photons to achieve 10 ps in TOF-PET", *Physics in Medicine and Biology*, vol. 61, no. 7, pp. 2802–2837, 2016. DOI: 10.1088/0031-9155/61/7/2802. [Online]. Available: https://doi.org/10.1088/0031-9155/61/7/2802.

[35] C. van Eijk, "Radiation detector developments in medical applications: inorganic scintillators in positron emission tomography", *Radiation Protection Dosimetry*, vol. 129, pp. 13–21, 1-3 2008. DOI: doi.org/10.1093/rpd/ncn043. [Online]. Available: https://doi.org/10.1093/rpd/ncn043.

[36] F. Gramuglia, "High-Performance CMOS SPAD-Based Sensors for Time-of-Flight PET Applications", Ph.D. dissertation, EPFL, 2022. DOI: 10.5075/epfl-thesis-8720.

[37] M. F. Bieniosek, J. W. Cates, and C. S. Levin, "A multiplexed TOF and DOI capable PET detector using a binary position sensitive network", *Physics in medicine and biology*, vol. 61, pp. 7639–7651, 2016. DOI: 10.1088/0031-9155/61/21/7639.

[38] G. Borghi, V. Tabacchini, R. Bakker, and D. R. Schaart, "Sub-3 mm, near-200 ps TOF/DOI-PET imaging with monolithic scintillator detectors in a 70 cm diameter tomographic setup", *Physics in Medicine and Biology*, vol. 63, no. 15, p. 155 006, 2018. DOI: 10.1088/1361-6560/aad2a6. [Online]. Available: https://doi.org/10.1088/1361-6560/aad2a6.

[39] J. Y. Yeom, R. Vinke, and C. S. Levin, "Side readout of long scintillation crystal elements with digital SiPM for TOF-DOI PET", *Medical Physics*, vol. 41, no. 12, p. 122 501, 2014. DOI: doi.org/10.1118/1.4901524. [Online]. Available: https://doi.org/10.1118/1.4901524.

[40] E. Yoshida, I. Somlai-Schweiger, H. Tashima, S. I. Ziegler, and T. Yamaya, "Parameter Optimization of a Digital Photon Counter Coupled to a Four-Layered DOI Crystal Block With Light Sharing", *IEEE Transactions on Nuclear Science*, vol. 62, no. 3, pp. 748–755, 2015. DOI: 10.1109/TNS.2015.2420577.

[41] R. Dolenec, "Time-of-Flight Positron Emission Tomography Using Cherenkov Radiation", Ph.D. dissertation, University of Ljubljana, 2012.

[42] E. Venialgo, "PET detector technologies for next-generation molecular imaging", Ph.D. dissertation, Delft University of Technology, 2019. DOI: https://doi.org/10.4233/uuid:427e3ce3-2b01-4fa0-9e80-bf0e9c033213.

[43] D. R. Schaart, G. Schramm, J. Nuyts, and S. Surti, "Time of Flight in Perspective: Instrumental and Computational Aspects of Time Resolution in Positron Emission Tomography", *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 5, pp. 598–618, 2021. DOI: 10.1109/TRPMS.2021.3084539.

[44] *Photomultiplier tube*, 2006. [Online]. Available: https://commons.wikimedia.org/wiki/File:Photomultipliertube.svg.

[45] H. P. K. K. E. Committee, "Phototmultiplier tubes. Basics and Applications", Hamamatsu, Tech. Rep., 2007. [Online]. Available: https://www.hamamatsu.com/content/dam/hamamatsu-photonics/sites/documents/99_SALES_LIBRARY/etd/PMT_handbook_v3aE.pdf.

[46] M. Pelletier and C. Pelletier, "RAMAN SPECTROSCOPY | Instrumentation", in *Encyclopedia of Analytical Science (Second Edition)*, P. Worsfold, A. Townshend, and C. Poole, Eds., Second Edition, Oxford: Elsevier, 2005, pp. 94–104, ISBN: 978-0-12-369397-6. DOI: https://doi.org/10.1016/B0-12-369397-7/00529-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B012369397700529X.

[47] E. Venialgo, C. Verrastro, D. Estryk, *et al.*, "PET calibration method of nonlinear position estimation algorithms for continuous NaI(Tl) crystals", in *2011 IEEE Nuclear Science Symposium Conference Record*, 2011, pp. 3359–3364. DOI: 10.1109/NSSMIC.2011.6152609.

[48] *Hamamatsu pmt*. [Online]. Available: https://www.hamamatsu.com/eu/en/product/optical-sensors/pmt/pmt_tube-alone.html.

[49] *Photek PMT*. [Online]. Available: https://www.photek.com/product-overview/.

[50] T. Frach, G. Prescher, C. Degenhardt, R. de Gruyter, A. Schmitz, and R. Ballizany, "The digital silicon photomultiplier — Principle of operation and intrinsic detector performance", in *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, 2009, pp. 1959–1965. DOI: 10.1109/NSSMIC.2009.5402143.

[51] L. H. C. Braga, L. Gasparini, L. Grant, *et al.*, "A Fully Digital 8 × 16 SiPM Array for PET Applications With Per-Pixel TDCs and Real-Time Energy Output", *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 301–314, 2014. DOI: 10.1109/JSSC.2013.2284351.

[52] A. Gola, C. Piemonte, and A. Tarolli, "Analog circuit for timing measurements with large area SiPMs coupled to LYSO crystals", in *2011 IEEE Nuclear Science Symposium Conference Record*, 2011, pp. 725–731. DOI: 10.1109/NSSMIC.2011.6154091.

