High speed electronics for SPAD image sensors used in TimeofFlight applications

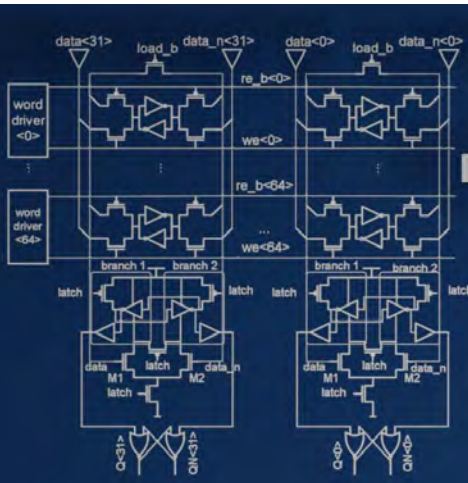
Carimatto, A.J.

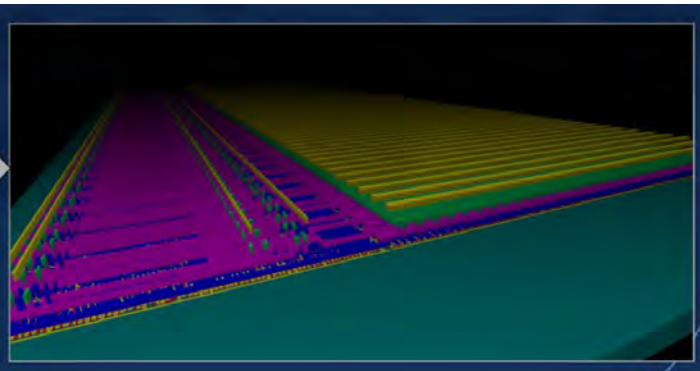
DOI
10.4233/uuid:228d9463-2c98-4cb6-b7f0-ac274e890edd

Publication date
2020

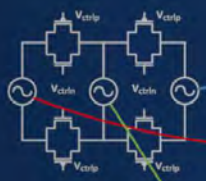
Document Version
Final published version

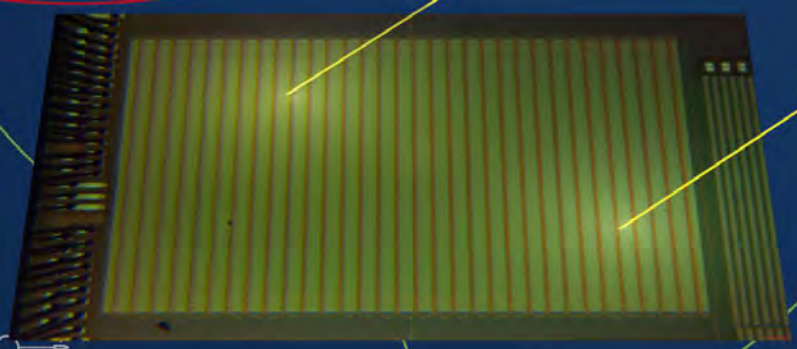
Citation (APA)
Carimatto, A. J. (2020). *High speed electronics for SPAD image sensors used in TimeofFlight applications*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:228d9463-2c98-4cb6-b7f0-ac274e890edd





High speed electronics for SPAD image sensors used in Time-of-Flight applications





Augusto José Carimatto

High speed electronics for SPAD image sensors used in Time-of-Flight applications

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
maandag 16 november 2020 om 14:20 uur

door

Augusto José CARIMATTO

Elektronisch Ingenieur, Universidad Tecnológica Nacional, Argentinië. Geboren te Ciudad Autónoma de Buenos Aires, Argentinië.

Dit proefschrift is goedgekeurd door de

promotor: Prof. dr. ir. E. E. E. Charbon

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. ir. E. E. E. Charbon	Technische Universiteit Delft promotor

Onafhankelijke leden:

Prof. dr. S. Stallinga,	Technische Univeristeit Delft
Prof. dr. ir. A. J. P. Theuswissen,	TU-Delft
Prof. dr. P. French	Imperial College
Prof. dr. E. Garutti,	University of Hamburg
Dr. M. Cazzaniga,	Intuitive Surgical
Dr. N. A. W. Dutton,	STMicroelectronics

High speed electronics for SPAD image sensors used in Time-of-Flight applications

Dissertation

Dissertation
for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus Prof. dr. ir. T.H.J.J. van der Hagen,
Chair of the Board for Doctorates
to be defended publicly on
Monday 16^{th}, November 2020 at 14:20

by

Augusto José CARIMATTO

Electronics Engineer, Universidad Tecnológica Nacional, Argentina. Born in Ciudad Autónoma de Buenos Aires, Argentina.

This dissertation has been approved by

promoter: prof. dr. ir. E. E. E. Charbon

Composition of the doctoral committee:

Rector Magnificus	chairman
Prof. dr. ir. E. E. E. Charbon	Technische Universiteit Delft promotor

Independent members:

Prof. dr. S. Stallinga,	Technische Univeristeit Delft
Prof. dr. ir. A. J. P. Theuswissen,	Technische Univeristeit Delft
Prof. dr. P. French	Imperial College
Prof. dr. E. Garutti,	University of Hamburg
Dr. M. Cazzaniga,	Intuitive Surgical
Dr. N. A. W. Dutton,	STMicroelectronics

TU Delft Delft University of Technology

Keywords: CMOS SPAD TDC PET LiDAR ARRAY ANN

Printed by:

Front & Back: designed by the author.

ISBN 000-00-0000-000-0

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

To my family and friends

Augusto J. Carimatto

Contents

Summary

Multi Digital Silicon Photon Multipliers (MD-SiPM), as image sensors, are utilized to calculate and estimate the properties of the incident light. These properties include spatial location of hits, intensity or number of photons and time of arrival. Some characteristics can be more important than others depending upon the problem at hand. Among endless applications where MD-SiPMs are used for, Positron Emission Tomography and LiDAR are the two that this thesis is focused on.

Positron Emission Tomography is an imaging technique to monitor functional information about tissue and organs, including early cancer lesions. This constitutes the main difference with structural techniques such as radiography, where, by means of X-rays, a projection of a section of the body under test is obtained. This topic is thoroughly discussed in chapter 1 and the results of the implementation are shown in chapter 4.

MD-SiPMs are employed in PET systems to indirectly capture gamma photons by using scintillators as intermediate transducer. Position, energy and time-of-arrival of the gamma photons are measured and calculated by MD-SiPMs. This information is utilized by algorithms to reconstruct a 3-D structure to represent the tissue under test. The quality of the final image is of paramount importance so as to help medical doctors to assess cancer risk and study other metabolic diseases. To this end, the characteristics and features of MD-SiPMs are essential.

The second application, addressed in this thesis, is LiDAR. This topographic technique is used to generate a graphical representation of a physical scene. A laser is employed to illuminate the whole scene and the reflected photons are detected by MD-SiPMs. Researchers have developed two different methods for this matter. In the first one, called flash method, the scene is evenly targeted by a laser by means of an optical diffuser. In the second one, called scanning method, the scene is swept employing a laser and two mobile mirrors that redirect the laser along X and Y axis. The characteristics of MD-SiPMs are crucial to obtain a high-quality image. Chapter 5 shows the design, implementations and their results.

Both of these applications mentioned here, along with many other applications where MD-SiPMs can be used for, generate massive amount of data. Many modules are included in these systems not only to reduce the volume of data by dismissing unimportant information, but also for processing and compressing information so that less bits need be transmitted. Chapter 2 is dedicated for read-out modules. The most common problems about read-out found in MD-SiPMs are explained and several techniques and solutions to cope with them are introduced.

Synchronization, timing, hit-counting capabilities, data transmission and many other features are highly desired to have in these optical systems. This thesis goes through all the building blocks of MD-SiPMs step by step, covering all the aspects

of the design and implementation. New techniques for photon counting and timing measurements are introduced along with their implementations in chapter 3.

The thesis continues in the next three chapters (4, 5 and 6) with three concrete MD-SiPM designs. The first sensor (Concolor) addresses PET applications. The design is explained in detail; energy and timing measurements for PET are shown. The second sensor (Panther) addresses LiDAR applications; many new ideas for hit-counting and timing measurements are introduced in this chapter, accompanied with results.

The thesis concludes in the last chapter (6) showing the most advanced design purposed by this essay, MindHive. This sensor holds most of the features of the previous two MD-SiPMs (proved to work) and it was designed aiming at vision applications where data can be processed in real-time by using an implementation of a four-layers feed-forward neural network. Results are shown for two different applications. At last, and outlook of the new generation of MD-SiPMs is discussed, setting the grounds for the upcoming smart image sensors based on SPADs and artificial intelligence.

Samenvatting

Multi Digitale Silicone Foton Vermenigvuldigers (MDSiPM) als beeldsensoren, worden gebruikt om de eigenschappen van het invallende licht te berekenen en te schatten. Deze eigenschappen omvatten ruimtelijke locatie van fotoninvallingen, intensiteit of aantal fotonen en aankomsttijd. Sommige kenmerken kunnen belangrijker zijn dan andere, afhankelijk van het probleem dat moet worden opgelost. Onder de eindeloze toepassingen waar MDSiPM's voor worden gebruikt, zijn Positron Emissie Tomografie en LiDAR de twee waarop dit proefschrift is gericht.