[53] S. Dolinsky, G. Fu, and A. Ivan, "Timing resolution performance comparison for fast and standard outputs of SensL SiPM", in *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC)*, 2013, pp. 1–6. DOI: 10.1109/NSSMIC.2013.6829520.

[54] E. A. G. Webster and R. K. Henderson, "A TCAD and Spectroscopy Study of Dark Count Mechanisms in Single-Photon Avalanche Diodes", *IEEE Transactions on Electron Devices*, vol. 60, no. 12, pp. 4014–4019, 2013. DOI: 10.1109/TED.2013.2285163.

[55] I. Rech, A. Ingargiola, R. Spinelli, *et al.*, "Optical crosstalk in single photon avalanche diode arrays: a new complete model", *Opt. Express*, vol. 16, no. 12, pp. 8381–8394, 2008. DOI: 10.1364/OE.16.008381. [Online]. Available: http://opg.optica.org/oe/abstract.cfm?URI=oe-16-12-8381.

[56] I. Goushcha, B. Tabbert, and A. O. Goushcha, "Optical and electrical crosstalk in pin photodiode array for medical imaging applications", in *2007 IEEE Nuclear Science Symposium Conference Record*, vol. 6, 2007, pp. 4348–4353. DOI: 10.1109/NSSMIC.2007.4437077.

[57] S. Jahromi and J. Kostamovaara, "Timing and probability of crosstalk in a dense CMOS SPAD array in pulsed TOF applications", *Opt. Express*, vol. 26, no. 16, pp. 20 622–20 632, 2018. DOI: 10.1364/OE.26.020622. [Online]. Available: http://opg.optica.org/oe/abstract.cfm?URI=oe-26-16-20622.

[58] A. Ingargiola, M. Segal, A. Gulinatti, *et al.*, "Optical crosstalk in SPAD arrays for high-throughput single-molecule fluorescence spectroscopy", *Nuclear instruments and methods in physics research*, vol. 9, no. 12, 255–258, 2018. DOI: doi.org/10.1016/j.nima.2017.11.070. [Online]. Available: https://doi.org/10.1016/j.nima.2017.11.070.

[59] C. Joshua W., G. Stefan, A. Etiennette, L. Paul, and S. L. Craig, "Improved single photon time resolution for analog SiPMs with front end readout that reduces influence of electronic noise", *Physics in Medicine and Biology*, vol. 63, no. 18, p. 185 022, 2018. DOI: 10.1088/1361-6560/aadbcd. [Online]. Available: https://doi.org/10.1088/1361-6560/aadbcd.

[60] J. Zhang, P. Maniawski, and M. V. Knoop, "Performance evaluation of the next generation solid-state digital photon counting PET/CT system", *EJNMMI Research*, vol. 8, pp. 8–97, Nov. 2018. DOI: 10.1186/s13550-018-0448-7.

[61] T. Frach, G. Prescher, and C. Degenhardt, "Silicon photomultiplier technology goes fully-digital", 2010.

[62] J. Jourdon, S. Lhostis, S. Moreau, *et al.*, "Hybrid bonding for 3D stacked image sensors: impact of pitch shrinkage on interconnect robustness", in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 7.3.1–7.3.4. DOI: 10.1109/IEDM.2018.8614570.

[63] Z. Wang, "Microsystems using three-dimensional integration and TSV technologies: Fundamentals and applications", *Microelectronic Engineering*, vol. 210, pp. 35–64, 2019, ISSN: 0167-9317. DOI: https://doi.org/10.1016/j.mee.2019.03.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167931719300553.

[64] S. E. Derenzo and M. W. W., "Critical instrumentation issues for resolution <2mm, high sensitivity brain PET", *Quantification of Brain Function, Tracer Kinetics and Image Analysis in Brain PET*, pp. 25–40, 1992.

[65] M. W. W., "Fundamental Limits of Spatial Resolution in PET", *Nuclear instruments and methods in physics research*, vol. 1, S236–S240, 2011. DOI: doi.org/10.1016/j.nima.2010.11.092.

[66] S. E. Derenzo, T. F. Budinger, R. H. Huesman, and J. L. Cahoon, "DYNAMIC POSITRON EMISSION TOMOGRAPHY IN MAN USING SMALL BISMUTH GERMANATE CRYSTALS", in *Sixth International Conference on Positron Annihilation*, 1982. [Online]. Available: https://escholarship.org/uc/item/1gg8f0cg.

[67] J. van Sluis, J. de Jong, J. Schaar, *et al.*, "Performance Characteristics of the Digital Biograph Vision PET/CT System", *Journal of nuclear medicine*, vol. 7, pp. 1031–1036, Jan. 2019. DOI: 10.2967/jnumed.118.215418.

[68] *Biograph Vision PET/CT*. [Online]. Available: https://www.siemens-healthineers.com/molecular-imaging/pet-ct/biograph-vision.

[69] J. S. Karp, V. Viswanath, M. J. Geagan, *et al.*, "PennPET Explorer: Design and Preliminary Performance of a Whole-Body Imager", *Journal of nuclear medicine*, vol. 1, pp. 136–143, Jan. 2020. DOI: 10.2967/jnumed.119.229997.

[70] *BiographQuadra PET*. [Online]. Available: https://www.siemens-healthineers.com/molecular-imaging/pet-ct/biograph-vision-quadra.

[71] *Vereos PET*. [Online]. Available: https://www.usa.philips.com/healthcare/product/HC882446/vereos-digital-petct-proven-accuracy-inspires-confidence#documents.

[72] *uExplorer PET*. [Online]. Available: https://usa.united-imaging.com/products/molecular-imaging/uexplorer/.

[73] C.-C. Chen, C.-L. Chen, W. Fang, and Y.-C. Chu, "All-Digital CMOS Time-to-Digital Converter With Temperature-Measuring Capability", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 9, pp. 2079–2083, 2020. DOI: 10.1109/TVLSI.2020.3007587.

[74] Y. J. Park and F. Yuan, "0.25–4 ns 185 MS/s 4-bit pulse-shrinking time-to-digital converter in 130 nm CMOS using a 2-step conversion scheme", in *2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2015, pp. 1–4. DOI: 10.1109/MWSCAS.2015.7282113.

[75] M. Kim, W.-Y. Shin, G.-M. Hong, *et al.*, "High-resolution and wide-dynamic range time-to-digital converter with a multi-phase cyclic Vernier delay line", in *2013 Proceedings of the ESSCIRC (ESSCIRC)*, 2013, pp. 311–314. DOI: 10.1109/ESSCIRC.2013.6649135.

[76] N. U. Andersson and M. Vesterbacka, "A Vernier Time-to-Digital Converter With Delay Latch Chain Architecture", *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 61, no. 10, pp. 773–777, 2014. DOI: 10.1109/TCSII.2014.2345289.

[77] K. Cui and X. Li, "A High-Linearity Vernier Time-to-Digital Converter on FPGAs With Improved Resolution Using Bidirectional-Operating Vernier Delay Lines", *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 8, pp. 5941–5949, 2020. DOI: 10.1109/TIM.2019.2959423.

[78] S.-H. Chung, K.-D. Hwang, W.-Y. Lee, and L.-S. Kim, "A high resolution metastability-independent two-step gated ring oscillator TDC with enhanced noise shaping", in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, pp. 1300–1303. DOI: 10.1109/ISCAS.2010.5537261.

[79] M. Perenzoni, H. Xu, and D. Stoppa, "Small area 0.3 pJ/conv, 45 ps time-to-digital converter for arrays of silicon photomultiplier interfaces in 150 nm CMOS", in *Electronics Letters*, vol. 51, 2015, pp. 1933–1935. DOI: 10.1049/el.2015.2761.

[80] A. Muntean, E. Venialgo, A. Ardelean, *et al.*, "Blumino: The First Fully Integrated Analog SiPM With On-Chip Time Conversion", *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 5, pp. 671–678, 2021. DOI: 10.1109/TRPMS.2020.3045081.

[81] S. Bickley, H. Chan, and B. Torgler, "Artificial intelligence in the field of economics", *Scientometrics*, vol. 127, pp. 2055–2084, 2022. DOI: 10.1007/s11192-022-04294-w. [Online]. Available: https://doi.org/10.1007/s11192-022-04294-w.

[82] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, "A guide to machine learning for biologists", *Nature Reviews Molecular Cell Biology*, vol. 23, pp. 40–55, 2022. DOI: doi.org/10.1038/s41580-021-00407-0. [Online]. Available: https://doi.org/10.1038/s41580-021-00407-0.

[83] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2009. DOI: http://aima.cs.berkeley.edu.

[84] D. Simon, *Evolutionary Optimization Algorithms, Biologically-Inspired and Population-Based Approaches to Computer Intelligence*. John Wiley and Sons, Inc., 2013.

[85] E. Varol Altay and B. Alatas, "Performance analysis of multi-objective artificial intelligence optimization algorithms in numerical association rule mining", *J Ambient Intell Human Computation*, vol. 11, pp. 3449–3469, 2020. DOI: https://doi.org/10.1007/s12652-019-01540-7. [Online]. Available: https://doi.org/10.1038/s41580-021-00407-0.

[86] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, pp. 436–444, 2015. DOI: doi.org/10.1038/nature14539. [Online]. Available: https://doi.org/10.1038/nature14539.

[87] C. Floyd, "An artificial neural network for SPECT image reconstruction", *IEEE Transactions on Medical Imaging*, vol. 10, no. 3, pp. 485–487, 1991. DOI: 10.1109/42.97600.

[88] M. T. Munley, C. E. Floyd, J. E. Bowsher, and R. E. Coleman, "An artificial neural network approach to quantitative single photon emission computed tomographic reconstruction with collimator, attenuation, and scatter compensation", *Medical physics.*, vol. 21, no. 12, 1994-12, ISSN: 0094-2405.

[89] A. J. Reader, G. Corda, A. Mehranian, C. d. Costa-Luis, S. Ellis, and J. A. Schnabel, "Deep Learning for PET Image Reconstruction", *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 1, pp. 1–25, 2021. DOI: 10.1109/TRPMS.2020.3014786.

[90] K. Gong, E. Berg, S. R. Cherry, and J. Qi, "Machine Learning in PET: From Photon Detection to Quantitative Image Reconstruction", *Proceedings of the IEEE*, vol. 108, no. 1, pp. 51–68, 2020. DOI: 10.1109/JPROC.2019.2936809.

[91] I. Häggström, C. R. Schmidtlein, G. Campanella, and T. J. Fuchs, "DeepPET: A deep encoder–decoder network for directly solving the PET image reconstruction inverse problem", *Medical Image Analysis*, vol. 54, pp. 253–262, 2019, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2019.03.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841518305838.