Positron emissie tomografie is een beeldvormende test om functionele informatie over weefsel en organen weer te geven, inclusief vroege kankerdetectie. Dit is het belangrijkste verschil met structurele technieken zoals radiografie, waarbij door middel van röntgenstralen een projectie van een deel van het te testen lichaam wordt verkregen. Dit onderwerp wordt uitvoerig besproken in hoofdstuk 1 en de resultaten van de implementatie worden weergegeven in hoofdstuk 4.

MDSiPM's worden gebruikt in PETsystemen om gammafotonen indirect vast te leggen door scintillatoren als tussenliggende omvormer te gebruiken. Positie, energie en aankomsttijd van de gammafotonen worden gemeten en berekend door MDSiPM's. Deze informatie wordt door algoritmen gebruikt om een 3Dstructuur te reconstrueren om het te testen weefsel weer te geven. De kwaliteit van het uiteindelijke beeld is van het allergrootste belang om artsen te helpen het risico op kanker te beoordelen en andere metabole ziekten te bestuderen. Daarom zijn de eigenschappen en kenmerken van MDSiPM's essentieel.

De tweede applicatie, behandeld in dit proefschrift, is LiDAR. Deze topografische techniek wordt gebruikt om een grafische weergave van een fysieke scène te genereren. Een laser wordt gebruikt om de hele scène te verlichten en de gereflecteerde fotonen worden gedetecteerd door MDSiPM's. Onderzoekers hebben hiervoor twee verschillende methoden ontwikkeld. In de eerste, de flitsmethode, wordt de scène gelijkmatig belicht door een laser door middel van een optische verspreider. In de tweede methode, de scanmethode, wordt de scène gescand met een laser en twee mobiele spiegels die de laser langs de X en Yas sturen. De kenmerken van MDSiPM's zijn cruciaal om een beeld van hoge kwaliteit te verkrijgen. Hoofdstuk 5 toont het ontwerp, de implementatie en hun resultaten.

Beide hier genoemde toepassingen, samen met vele andere toepassingen waar MDSiPM's voor kunnen worden gebruikt, genereren een enorme hoeveelheid gegevens. Veel modules zijn in deze systemen opgenomen om niet alleen het gegevensvolume te verminderen door onbelangrijke informatie te negeren, maar ook om informatie te verwerken en te comprimeren zodat er minder bits moeten worden verzonden. Hoofdstuk 2 is gericht op uitleesmodules. De meest voorkomende problemen met het uitlezen van MDSiPM's worden uitgelegd en er worden verschillende technieken, oplossingen en ideeën geïntroduceerd om hiermee om te gaan.

Synchronisatie, timing, mogelijkheden tot invaltellingen, datatransmissie en vele andere kenmerken zijn zeer gewenst in deze optische systemen. Dit proefschrift bespreekt stap voor stap alle bouwstenen van MDSiPM's en behandelt alle aspecten van het ontwerp en de implementatie. Nieuwe technieken voor het tellen van foto's en tijdsmetingen worden samen met hun implementaties geïntroduceerd in hoofdstuk 3.

Het proefschrift gaat verder in de volgende drie hoofdstukken (4, 5 en 6) met drie concrete MDSiPMontwerpen. De eerste sensor (Concolor) richt zich op PET toepassingen. Het ontwerp wordt in detail uitgelegd; energie en tijdsmetingen voor PET worden getoond. De tweede sensor (Panther) richt zich op LiDAR toepassingen; In dit hoofdstuk worden veel nieuwe ideeën voor het tellen van foton invallingen en tijdsmetingen geïntroduceerd, vergezeld van meetresultaten.

Het proefschrift eindigt in het laatste hoofdstuk (6) en laat het meest geavanceerde ontwerp zien van dit proefschrift, MindHive. Deze sensor bevat de meeste kenmerken van de vorige twee MDSiPM's (bewezen te werken) en is ontworpen voor zicht toepassingen waarbij gegevens onmiddellijk kunnen worden verwerkt door gebruik te maken van een vierlaagse voorwaarts gekoppeld neuraal netwerk. Resultaten worden getoond voor twee verschillende toepassingen. Tenslotte worden de vooruitzichten van de nieuwe generatie MDSiPM's besproken, die de basis vormt voor de aanstaande slimme beeldsensoren op basis van SPAD's en kunstmatige intelligentie.

Resumen

Los Foto Multiplicadores Digitales de Estado Sólido (MD-SiPM por sus siglas en inglés), como sensores de imagen, se utilizan para calcular y estimar las propiedades de la luz incidente. Estas propiedades incluyen la localización espacial de cada evento, su intensidad o número de fotones y su tiempo de arribo. Algunas de estas características pueden ser más importantes que otras dependiendo del problema que se persigue resolver. Entre un sinfin de aplicaciones donde se utilizan los MD-SiPMs, la Tomografía por Emisión de Positrones (PET por sus siglas en inglés) y LiDAR son las dos en las cuales se centra esta tesis.

La Tomografía por Emisión de Positrones es un estudio imagenológico para obtener información funcional de tejidos y órganos. Esto constituye la mayor diferencia con técnicas estructurales como radiografía, donde por medio de rayos X, se realiza una proyección de las secciones del cuerpo. Este tema se discute extensamente en el capítulo 1 y los resultados de las implementaciones se encuentran en el capítulo 4.

Los MD-SiPMs se emplean en sistemas PET para capturar fotones Gamma indirectamente por medio de centelladores utilizados como transductores intermedios. La posición, energía y tiempo de arribo de los fotones Gamma se miden y calculan por medio de los MD-SiPMs. Esta información es usada por varios algoritmos para reconstruir estructuras 3D para representar el tejido bajo estudio. La calidad de la imagen final es sumamente importante ya que ayuda a los profesionales médicos a evaluar el riesgo de cáncer y estudiar enfermedades metabólicas. Con este objetivo, las características de los MD-SiPMs son esenciales.

La segunda aplicación de interés en esta tesis es LiDAR. Esta técnica topográfica se usa para generar una representación gráfica de una escena física. Por medio de un laser, se ilumina la escena completa; los fotones reflejados se detectan usando MD-SiPMs. Los investigadores han desarrollado dos formas diferentes para conseguirlo. En la primera, llamada método flash, la escena se ilumina uniformemente por un laser difuminado. En la segunda forma, llamada método por escaneo, se barre la escena con un laser y dos espejos que lo reflejan en los ejes X e Y. Las características de los MD-SiPMs son cruciales para obtener una imagen de alta definición. En el capítulo 5 se incluyen sendas implementaciones y sus resultados.

Ambas de estas aplicaciones mencionadas aquí, junto con muchas otras aplicaciones donde se pueden utilizar MD-SiPMs, generan una cantidad masiva de información. Existen muchos módulos que se incluyen en estos sistemas no solamente para reducir el volumen de datos desechando información irrelevante, sino también procesando y comprimiendo la información, de este modo reduciendo el número de bits que necesitan ser transmitidos. El capítulo 2 está dedicado a los módulos de comunicación. El lector podrá ver los problemas más comunes que se encuentran

en los MD-SiPMs acerca de las técnicas de transmisión de datos así como también varias soluciones e ideas para resolverlos.

La sincronización, medición temporal, capacidades de conteo de fotones y otras características son altamente deseables en estos sistemas ópticos. Esta tesis cubre todos los tópicos relacionados con los bloques constitutivos de los MD-SiPMs paso por paso, tocando todos los aspectos de su diseño e implementación. En el capítulo 3 se presentan nuevas técnicas de fotoconteo y medición temporal y sus implementaciones.

La tesis continua en los siguientes tres capítulos (4, 5 and 6) con tres diseños concretos utilizando MD-SiPMs. El primer sensor (Concolor) está enfocado en aplicaciones PET. Se explica el diseño en detalle, acompañado por mediciones de tiempo y energía para PET. El segundo sensor (Panther) se enfoca en aplicaciones para LiDAR. Se presentan varias ideas nuevas para fotoconteo y mediciones temporales, también acompañado de los resultados.

La tesis concluye con el ultimo capítulo (6) donde se muestra el diseño más avanzado desarrollado para este ensayo (MindHive). Este sensor contiene las características ya empleadas en los dos anteriores MD-SiPMs (testeadas) y fue diseñado para aplicaciones de visión donde la información se procesa en tiempo real utilizando una implementación de cuatro capas de una red neuronal de propagación directa (feed-forward). Se muestran los resultados para dos aplicaciones diferentes. Para concluir, se discute el futuro de la nueva generación de los detectores MD-SiPMs, sentando los cimientos para los sensores de imagen emergentes basados en SPADs e inteligencia artificial.

1

Introduction

Augusto José Carimatto

"Those who can imagine anything, can create the impossible."

Alan Turing

"The enchanting charms of this sublime science reveal only to those who have the courage to go deeply into it."

Carl Friedrich Gauss

"There cannot be a language more universal and more simple, more free from errors and obscurities...more worthy to express the invariable relations of all natural things than Mathematics. It interprets all phenomena by the same language."

Joseph Fourier

"You cannot hope to build a better world without improving the individuals. To that end, each of us must work for our own improvement."

Marie Curie

This thesis explores all the aspects of time-of-flight image sensors based on Multi Digital Silicon Photo Multipliers (MD-SiPM) and introduces three designs to cover the main applications nowadays that are Positron Emission Tomography (PET) and LiDAR. This essay sets the grounds for the next upcoming sensors.

1

1.1. Image sensors

Image sensors, particularly Multi-channel Digital Silicon Photon Multipliers (MD-SiPM) that are presented in this thesis, are used for photon counting and estimation of the three intrinsic characteristics of the incident light which are spatial position, intensity or number of photons and time-of-arrival. The application where these MD-SiPMs are designed for dictates which aspects of the aforementioned information are the most important; the sensor will therefore be designed for that matter enhancing the key features to estimate the properties that are most required. In this thesis, the term image sensor is used to refer to all kind of CMOS imagers using SPADs.

Historically, PhotoMultiplier Tubes (PMT) were the first devices used for photon detection. The structure of PMTs comprises a photocathode, a chain of dynodes and the anode as shown in Fig. 1.1. The photo-cathode material releases electrons when is hit by photons due to the photo-electric phenomena (chapter 9 of [1]). The dynodes, biased with high positive voltages, accelerate the free electrons making them collide with the next dynode; thus, generating an avalanche of electrons. As a consequence of this process, at the end of the PMT, a current pulse, which is a multiplication of the first electron, is generated and it can be read by external electronics. The main characteristics of PMTs to be considered are [2]:

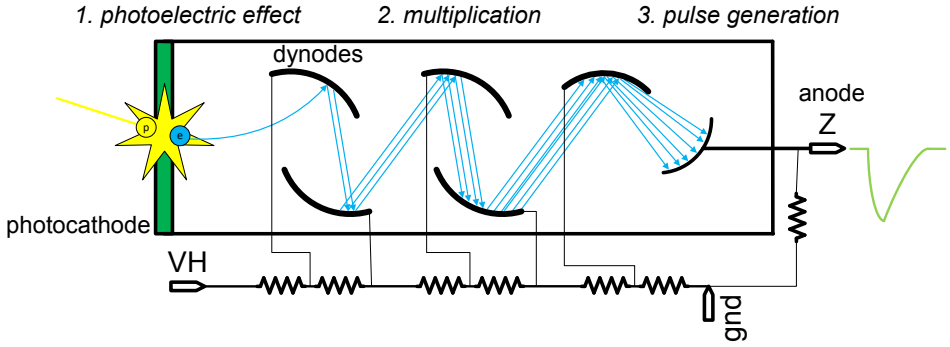


Figure 1.1: Diagram of PMTs. Whole sequence from the moment a photon hits the photocathode until the pulse is generated.

- Quantum efficiency: number of electrons generated over number of incident photons (wavelength dependent).
- Transport time: time needed by a generated electron to become a current pulse at the end of the dynode chain.
- Time spread: jitter of the the transport time.
- Cathode material: this will determine the wavelength spectrum that the PMT is sensitive to.
- Dark count rate: spurious pulses generated in absence of light.
- Collection efficiency: the percentage of electrons generated by the photocathode that generate a pulse at the anode. It is related to the applied voltage.

PMTs were a success and were employed in many applications such as Gamma Camera, Spectroscopy and Positron Emission Tomography among others. However, their intrinsic limitations such as cost, sensitivity to magnetic fields, size and required high voltages to operate, gave an opportunity to other emerging sensors like CCDs or CMOS-SPAD detectors, which are the main focus of this thesis. It should be mentioned though, that some of those once-upon-a-time disadvantages of PMTs have been mitigated over time.

MD-SiPMs are fabricated in CMOS technology and use Single Photon Avalanche Diode (SPAD) as their photo-sensing component. SPADs are P-N devices that are reverse-biased and can generate electrical pulses when they get hit by photons. The basic structure of SPADs are shown in Fig. 1.2 [3].

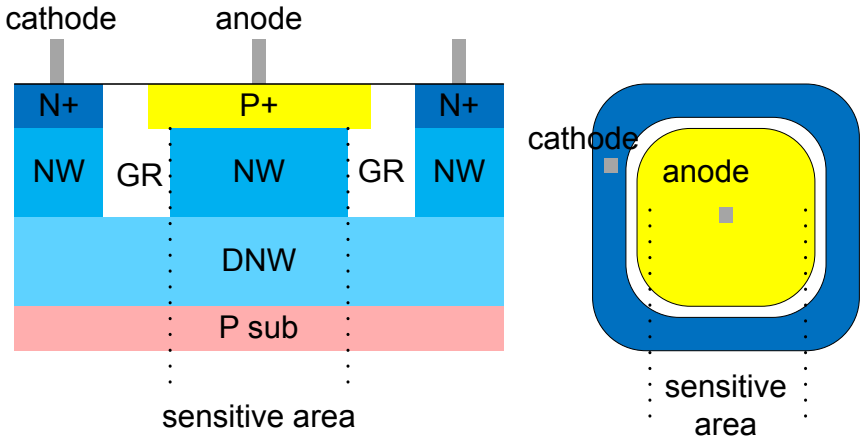


Figure 1.2: SPAD structure. The P-N junction can be implemented by any combination of P and N materials: P+ / NW, P+ / DNW, P+ / BNW, PW / DNW, etc.. Profile view on the left, top view on the right. Hereafter SPADs are yellow-colored in all the plots.

1.2. Applications

SPAD-based image sensors have been employed in many applications such as PET, fluorescence lifetime microscopy, 3-D imaging, LiDAR, high energy physics, etc..

1.2.1. LiDAR

3-D imaging is a topographic method to create a 3-D graphical representation of a physical target. The working principle is based on the illumination of a scene with a pulsed laser and the detection of the photons that reflect off the target. By calculating the time-of-arrival of these photons, it is possible to create a representation model with X, Y, and depth information. Z information is calculated as

$$Z = c\frac{\Delta t}{2} \tag{1.1}$$

1

,

where Δt is the time-of-flight of the photons. LiDAR is mainly used for 3-D mapping and its applications are endless; the most prominent ones being self-driving vehicles, space mapping, navigation, transport, civil engineering, robotics, etc.. Fig. 1.3 shows the basics of LiDAR systems along with all the components.

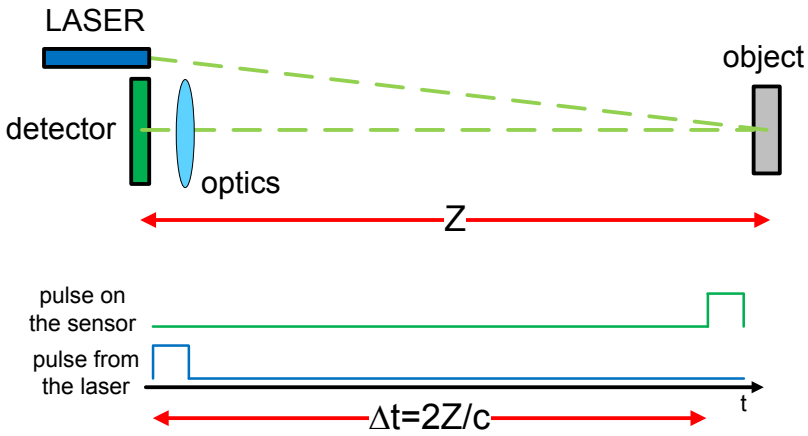


Figure 1.3: Basic diagram of the time-of-flight computation at the core of LiDAR applications. The optics projects the image of the scene into the sensor.

The measurement sequence is as follows. The laser photons are shot synchronously with the clock of the system and go through the output optics towards the scene. The photons reflect and/or scatter on different surfaces and obstacles. Part of those reflected photons travel back to the input optics and they are focused onto an optical sensor. In this thesis, the optical detectors are MD-SiPMs, that have been largely proved to work in this application [4]. The detection of the photon comprises X and Y position and time information; depending on the system, this is processed on-chip or off-chip or a combination of the two in order to generate a 3-D map of the scene [5]. The resolution of the final representation depends on the resolution of X and Y axis, as well as on the time resolution that will be translated into space resolution in the Z axis. There are other constraints that these systems usually have, such as background noise for open-space environments and eye safety when they are operating in an environment with human interaction. Eye safety will limit the maximum laser power that can be used for a given laser optics and, ultimately, the maximum distance that can be measured [6]. The wavelength of choice for LiDAR systems are always in the non-visible spectrum to not interfere with humans' vision. Large efforts have been made into coping with the main problems of LiDAR which are signal-to-noise and background ratio (SNBR), multiple-reflection paths and system-to-system interference. There exist two main approaches for LiDAR systems, which are flash method and scanning method; both explained in the following. An animation of how LiDAR operates is available in [7].

Methods

Scanning method: for this method, the laser is focused only onto one point of the scene [8]. The laser is swept along X and Y axis using mirrors controlled by electronics in order to cover the whole field of view. Fig. 1.4 shows the scheme of this method. In this case, since the X and Y positions are known, the MD-SiPM is used for detection and for depth (Z) calculation. In reality the laser does not sweep with discrete steps but it rather does so continuously as Fig. 1.5 shows.