[92] K. Gong, J. Guan, C.-C. Liu, and J. Qi, "PET Image Denoising Using a Deep Neural Network Through Fine Tuning", *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 153–161, 2019. DOI: 10.1109/TRPMS.2018.2877644.

[93] S. I. Kwon, R. Ota, E. Berg, *et al.*, "Ultrafast timing enables reconstruction-free positron emission imaging", *Nature Photonics*, vol. 15, pp. 914–918, 12 2021. DOI: 10.1038/s41566-021-00871-2. [Online]. Available: https://doi.org/10.1038/s41566-021-00871-2.

[94] E. Venialgo, S. Mandai, T. Gong, D. R. Schaart, and E. Charbon, "Time estimation with multichannel digital silicon photomultipliers", *Physics in medicine and biology*, vol. 60 6, pp. 2435–52, 2015.

[95] E. Berg and S. R. Cherry, "Using convolutional neural networks to estimate time-of-flight from PET detector waveforms", *Physics in Medicine and Biology*, vol. 63, no. 2, 02LT01, 2018. DOI: 10.1088/1361-6560/aa9dc5. [Online]. Available: https://doi.org/10.1088/1361-6560/aa9dc5.

[96] P. Carra, M. G. Bisogni, E. Ciarrocchi, *et al.*, "A neural network-based algorithm for simultaneous event positioning and timestamping in monolithic scintillators", *Physics in Medicine and Biology*, vol. 67, no. 13, p. 135 001, 2022. DOI: doi.org/10.1088/1361-6560/ac72f2.

[97] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks", *Towards data science*, vol. 6, no. 12, pp. 310–316, 2017.

[98] B. Karlik and A. V. Olgac, "Performance analysis of various activation functions in generalized MLP architectures of neural networks", *International Journal of Artificial Intelligence and Expert Systems*, vol. 1, no. 4, pp. 111–122, 2011.

[99] F. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi, "Learning activation functions to improve deep neural networks", *arXiv preprint arXiv:1412.6830*, 2014.

[100] S. Gómez, D. Sánchez, J. Mauricio, *et al.*, "Multiple Use SiPM Integrated Circuit (MUSIC) for Large Area and High Performance Sensors", *Electronics*, vol. 10, no. 8, 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10080961. [Online]. Available: https://www.mdpi.com/2079-9292/10/8/961.

[101] F. Gramuglia, A. Muntean, E. Venialgo, *et al.*, "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.

[102] M.-J. Lee, A. R. Ximenes, P. Padmanabhan, *et al.*, "High-Performance Back-Illuminated Three-Dimensional Stacked Single-Photon Avalanche Diode Implemented in 45-nm CMOS Technology", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, pp. 1–9, 2018. DOI: 10.1109/JSTQE.2018.2827669.

[103] W. Shen, T. Harion, H. Chen, *et al.*, "A Silicon Photomultiplier Readout ASIC for Time-of-Flight Applications Using a New Time-of-Recovery Method", *IEEE Transactions on Nuclear Science*, vol. 65, no. 5, pp. 1196–1202, 2018. DOI: 10.1109/TNS.2018.2821769.

[104] T Harion, K Briggl, H Chen, *et al.*, "STiC — a mixed mode silicon photomultiplier readout ASIC for time-of-flight applications", *Journal of Instrumentation*, vol. 9, no. 02, pp. C02003–C02003, 2014. DOI: 10.1088/1748-0221/9/02/c02003. [Online]. Available: https://doi.org/10.1088/1748-0221/9/02/c02003.

[105] R. Becker, C. Casella, S. Corrodi, *et al.*, "Studies of the high rate coincidence timing response of the STiC and TOFPET ASICs for the SAFIR PET scanner", *Journal of Instrumentation*, vol. 11, no. 12, P12001–P12001, 2016. DOI: 10.1088/1748-0221/11/12/p12001. [Online]. Available: https://doi.org/10.1088/1748-0221/11/12/p12001.

[106] A. Comerma, D. Gascón, L. Freixas, *et al.*, "Flextot - current mode ASIC for readout of common cathode SiPM arrays", in *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC)*, 2013, pp. 1–2. DOI: 10.1109/NSSMIC.2013.6829761.

[107] P. Jarron, E. Auffray, S. Brunner, *et al.*, "Time based readout of a silicon photomultiplier (SiPM) for time of flight positron emission tomography (TOF-PET)", in *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, 2009, pp. 1212–1219. DOI: 10.1109/NSSMIC.2009.5402391.

[108] *A brief Introduction to Silicon Photomultiplier (SiPM) Sensors*, AND9795/D, Rev. 3, ONSEMI, Jan. 2019.

[109] S. J. Kim, D. Kim, and M. Seok, "Comparative study and optimization of synchronous and asynchronous comparators at near-threshold voltages", in *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2017, pp. 1–6. DOI: 10.1109/ISLPED.2017.8009169.

[110] M. Bazes, "Two novel fully complementary self-biased CMOS differential amplifiers", *IEEE Journal of Solid-State Circuits*, vol. 26, no. 2, pp. 165–168, 1991. DOI: 10.1109/4.68134.

[111] B. Chappell, T. Chappell, S. Schuster, *et al.*, "Fast CMOS ECL receivers with 100-mV worst-case sensitivity", *IEEE Journal of Solid-State Circuits*, vol. 23, no. 1, pp. 59–67, 1988. DOI: 10.1109/4.257.