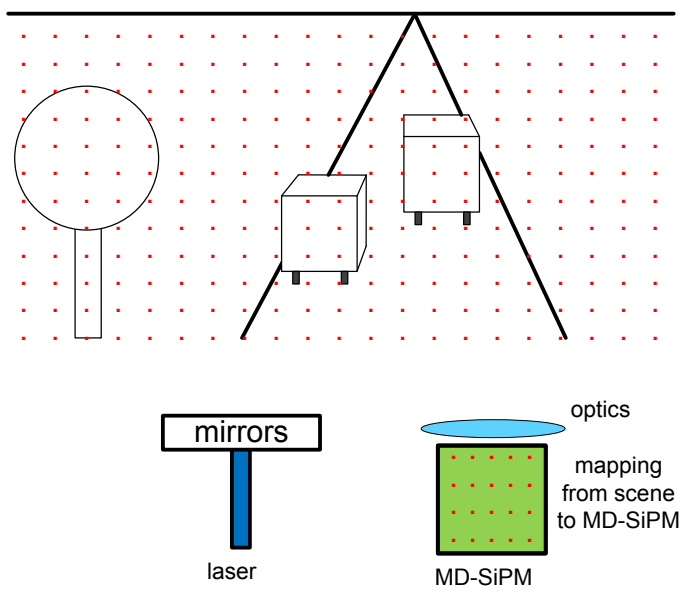


Figure 1.4: Diagram of scanning mode. The laser sweeps the scene along X and Y axis. The MD-SiPM is used to calculate Z (depth) information.

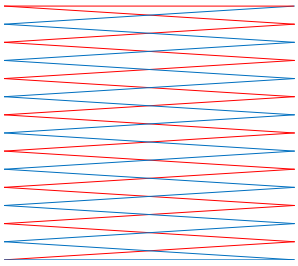


Figure 1.5: Possible trajectory of the laser in scanning mode. The laser moves continuously drawing a closed curve. Blue and red colors are to denote the forth and back trajectory.

The main drawback of this method is the fact that it intrinsically uses movable mechanical parts that can wear off and age with time. On the other hand, the levels of SNR are higher than those of its counterpart: Flash.

Flash method: contrasting to scanning method, for flash method, the whole scene is simultaneously illuminated by the laser [9]. The photons coming from the laser cover the whole field of view. Under these conditions, X and Y information are no longer known (like in scanning method), and the MD-SiPM is then used to measure X, Y and depth (Z) of every impinging photon. The optics coupled to the photo sensor has to be properly aligned and calibrated to make a direct mapping from (x;y) of the chip to (x;y) of the scene.

1.2.2. Positron Emission Tomography (PET)

Positron Emission Tomography is a non-invasive nuclear medical technique to assess metabolic activity in the body. A map of functional activity can be used for cancer early detection or organs malfunctions diagnosis. In contrast with structural tomographic methods where a map of density and location of tissue, bones and organs is obtained, in PET, the activity of the tissue and functional information can be observed in the final image. The basic principle of Positron Emission Tomography is the detection of two gamma photons that are simultaneously generated by positron-electron annihilations [10]. Positrons are released from a radiotracer held in a fluorodeoxyglucose (FDG) molecule that is injected into the body of the patient. Different carrier molecules are designed to cluster in the organs of interest. The detections of the system are used along with a reconstruction algorithm to generate a 3-D model of the tissue that had absorbed the radiotracer. Several algorithms are employed to generate the volume as Backprojection or iterative algorithms that model the different parts of the system to iteratively calculate the shape of the source that could have been generated by such data as explained in [11].

The whole PET system

Fig. 1.6 shows how a complete PET system looks like. The patient, tied to a stretcher, is placed in the center of the system. Billions of pairs of gamma photons are generated from the decay of the radiotracer and hit the detectors. These lines described by the trajectory of both gamma photons moving in opposite directions are called Lines-of-Response (LoR). If both gamma photons are detected by the sensors, their time-of-arrival can be used to calculate the place where the annihilation that generated the photons occurred. This calculation along with its uncertainty is saved in a long list with all the LoRs detected throughout the whole measurement. There are several reconstruction algorithms that can generate 3-D models of the tissue using this list and the geometry of the system as inputs.

In this thesis the detection process is made by means of SPAD sensors which will be explained in this section. SPAD sensors can indirectly capture gamma photons by using a scintillator that can generate visible photons after the absorption process. The full sequence is shown in Fig. 1.7.

Decay: the FDG molecule releases a positron that interacts with the surrounding tissue in the body. The average distance that positrons move is few millimeters and it depends on the density of the tissue and the their initial kinetic energy.

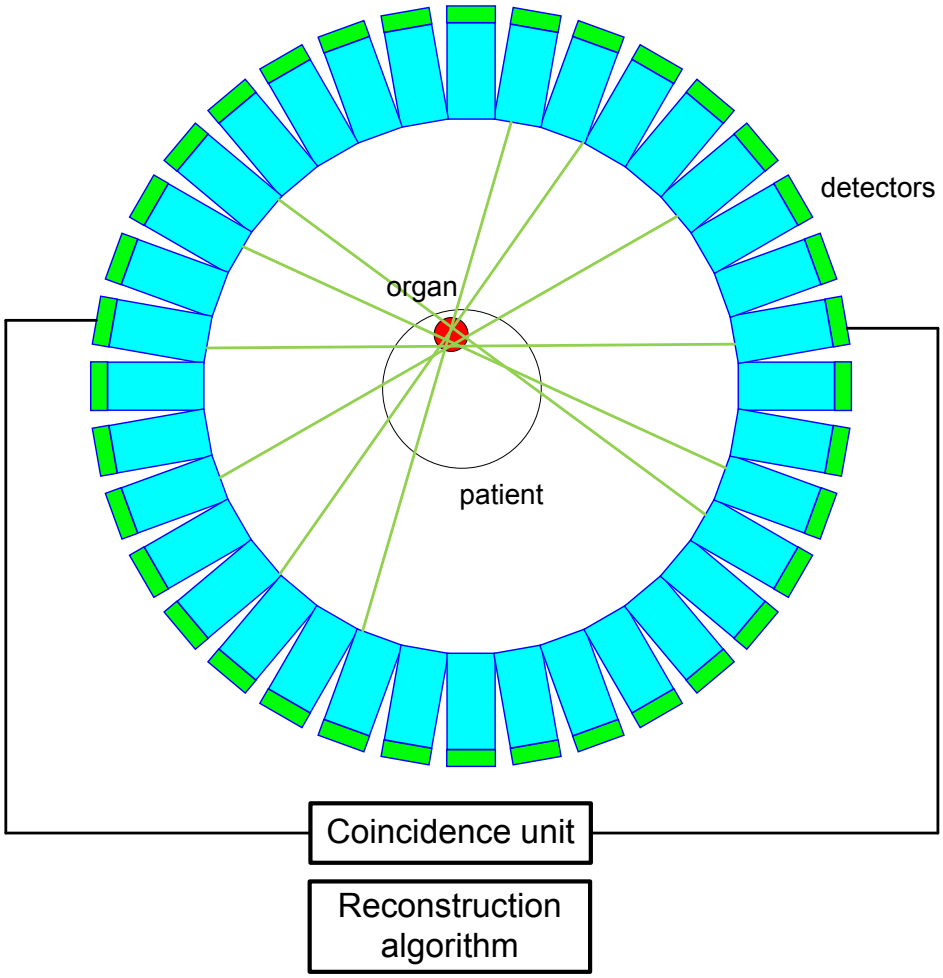


Figure 1.6: Diagram of a whole PET system. Concentric ring of modules to detect gamma photons and calculate the two points from which LoRs can be estimated.

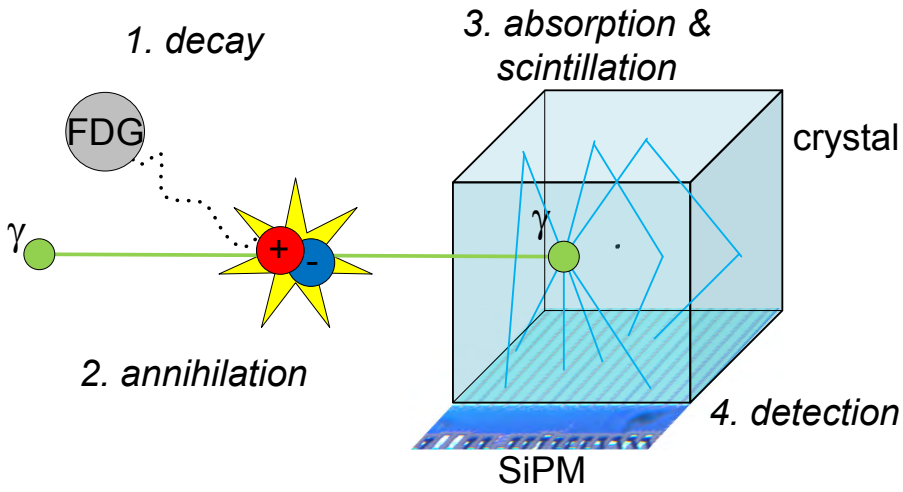


Figure 1.7: Whole process of gamma photon detection when a SPAD sensor is used. The MD-SiPM is required to provide estimation of the energy of the impinging gamma photon and its time-of-arrival.

Annihilation: the free positron interacts with an electron and the process, called annihilation, takes place. The pair of particles transform into two gamma photons that move along the same line but in opposite directions. The angle described by the gamma photons is not exactly 180º due to the fact that the initial momentum of the e+ e- pair might not be nil.

Absorption and scintillation: the generated gamma photons go through the crystals and, with certain probability, will be absorbed. Right after the absorption, another process takes place in the crystal: scintillation. Thousands of visible photons, a number that is proportional to the absorbed energy, are generated.The scintillation photons are emitted isotropically. If the crystal is wrapped with reflecting material, the photons reaching the surface are reflected back inside the crystal. Multiple reflection can take place until the photon reaches the facet coupled to the photodetector.

Detection: One of the faces is not covered by reflectors but by a MD-SiPM. The photons impinge the sensor and are detected with a certain probability. If the detection occurs, it is said there is a gamma event. The three main parameters of the gamma events are then estimated. This information comprises: position of, time-of-arrival and energy deposited into the crystal.

Singles and coincidences: every detection of a gamma photon is called single event and every time two gamma photons are detected simultaneously within a predefined time window it is said there is a coincidence event.

Estimation of the position of the annihilation

As said before, LoRs are used by algorithms to perform 3-D reconstructions. LoRs are defined once a coincidence event is detected. The position of the LoR depends on the place the detector was hit by the gamma photon and the relative position of these to the center of the system. By using the arrival time in both the sensors it is possible to estimate the actual position of the annihilation that ultimately gives the position where the FDG has been absorbed. This position, relative to the center of the LoR is calculated as:

$$x = c.\frac{t_0 - t_1}{2}, \tag{1.2}$$

where c is the speed of light and $t_{0/1}$ is the time-of-arrival of the gamma photon in the first or second detector.

1.3. MD-SiPMs

1.3.1. Implemented designs

Along with theory, conceptual and practical ideas, three full implementations of detectors have been designed for this thesis.

EndoToF chip: this sensor was not designed for this thesis but it was characterized for PET applications. It is constituted by an array of 9x18 MD-SiPMs with a bank of 432 TDCs that are shared along the array facilitating multiple indirect captures of gamma events by using a scintillator. Read-out electronics is included to send the information out the chip in two different modes. The first mode (frame-based) is meant to read every pixel and every TDC. The second mode (event-driven) is meant to be used for PET; it is possible to read up to 4 gamma events every 6.4 μs. The chip is part of a whole project called EndoTOFPET-US [12] that was designed to be used in a system with two different sensors to detect prostate and pancreatic cancer at early stages. The system has two detectors, one external with a shape of a plate that has an array of 12x12 independent detector modules, and one endoscopic detector that uses a scintillator coupled to this sensor plus an FPGA to read, process and send the information of the gamma events.

Concolor: it is a 2x2 MD-SiPM 2-D sensor that was designed for multiple applications that include PET and LiDAR. Every MD-SiPM has 64x128 SPADs with all the electronics associated to properly operate the sensor. Full integration of digital read-out has been included. A bank of Sliding Scale TDCs was included per MD-SiPM in order to time-stamp events with one of the lowest DNL and INL ever shown in these type of sensors. Frame-based and event-driven modes are available and a synchronization input makes the system scalable to operate jointly with more sensors of the same or different type. A LYSO scintillator was coupled to this sensor to perform measurements for PET. Concolor was the first MD-SiPM in 40nm technology. The keyword of this sensor is **PET**.

1

Panther: this design was mostly aimed at LiDAR. It was proven to work in flash mode. The sensor has one MD-SiPM of 64x64 SPADs. Panther is a 3-D design that has been used to test different type of SPADs and architectures. The electronics is capable of time-stamping events with a bin size up to 20ps and it can store hits and hit addresses to reconstruct 3-D scenes. Along with the same sensor, there is an auxiliary small system to perform very basic measurements. They keyword of this sensor is **LiDAR**.

MindHive: the last and most advanced design of this thesis, fabricated in TSMC 40nm technology, it includes all the features of the previous two sensors with enhancements in time capabilities and hit counting. MindHive resembles a honeycomb where every cell is equipped with smart electronics and can operate independently from the rest. A 4-layers neural network provides the sensor of versatility never implemented before in SPAD sensors. MindHive can address many vision computing applications by training the neural network for that matter. The chip is still under test. They keyword of this sensor is **Vision**.

1.3.2. Single-Photon Avalanche Diode (SPAD)

Introduction: SPADs are diodes used as the main transducer in image sensors to detect light photons that are converted into electrical pulses [13]. While APDs are diodes working in reversed bias mode below the breakdown voltage, SPADs operate above their breakdown voltage. A SPAD is characterized by 4 phases during detection of a photon: detection or seeding, avalanche, quenching and recharge.

Fig. 1.8 shows SPAD's operation when the SPAD is serially connected to a resistor R. V_{op} is the total voltage between the cathode and ground. The diode has all the voltage applied between its terminals and Z is equal to ground. At t_0, a thermal event or a photon creates a hole-electron pair and, with certain probability, the avalanche begins. The current increases exponentially as so does the voltage over R, so Z increases and once it overpasses the threshold of the buffer, Z_b becomes logical '1'. The process continues until the voltage across the diode is lower than the break-down voltage, defined by the geometry and composition of the diode. At time t_1, the avalanche is no longer sustainable and it is quenched. The SPAD now behaves as a capacitor; it thus starts charging with an RC constant until it reaches V_{op} between its terminals again. Z goes down until 0 and Z_b changes to low state as soon as Z goes below the threshold. The SPAD is ready for operation after this recharge process (t_2). During the time the quenching and the recharge take place, the SPAD cannot detect a new perturbation; this time is called dead time and is around tens of nanoseconds.

The information obtained from the SPAD essentially comprises the photon event in itself or the photon detection, and the time-of-arrival of the detected photon. The photon detection is usually stored in an internal memory and the time-of-arrival is measured by a time module, most commonly a Time-to-Digital Converter (TDC). The main characteristics of SPADs include:

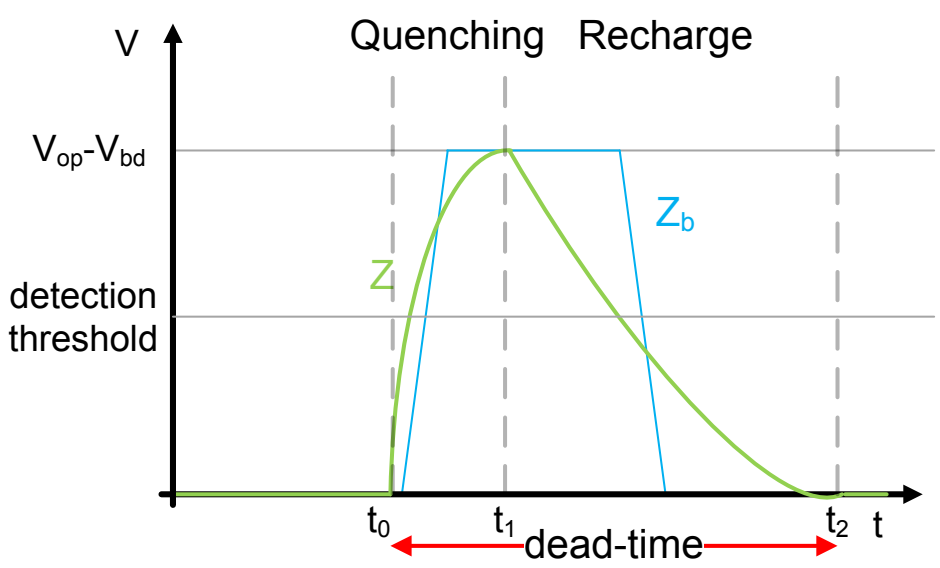


Figure 1.8: Typical waveform of a SPAD. The SPAD fires at t_0, is quenched at t_1 and it is ready for operation at t_2 after recharge.

Dark Count Rate (DCR): in the absence of light, SPADs fire due to thermal events (chapter 11 of [1]). The two sources of thermal events are band-to-band tunneling and trap assisted events. The number of avalanches generated per second in darkness conditions is called Dark Count Rate. Some applications are far more affected than others by this effect. LiDAR systems have to cope with background noise and, in general, DCR can be neglected. Other applications where photons are scarce like FLIM, low-DCR SPADs are highly desirable. Quantum random number generators are extremely sensitive to DCR, which must be suppressed. Typical values of DCR are in the range of 0.1 to 10 cps per μm^2.

Fill factor (FF): SPADs have an active area that is sensitive to light. Plenty of research has been done and is being done to try to maximize it. The fill factor is the ratio between active area and the total area of the SPAD. The fill factor can be increased by the use of microlenses that focus light into the active area of the SPAD [14].

Photon Detection Efficiency (PDE): is the probability of a SPAD to detect a photon. A photon impinging the SPAD has a probability to create a hole-electron pair that depends on several factors as structure of the SPAD and the wavelength of the incident light and excess bias voltage, the voltage at which the SPAD is biased above breakdown. Typical PDE plots for SPADs can be found in [14].

Photon Detection Probability (PDP): PDE, explained in the previous paragraph, is referred to the sensitive area of the SPAD. The PDP of the full sensor accounts for both the FF and the PDE. $PDE = FF * PDP$.