[112] A. Sachdeva, "Design of Low-Threshold Comparator for Improved Timing-Resolution Analog/Digital SiPM", Master thesis, 2018.

[113] M. Z. Straayer and M. H. Perrott, "A Multi-Path Gated Ring Oscillator TDC With First-Order Noise Shaping", *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1089–1098, 2009. DOI: 10.1109/JSSC.2009.2014709.

[114] J. Chen, H. Yumei, and H. Zhiliang, "A multi-path gated ring oscillator based time-to-digital converter in 65 nm CMOS technology", *Journal of Semiconductors*, vol. 34, p. 035 004, 2013.

[115] A. Muntean, "Design of a fully digital analog SiPM with sub-50 ps time conversion", Master thesis, 2017.

[116] S. Burri, C. Bruschini, and E. Charbon, "LinoSPAD: A Compact Linear SPAD Camera System with 64 FPGA-Based TDC Modules for Versatile 50 ps Resolution Time-Resolved Imaging", *Instruments*, vol. 1, no. 1, 2017, ISSN: 2410-390X. DOI: 10.3390/instruments1010006. [Online]. Available: https://www.mdpi.com/2410-390X/1/1/6.

[117] *Direct Time-of-Flight Depth Sensing Reference Designs*, TND6341/D, Rev. 3, ONSEMI, Aug. 2021.

[118] B. Piatek, "What is an SiPM and how does it work", Hamamatsu, Tech. Rep., Oct. 2016.

[119] Y. Kong, G. Pausch, K. Romer, *et al.*, "Linearization of Gamma Energy Spectra in Scintillator-Based Commercial Instruments", *IEEE Transactions on Nuclear Science*, vol. 57, no. 3, pp. 1430–1434, 2010. DOI: 10.1109/TNS.2009.2033684.

[120] S. Tsigaridas, M. Beuzekom, H. Graaf, *et al.*, "Timewalk correction for the Timepix3 chip obtained with real particle data", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 930, pp. 185–190, 2019, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2019.03.077. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016890021930419X.

[121] J. Du, J. P. Schmall, M. S. Judenhofer, K. Di, Y. Yang, and S. R. Cherry, "A Time-Walk Correction Method for PET Detectors Based on Leading Edge Discriminators", *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 1, no. 5, pp. 385–390, 2017. DOI: 10.1109/TRPMS.2017.2726534.

[122] S. Mandai, V. Jain, and E. Charbon, "A $780 \times 800 \mu m^2$ Multichannel Digital Silicon Photomultiplier With Column-Parallel Time-to-Digital Converter and Basic Characterization", *IEEE Transactions on Nuclear Science*, vol. 61, no. 1, pp. 44–52, 2014. DOI: 10.1109/TNS.2013.2294022.

[123] A. Carimatto, S. Mandai, E. Venialgo, *et al.*, "11.4 A 67,392-SPAD PVTB-compensated multi-channel digital SiPM with 432 column-parallel 48ps 17b TDCs for endoscopic time-of-flight PET", in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015, pp. 1–3. DOI: 10.1109/ISSCC.2015.7062996.

[124] W. Jiang, Y. Chalich, and D. M. Jamal, "Sensors for Positron Emission Tomography Applications", *Sensors MDPI*, vol. 19, no. 22, 2019. DOI: 10.3390/s19225019.

[125] V. C. Spanoudaki and C. S. Levin, "Photo-Detectors for Time of Flight Positron Emission Tomography (ToF-PET)", *Sensors*, vol. 10, no. 11, pp. 10 484–10 505, 2010, ISSN: 1424-8220. DOI: 10.3390/s101110484. [Online]. Available: https://www.mdpi.com/1424-8220/10/11/10484.

[126] S. Seifert, H. T. van Dam, and D. R. Schaart, "The lower bound on the timing resolution of scintillation detectors", *Physics in Medicine and Biology*, vol. 57, no. 7, pp. 1797–1814, 2012. DOI: 10.1088/0031-9155/57/7/1797. [Online]. Available: https://doi.org/10.1088/0031-9155/57/7/1797.

[127] M. W. Fishburn and E. Charbon, "System Tradeoffs in Gamma-Ray Detection Utilizing SPAD Arrays and Scintillators", *IEEE Transactions on Nuclear Science*, vol. 57, no. 5, pp. 2549–2557, 2010. DOI: 10.1109/TNS.2010.2064788.

[128] E. Venialgo, S. Mandai, T. Gong, D. Schaart, and E. Charbon, "Practical time mark estimators for multichannel digital silicon photomultipliers", in *2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2015, pp. 1–3. DOI: 10.1109/NSSMIC.2015.7582138.

[129] F. Gramuglia, A. Muntean, C. A. Fenoglio, *et al.*, "Architecture and Characterization of a CMOS 3D-Stacked FSI Multi-Channel Digital SiPM for Time-of-Flight PET Applications", in *2021 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2021, pp. 1–2. DOI: 10.1109/NSS/MIC44867.2021.9875625.

[130] C. Zhang, S. Lindner, I. M. Antolovic, M. Wolf, and E. Charbon, "A CMOS SPAD Imager with Collision Detection and 128 Dynamically Reallocating TDCs for Single-Photon Counting and 3D Time-of-Flight Imaging", *Sensors*, vol. 18, no. 11, 2018, ISSN: 1424-8220. DOI: 10.3390/s18114016. [Online]. Available: https://www.mdpi.com/1424-8220/18/11/4016.