Jitter: SPADs are employed for photon detection and to measure their time-of-arrivals. The jitter of the SPAD is a very important parameter because establishes a hard lower-bound jitter of the system for a single photon. Any electronics added in the chain can only worsen this time with no possibilities of improving it. Although it is known that reducing the threshold of activation improves the jitter of the system [15], this method uses early detection of the avalanche and it does not change the intrinsic jitter of the SPAD. Typical values for the jitter are around 100ps [16].

Cross-talk: SPADs are connected and combined in very large structures and matrices to create large sensitive sensors. Groups of 64x64, 512x512 [17] or even 1024x1000 [18] are now common. The SPADs are not perfectly isolated and a firing SPAD can disturb its surroundings and can make other SPADs fire too. This effect is called cross-talk. There are two types of cross-talk based on the source of the disturbance. If this disturbance is photon-based, it is named ”optical cross-talk”. In case it is voltage/charge/current-based, it is called electrical cross-talk. It is interesting to notice that optical cross-talk can only be positive: a photon, generated during an avalanche in a SPAD, can trigger an avalanche in a neighbor SPAD; thus leading to over-counting. On the contrary, electrical cross-talk can be either positive or negative since different amplitudes of voltage, currents or charges might vary the applied voltage to another SPAD in unexpected ways; leading then to a temporary PDE variation that can result in over-counting or under-counting.

Masking: the DCR of SPADs follow a very particular shape within a chip as shown in Fig. 1.9. This includes most of the population with a stable DCR until the curve reaches a point, called ”the knee”, where the rate of DCR starts increasing. At the end, there are SPADs, called ”screamers”, that fire out of control. In a large system, SPADs are combined in several structures that will be analyzed in the next sections. If one of the SPADs used in those structures fires repeatedly, it can reduce the availability of the electronics that also serves the rest of the SPADs. This situation is usually prevented by masking the screamers with special electronics. The masking method is called ”electrical mask” when it blocks the pulses from the SPAD to prevent the saturation of the electronics; and it is called ”optical mask” when the electronics acts on the SPAD itself to prevent it from firing.

Quenching is the process through which the SPAD avalanche is stopped. In order to make this happen, the voltage of the SPAD has to drop below the breakdown voltage. There are basically two methods to do it: passive and active. In the first case, a simple resistor, as explained, is used to lower the voltage across the SPAD. A transistor, properly biased, can also be used to mimic a resistor behavior. This alternative is very flexible as the resistance can be controlled by the voltage applied to the gate of the transistor. The second method uses an active circuit

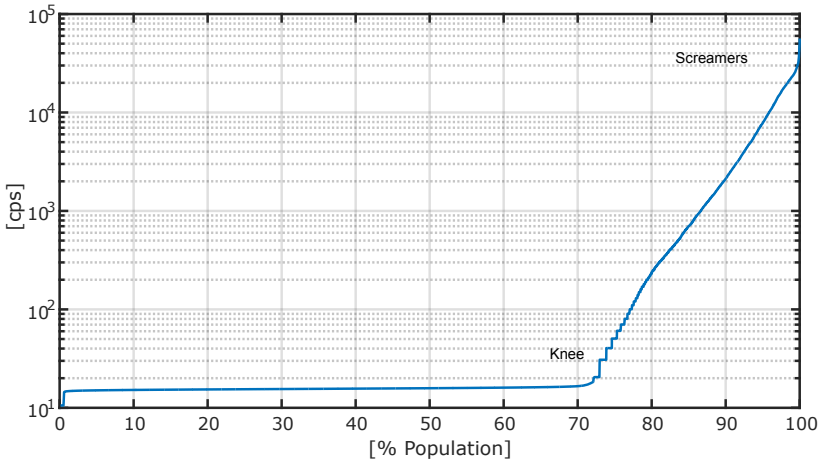


Figure 1.9: Typical DCR of image sensors. The first portion exhibits a flat DCR, knee and screamers can be observed in the right portion.

that detects the avalanche at early stage and sets the anode voltage below the break-down voltage to quickly quench the avalanche.

Recharge is the process to set the voltage across the SPAD to $V_{bd} + V_{EB}$ again. The voltage of the anode has to return to ground. The recharge can be done passively by using a resistor that usually is the same resistor for quenching, or optionally, it can be done actively by means of a transistor that pulls down the voltage to ground. Active recharge can be synchronous with the hit by placing a module that closes the recharge transistor right upon hit detections. Asynchronous mode is also possible by using the same transistor to pull down the voltage node, but this time the activation is done through an unrelated logic that could be, for instance, a global reset. More details are explained in [19]. On the other hand, passive recharge can only be synchronous with hits.

Interdependence of characteristics: though desirable, SPADs are rarely designed to improve all their characteristics because in many cases the improvement of one of them would be detrimental for another characteristic. For instance, to improve the cross-talk, SPADs should be placed farther from each other with the consequent reduction in the fill factor [20].

There are different schemes of quenching and recharge that are shown here. It should be noted that the simplest read-out scheme for a SPAD can use a resistor for quenching and recharge as shown in Fig. 1.8. Nevertheless, this has a drawback: both the quenching constant and the recharge constant depend on the same variable, which is the resistor R. In order to analyze these two constants, we have to keep in mind that the goal of the SPAD is to detect photons and to

measure their time-of-arrival. It is desirable to achieve high time resolution and low dead time so as to count as many events as possible. The time resolution is given by the jitter from the moment an avalanche is triggered until the time the pulse reaches the threshold voltage. Minimizing that time then improves the jitter because the longer that time is, more spurious spikes can vary the crossing point for the threshold. This means that if passive quenching is used, the resistor has to be as big as possible. As R increases, the recharge time gets longer and longer; thus extending the recharge time and the dead time. As a consequence the hit-counting capabilities lowers. From the recharge perspective R should be as small as possible. As a conclusion, only one resistor is not enough to achieve both conditions at the same time. Fig. 1.10 shows different combinations of quenching and recharge schemes. The scheme *b* is one of the simplest ones used in image sensors but it has the drawback already explained. Scheme *c* has active quenching and passive or active recharge depending on the mode of operation over that transistor. This scheme is very useful because it accepts a logic operating on Vq to implement the masking feature. If that voltage is held to V_{EB} the SPAD is not biased and will not fire. Schemes *d* and *e* are very similar. They implement passive quenching, passive or active recharge.

Concolor and MindHive use variations of the scheme *d*. This circuitry was chosen in order to provide adjustable passive quenching and active recharge. In general, the quenching transistor is biased to ground so the resistance created is $\rightarrow \infty$ thus increasing the quenching slope to its maximum. The recharge transistor is controlled by a logic that can operate synchronously or asynchronously with the avalanche. This structure also enables what we call "virtual optical masking". Unlike other methods of masking where the bias is manipulated to prevent the SPAD from firing, in this case the SPAD is allowed to fire but it is never recharged; the recharge logic is connected to a 1-bit memory that has the masking information. Since the quenching resistor is very high, it takes very long time for the SPAD to recharge (millions of times longer than the frames used in the system so it is virtually masked).

Paralyzable and non-paralyzable detectors: this general concept is not restricted to MD-SiPMs but it rather applies to any kind of detectors used in many applications. The dead time of any detector is the time it takes for the device to become ready again after a hit. During this time, new events cannot be detected; then, leading to under-counting. The probability of missing a new event is related to the number of events per second the device is detecting. Here, it is required to distinguish two different cases. Non-paralyzable devices discard the events during their dead time and continue operating normally after they recover. On the other hand, paralyzable devices prolong their dead time if they receive new hits, both described in chapter 4 of [1]. The probability of missing a hit by non-paralyzable devices is given by Eq. 1.3.

$$P(X) = m\tau, \tag{1.3}$$

where X is variable "the detector is busy", m is the recorded event rate and τ

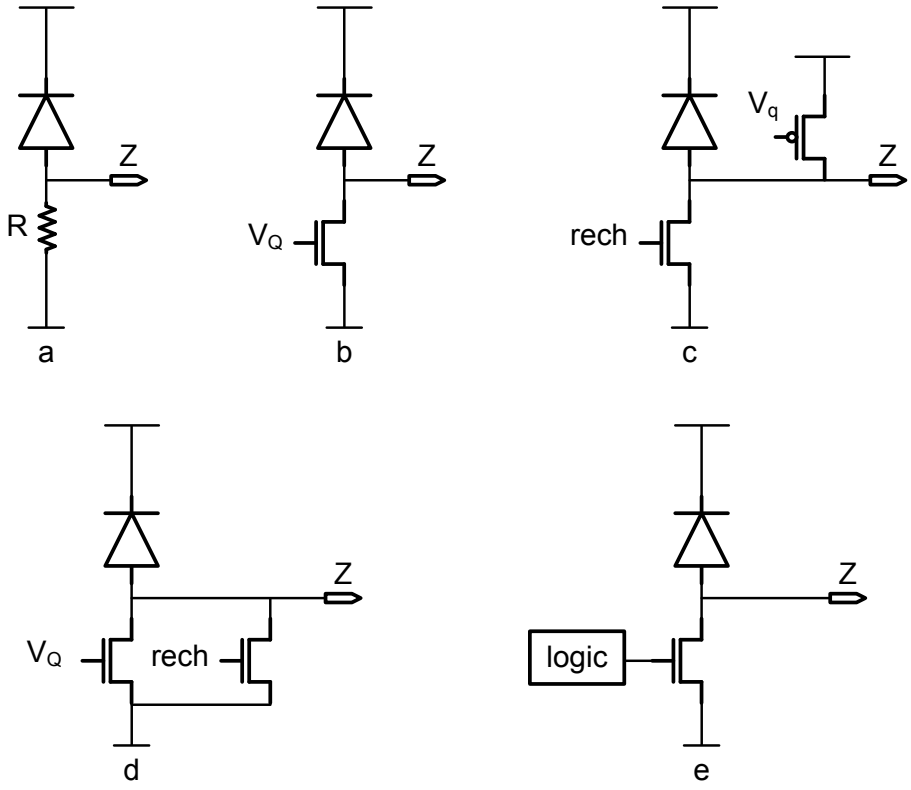


Figure 1.10: a) passive quenching and passive recharge, b) passive quenching and passive recharge using transistor-based resistor, c) active quenching, passive recharge, d) passive quenching active recharge, e) passive quenching and active recharge using one transistor.

is the dead time of the device. The number of missed hits is given by $q = nm\tau$, where n is the real value of photons hitting the sensor, q is the difference between the real number of photons hitting the sensor and the number of detected photons, namely $q = n - m$. Combining both expressions, it leads to Eq. 1.4.

$$n = \frac{m}{1 - m\tau}, \tag{1.4}$$

Fig. 1.11 shows the curve of this equation.

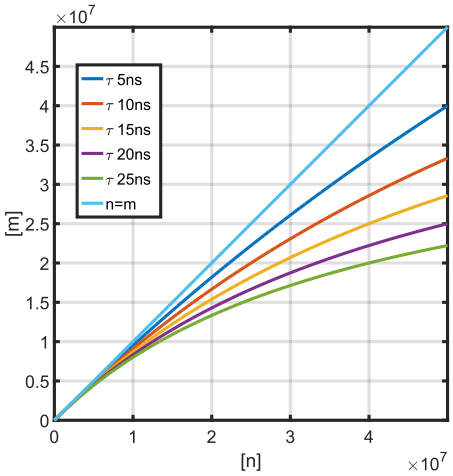


Figure 1.11: Model of non-paralyzable model for different dead times. n is the actual number of hits per second and m is the apparent number of hits per second.

For paralyzable devices, the missed hits q is calculated as $q = nP(Y)$ where $P(Y)$ is the probability of getting a hit within the dead time. To calculate $P(Y)$, it is required to know the probability density function of interval times between two events that is given by Eq. 1.5.

$$p(i) = ne^{-nt} \tag{1.5}$$

Then, $P(Y) = \int_{t=0}^{\tau} p(i).dt = \int_{t=0}^{\tau} ne^{(-nt)}.dt = 1 - e^{(-n\tau)}$

The rate of missed hits can be expressed as $q = n - m$. Therefore, the final expression is given by Eq. 1.6. Fig. 1.12 shows the representation of this theoretical curve.

$$m = ne^{(-n\tau)} \tag{1.6}$$

It is important to remark that, unlike non-paralyzable devices, the hit-counting starts decreasing in paralyzable devices as they go towards saturation. SPADs using active recharge have a much more abrupt slope than SPADs using passive recharge. The difference is that the passive recharge allows the SPAD to fire again yet not triggering the read-out buffer with the consequent paralyzation of the device. As

a consequence, SPADs with passive recharge under counts hits as the activity approaches saturation. This was another reason to have chosen active recharge for the designs presented here.

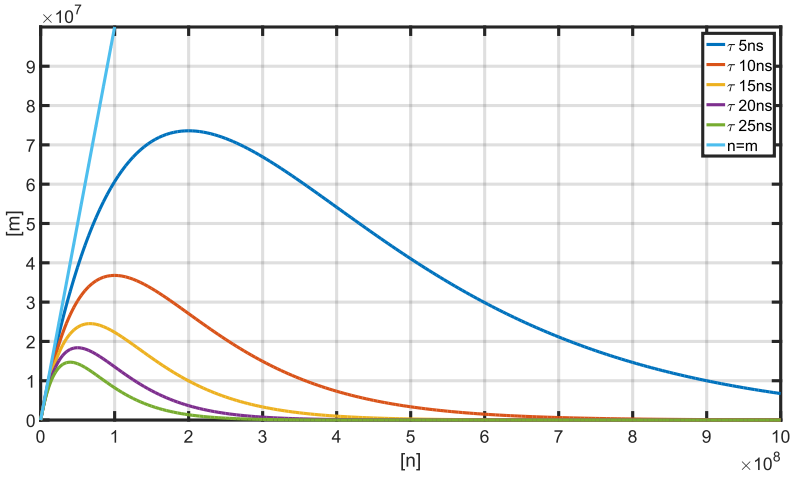


Figure 1.12: Simulation of paralyzable model for different dead times. Counting goes down as the detector approaches saturation. n is the actual number of hits per second and m is the apparent number of hits per second

1.3.3. Time-to-Digital Converters (TDC)

Generally speaking, Time-to-Digital converters are analog-to-digital devices utilized to measure and digitize time intervals. These time intervals can be represented by the width of a pulse or the time difference between two signals usually called *start* and *stop*. In many cases, they include a *reset* signal and a *latch* signal. See the scheme shown in Fig. 1.13. The output of a TDC is a N-bit word that, under a defined system coding, represents the measured time.

In image sensors, these devices are used to time-stamp events coming from the optical sensors, namely SPADs in this case.

TDCs normally include a coarse counter that counts the periods of an external reference, and a fine module that can count or detect the phases of the reference. The final count is a combination of these two counters. Whereas this is one common typical architecture of TDCs, many other types of TDCs have also been researched [21]. The main characteristics of TDCs are listed below.

Range: the maximum time that a TDC can count described in number of bits. The actual time normally depends on other constraints of the TDC and the frequency of the reference.

Resolution: this is the smallest time that the TDC can resolve. If the structure is coarse-fine counters, then the bin size will be the smallest delay, step or tap of

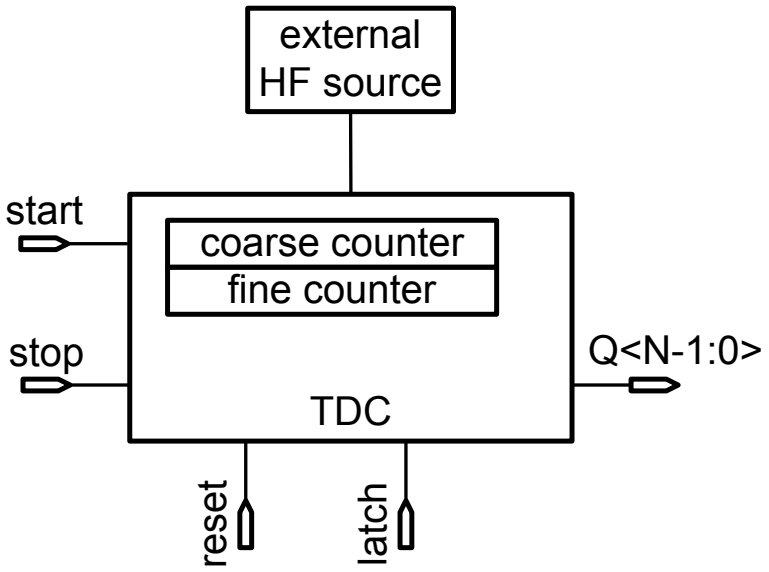


Figure 1.13: General scheme of a TDC. Although this is the most typical TDC for image sensors, there exist all kind of architectures, i.e. based on charge of capacitors or based on metastability of flip-flops.

the fine counter. With the current technologies, it is very common to reach 30ps, 20ps or even 10ps or less, as the TDCs introduced by [22].

Precision: non-uniformities in the layout, cross-talk between lines, limited noise-rejection to the power supply, thermal noise and many other effects will create many sources of jitter. The FWHM of the total jitter is the resolution of the TDC. It is the deviation with which a TDC can measure time.

Accuracy: is the quantity by which the TDC is predictably off the ideal time.

Time-of-Conversion (ToC): TDCs need time to make a full conversion and deliver the bits with the information. The shorter this time is, the higher the duty cycle of the TDC could be.

Differential non-linearity (DNL): the bins of the TDC have not all the same value and the difference affects the measurements of the TDC. DNL is calculated for every bin as

$$D(i) = \frac{T(i+1) - T(i)}{T_{ideal}},$$

where $T(i)$ is the time of the i_{th} bin and T_{ideal} is calculated as the average bin size of the system:

$$T_{ideal} = \sum T(i)/N,$$

where $\sum T(i)$ is the range of the TDC in seconds and N is the number of bins. The specifications of a TDC may include the maximum and the minimum of this set. DNL of TDC = $\{min(D) : max(D)\}$.

Integral non-linearity (INL): the accumulation of differential non-linearities leads to an integral non-linearity that can be calculated as $I(i) = \sum_{x=0}^{i} D(x)$. INL corresponds to the absolute shift between a bin and the actual value. The specifications of a TDC may include the maximum and the minimum values of this set. INL of TDC = $\{min(I) : max(I)\}$. Provided that the INL is monotonically increasing, the inverse of INL can be used as a LUT to correct the absolute value of the measurement.