[131] C. Zhang, S. Lindner, I. M. Antolović, J. Mata Pavia, M. Wolf, and E. Charbon, "A 30-frames/s, 252 × 144 SPAD Flash LiDAR With 1728 Dual-Clock 48.8-ps TDCs, and Pixel-Wise Integrated Histogramming", *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 1137–1151, 2019. DOI: 10.1109/JSSC.2018.2883720.

[132] E. Manuzzato, L. Gasparini, M. Perenzoni, *et al.*, "A 16 × 8 Digital-SiPM Array With Distributed Trigger Generator for Low SNR Particle Tracking", in *ESSCIRC 2019 - IEEE 45th European Solid State Circuits Conference (ESSCIRC)*, 2019, pp. 75–78. DOI: 10.1109/ESSCIRC.2019.8902571.

[133] Y. Haemisch, *Digital Photon Counting (DPC) - a scalable light detection technology with high temporal resolution*, 2013.

[134] S. Lindner, "Time-resolved Single-photon Detector Arrays for High Resolution Near-infrared Optical Tomography", Ph.D. dissertation, EPFL, 2018. DOI: 10.5075/epfl-thesis-8815.

[135] D. Hwang, S. K. Kang, K. Y. Kim, *et al.*, "Generation of PET Attenuation Map for Whole-Body Time-of-Flight 18F-FDG PET/MRI Using a Deep Neural Network Trained with Simultaneously Reconstructed Activity and Attenuation Maps", *Journal of Nuclear Medicine*, vol. 60, no. 8, pp. 1183–1189, 2019. DOI: 10.2967/jnumed.118.219493.

[136] M. A. Miller, "Vereos Digital PET/CT Performance", Philips, Tech. Rep., Jun. 2016.

[137] J. Naranjo-Torres, M. Mora, R. Hernández-García, R. J. Barrientos, C. Fredes, and A. Valenzuela, "A Review of Convolutional Neural Network Applied to Fruit Image Processing", *Applied Sciences*, vol. 10, no. 10, 2020, ISSN: 2076-3417. DOI: 10.3390/app10103443. [Online]. Available: https://www.mdpi.com/2076-3417/10/10/3443.

[138] K. Gong, J. Guan, K. Kim, *et al.*, "Iterative PET Image Reconstruction Using Convolutional Neural Network Representation", *IEEE Transactions on Medical Imaging*, vol. 38, no. 3, pp. 675–685, 2019. DOI: 10.1109/TMI.2018.2869871.

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

[140] *Geant4*. [Online]. Available: https://geant4.web.cern.ch/.

[141] J. Allison, K. Amako, J. Apostolakis, *et al.*, "Geant4 developments and applications", *IEEE Transactions on Nuclear Science*, vol. 53, no. 1, pp. 270–278, 2006. DOI: 10.1109/TNS.2006.869826.

[142] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, and V. B. S. Prasath, "Choosing Mutation and Crossover Ratios for Genetic Algorithms—A Review with a New Dynamic Approach", *Information*, vol. 10, no. 12, 2019, ISSN: 2078-2489. DOI: 10.3390/info10120390. [Online]. Available: https://www.mdpi.com/2078-2489/10/12/390.

[143] K. Man, K. Tang, and S. Kwong, "Genetic algorithms: concepts and applications [in engineering design]", *IEEE Transactions on Industrial Electronics*, vol. 43, no. 5, pp. 519–534, 1996. DOI: 10.1109/41.538609.

[144] J. Holland, *Adaptation in Neural and Artifical Systems: An Introductory Analysis with Applications to Biology, Control, and Artifical Intelligence*. MIT Press, 1992. [Online]. Available: https://ieeexplore.ieee.org/book6267401.

[145] K. Guo, M. Yang, and H. Zhu, "Application research of improved genetic algorithm based on machine learning in production scheduling", *Neural Computing and Applications*, vol. 32, Apr. 2020. DOI: 10.1007/s00521-019-04571-5.

[146] H. Gupta and D. Singh, "Speech Feature Extraction and Recognition Using Genetic Algorithm", vol. 9001, Feb. 2008.

[147] K. Ashish, A. Dasari, S. Chattopadhyay, and N. B. Hui, "Genetic-neuro-fuzzy system for grading depression", *Applied Computing and Informatics*, vol. 14, no. 1, pp. 98–105, 2018, ISSN: 2210-8327. DOI: https://doi.org/10.1016/j.aci.2017.05.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210832717301242.

[148] N. Papanna, A. R. M. Reddy, and M. Seetha, "EELAM: Energy efficient lifetime aware multicast route selection for mobile ad hoc networks", *Applied Computing and Informatics*, vol. 15, no. 2, pp. 120–128, 2019, ISSN: 2210-8327. DOI: https://doi.org/10.1016/j.aci.2017.12.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210832717302302.

[149] A. A. AbdulHamed, M. A. Tawfeek, and A. E. Keshk, "A genetic algorithm for service flow management with budget constraint in heterogeneous computing", *Future Computing and Informatics Journal*, vol. 3, no. 2, pp. 341–347, 2018, ISSN: 2314-7288. DOI: https://doi.org/10.1016/j.fcij.2018.10.004.

[150] A. Piszcz and T. Soule, "Genetic programming: Optimal population sizes for varying complexity problems", vol. 1, Jan. 2006, pp. 953–954. DOI: 10.1145/1143997.1144166.

[151] S. P. Gotshall and B. Rylander, "Optimal Population Size and the Genetic Algorithm", 2002. DOI: 10.1.1.105.2431. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.105.2431.

[152] P. Diaz-Gomez and D. Hougen, "Initial Population for Genetic Algorithms: A Metric Approach.", Jan. 2007, pp. 43–49.

[153] R. Oladele and J. Sadiku, "Genetic Algorithm Performance with Different Selection Methods in Solving Multi-Objective Network Design Problem", *International Journal of Computer Applications*, vol. 70, pp. 5–9, May 2013. DOI: 10.5120/12012-7848.

[154] N. Razali and J. Geraghty, "Genetic Algorithm Performance with Different Selection Strategies in Solving TSP", vol. 2, Jan. 2011.

[155] P. Kora and P. Yadlapalli, "Crossover Operators in Genetic Algorithms: A Review", *International Journal of Computer Applications*, vol. 162, pp. 34–36, Mar. 2017. DOI: 10.5120/ijca2017913370.

[156] P. K. Diederik and B. Jimmy, "Adam: A Method for Stochastic Optimization", 2015. DOI: doi.org/10.48550/arXiv.1412.6980.

[157] *Low breakdown voltage type MPPC for scintillation detector*, S14160/S14161 series, Hamamatsu, Jun. 2020.

[158] R. Eki, S. Yamada, H. Ozawa, *et al.*, "9.6 A 1/2.3inch 12.3Mpixel with On-Chip 4.97TOPS/W CNN Processor Back-Illuminated Stacked CMOS Image Sensor", *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 64, pp. 154–156, 2021.

[159] S. Choi, J. Lee, K. Lee, and H.-J. Yoo, "A 9.02mW CNN-stereo-based real-time 3D hand-gesture recognition processor for smart mobile devices", in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 220–222. DOI: 10.1109/ISSCC.2018.8310263.

[160] S. Park, S. Choi, J. Lee, M. Kim, J. Park, and H.-J. Yoo, "14.1 A 126.1mW real-time natural UI/UX processor with embedded deep-learning core for low-power smart glasses", in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, 2016, pp. 254–255. DOI: 10.1109/ISSCC.2016.7418003.

Chip gallery

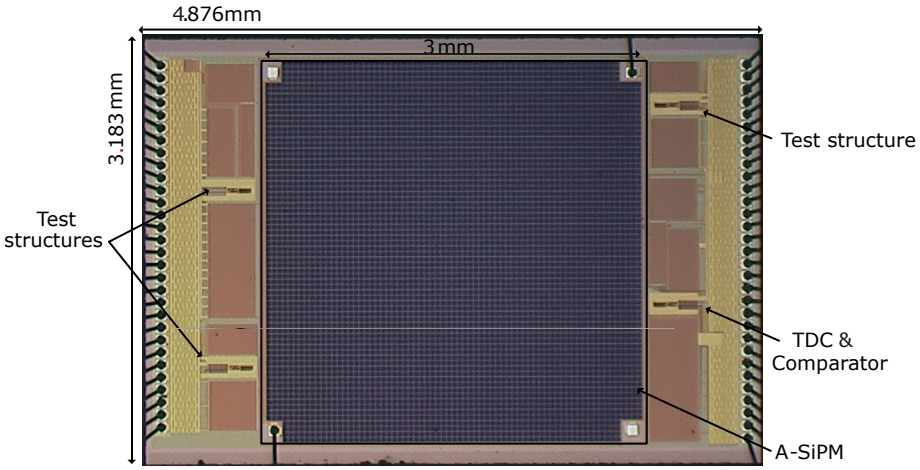


Figure 5.1: Blumino - Chapter 2.

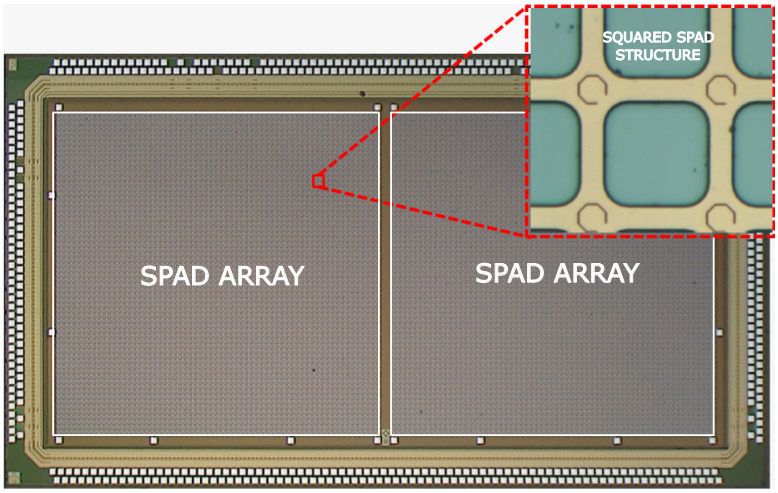


Figure 5.2: Blueberry - Chapter 3.

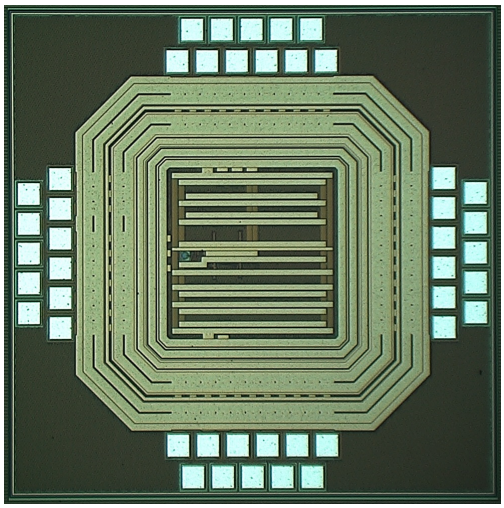


Figure 5.3: Blueberry TDC - Chapter 3.

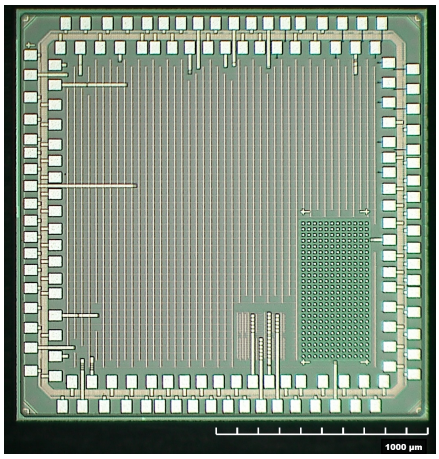


Figure 5.4: Smarty (part of the Sansa SoC) - Chapter 4.

List of publications

Journal articles

- A. Muntean, E. Venialgo, A. Ardelean, A. Sachdeva, E. Ripiccini, D. Palubiak, C. Jackson and E. Charbon, "Blumino: The First Fully Integrated Analog SiPM With On-Chip Time Conversion", *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 5, p. 671-678, 2021, DOI: 10.1109/TRPMS.2020.3045081.
- A. Muntean, et. al., "Smarty: A Fully Integrated and Reconfigurable artificial neural network for Positron Emission Tomography" - in preparation.

Proceedings articles

- A. Muntean, F. Gramuglia, E. Venialgo, C. Bruschini and E. Charbon, "Tradeoffs in Cherenkov Detection for Positron Emission Tomography", *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2018, DOI: 10.1109/NSSMIC.2018.8824430.
- A. Muntean, E. Venialgo, S. Gnecchi, C. Jackson and E. Charbon, "Towards a fully digital state-of-the-art analog SiPM", *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2018, DOI: 10.1109/NSSMIC.2017.8533036. - first place, student paper award.
- A. Muntean, A. Sachdeva, E. Venialgo, S. Gnecchi, D. Palubiak, C. Jackson and E. Charbon, "A Fully Integrated State-of-the-Art Analog SiPM with on-chip Time Conversion", *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2019, DOI: 10.1109/NSSMIC.2018.8824662.
- F. Gramuglia*, A. Muntean*, E. Venialgo, M.J. Lee, S. Lindner, M. Motoyoshi, A. Ardelean, C. Bruschini and E. Charbon, "CMOS 3D-Stacked FSI Multi-Channel Digital SiPM for Time-of-Flight PET Applications", *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2020, DOI: 10.1109/NSS/MIC42677.2020.9507833. (*equally contributing authors)

- F. Gramuglia*, A. Muntean*, C. A. Fenoglio, E. Venialgo, M. J. Lee, S. Lindner, M. Motoyoshi, A. Ardelean, C. Bruschini and E. Charbon, "Architecture and Characterization of a CMOS 3D-Stacked FSI Multi-Channel Digital SiPM for Time-of-Flight PET Applications", *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2021, DOI: 10.1109/NSS/MIC44867.2021.9875625. (*equally contributing authors)

Workshops and poster presentations

- A. Muntean, "Fully integrated analog SiPM with on-chip time conversion", *International SPAD Sensors Workshop*, 2020, workshop.
- A. Muntean, "Towards High-performance SPAD based detectors for Positron Emission Tomography", *Colloquia in Intelligent Sensing, Measurement and Actuators*, 2021, workshop.
- A. Muntean, "Blumino: the first fully integrated analog SiPM with on-chip discrimination and time conversion", *International SPAD Sensor Workshop*, 2022, poster.

Awards

First Place Student Paper Award, "Towards a fully digital state-of-the-art analog SiPM", IEEE Nuclear Science Symposium Sydney, NSW, Australia, 2018.

Patent

A. Muntean, E. Charbon, Imaging system with silicon photomultipliers and method for operating thereof, Patent Application Nr. PCT/EP2021/071002, 27 July, 2021.

Curriculum vitae

Andrada Alexandra Muntean

1992 Born in Timișoara, Romania

Education

2017 - 2022 **Ecole Polytechnique Fédérale de Lausanne (EPFL)**
Advanced Quantum Architecture Laboratory (AQUA)
Lausanne, Switzerland
Ph.D. in Microelectronics
Thesis: "*Integrated electronics for time-of-flight positron emission tomography photodetectors*"

2015 - 2017 **Delft University of Technology (TUDelft)**
Circuit and Systems Group (CAS)
Delft, The Netherlands
M.Sc. in Microelectronics
Thesis: "*Design of a fully digital analog SiPM with sub-50 ps time conversion*"

2011 - 2015 **Politehnica University of Timișoara (UPT)**
Timișoara, Romania
B.Sc in Electronics Engineering
Thesis: "*Distance estimation through stereoscopy*"

Professional Experience

2016 - 2016 **NASA Jet Propulsion Laboratory**
Pasadena, California, USA
Advanced UV/Vis/NIR Detector Arrays, Systems and Nanoscience Group
Intern
Project: "*Characterization of a compact far UV spectrometer*"

2014 - 2015 **Continental Automotive Romania**
Timișoara, Romania
Software Developer