Area: this specification is very important for image sensors as any area used for electronics might take space from SPADs; consequently reducing the sensitivity of the system. In 3-D systems the area used by TDCs might reduce the space for processing units.

Power: another important specification of TDCs. Image sensors have many TDCs along the chip and high power consumption might bring problems in the power distribution, heat generation, IR drops or, depending on the application, problems in the power distribution in the whole system. TDCs, oscillators and distribution of timing signals might have a significant impact in the final required power.

Trade-offs: as it happens to many components in general, it is very common to have to decide on some typical trade-offs that rise when designing TDCs. It is impossible in practice to maximize all the characteristics of this component. There some features that clash into direct competition; such as range vs. area, bin-size vs. power, and DNL vs. area. The rule of thumb is to enhance those features according to the application and to the device or phenomena to be measured by the TDCs, namely SPADs in this case. The characteristics of the SPADs should be taken into account at the moment of designing the TDCs of the system. For instance, knowing the jitter of the SPAD, the TDC should be designed to have an according bin size. If the bin is too wide, the jitter of the total system will be worsened. If it is too narrow, the TDC would be over-designed, thus taking more area and using more power than it should.

1.3.4. Architectures of MD-SiPMs

In this subsection, the reader will be walked through the main architectures and key points of MD-SiPMs. Fig. 1.14 shows the diagram of MD-SiPMs where the main components are:

Sensitive Matrix: a SPAD array of NxM forms the photon sensitive module of the MD-SiPM. SPADs and serving electronics are designed according to the specifications of the application. The shape of the matrix is usually rectangular but

other layouts are also common like a honeycomb in case the SPADs are rounded. Quenching, reset and recharge circuits, along with masking memory and read-out modules are attached to the SPADs. The matrix can be divided in columns or panels of different shapes to display specific features. A configuration module is desired to enable changes in the behavior of the matrix. Those can include: speed of quenching, type of reset, masking map, etc.

TDC bank: in order to time-stamp the events, MD-SiPMs are connected to a TDC bank that measures time of the events against a reference. In order to add flexibility to the system, the bank can usually be parametrized to change range, bin size, mode of operation (all-time running mode stand-by mode), frequency and other aspects encompassing different needs. Some degree of versatility in its configuration is required because the consumption of this unit can be a representative part of the total power budget.

Event discard unit: it is very likely that not all the events are important or useful for a given task. These non-useful events must be discarded at early stages as soon as they are marked as unimportant. This event-chopping technique can help to relax the communication constraints and lowering the workload of the processing unit. Better efficiency in terms of power and data volume can be achieved if this module is scattered along the sensor; the information that is not important can therefore be dismissed at the right moment so that the system can avoid its futile propagation (Fig. 1.15). For debugging purposes, the system should be able to shut off this unit at will.

Processing unit: all the information of the events, namely number of hits, location and time-of-arrivals, is processed in order to get the values that are interesting for the given application. This unit is not always located in one place, but rather is distributed along the chain to process the information as early as possible. As examples, we could mention Anger calculations for PET and histogram-on-chip for LiDAR. For debugging purposes, this unit should be able to provide raw information off the chip as well.

1.3.5. Interconnectivity and practical problems of MD-SiPM

SPADs-TDCs interconnection analysis

Ideally, it is desirable to have a TDC per SPAD, or an equivalent system such as multiple FIFOs and TDCs to time-stamp every event that occurs on the matrix. The interconnection scheme, area and power to achieve it, along with the low-duty time of the components, make this scheme highly inefficient. A good design of a MD-SiPM defines a proper interconnection between SPADs and TDCs to maximize the serving time and availability of the electronics while minimizing the *cost* function which includes area and power among others.

There are many techniques to design such interconnection. Fixed solutions might include a TDC bank of M TDCs serving a matrix of NxN SPADs [23]; thus,

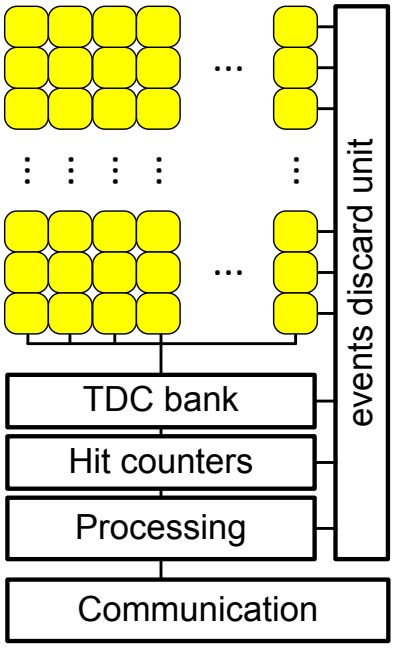


Figure 1.14: Diagram of generic MD-SiPMs. Components that are never missing in a MD-SiPM are SPAD arrays, TDCs, processing units and communication systems.

resulting in NxN/M SPADs per TDC. Dynamic solutions can redirect TDCs connections to SPADs as required. Other structures, as ping-pong TDCs or memories to increase availability are also possible. The relation between SPADs and TDCs will therefore not be one-to-one as the ideal case, but rather many-to-one or even many-to-many. This could create all sort of collisions that make the information of the address of the SPAD that triggered a given TDC be lost if especial electronics is not implemented.

This lack of full interconnection leads to a statistical problem that deserves to be analyzed. As SPADs are clustered to get access to a TDC to time-stamp the events, there are several shared segments of this path that can be busy at given point in time. The physical connectivity between SPADs and TDCs is called *timing line* and is extensively explained in chapter 3. Even when dynamic configurations of TDCs or memories are implemented, TDCs or timing lines can be busy when SPADs connected to them fire. In general, unless a specific circuit be included, the timing information of these events is lost.

The probability that an event is lost at a time t is:

$$P(t) = P_{TL}(t) \cup P_{TDC}(t) \cup P_{mem}(t),$$

where $P_{TL}(t)$ is the probability that the timing line is busy at a time t, $P_{TDC}(t)$ is the probability that the TDC is busy at a time t and $P_{mem}(t)$ is the probability that the memory is busy at a time t.

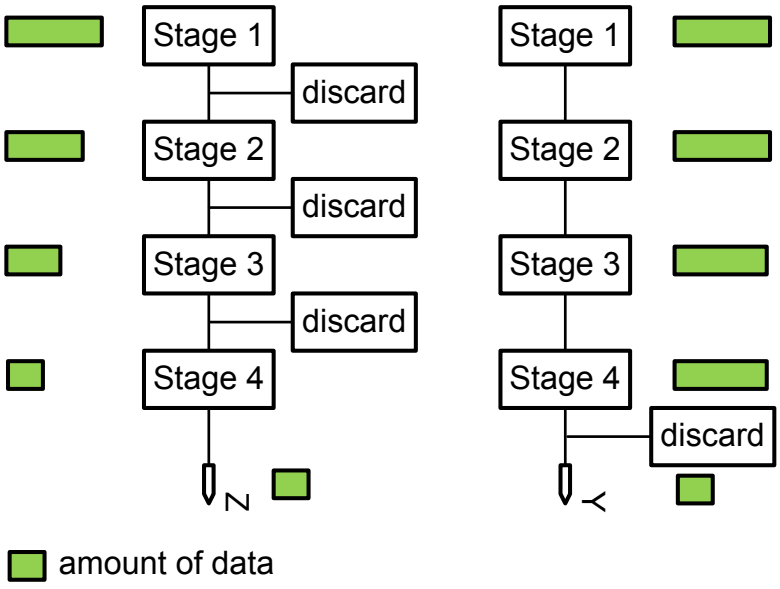


Figure 1.15: Data are progressively discarded as soon as they be found not to be useful (left). Data are filtered out at the end (right). The first strategy can save power and uses less throughput along the sensor. Additionally, the complexity of a module that discards events in the all-at-once strategy usually is much higher than in the scattered strategy.

These probabilities can be highly dependent and also other aspects of the modules should be considered like time-of-conversion, specific inter-group collisions or intra-group collisions. In order to simplify the problem, as first approximation we consider a point-to-point timing line with no delay so $P_{TL}(t) = 0$ and the memory is considered to have infinite depth with negligible delay, then $P_{mem}(t) = 0$. Time of conversion of the TDCs is neglected and TDCs can be used only once during the frame. As a result,

$$P(t) = P_{TDC}(t),$$

For the first photon, the occupancy of the TDCs is $O_{TDC} = 0$. For the second photon $O_{TDC} = \frac{1}{M}$. For the third photon $O_{TDC} = \frac{2}{M}$ or $O_{TDC} = \frac{1}{M}$ if the second photon hits the same group as the first. For the fourth photon $O_{TDC} = \frac{3}{M}$ or $O_{TDC} = \frac{2}{M}$ or $O_{TDC} = \frac{1}{M}$; this depends on how the previous photons hit the groups. In general, after p photons ($p < M$), the occupancy O can get any of the values of the set:

$$\left\{\frac{p}{M}; \frac{p-1}{M}; \ldots; \frac{1}{M}\right\}.$$

In the following, Fig. 1.16 shows the simulations of TDC occupancy for different value of the parameter of M. It is clear from the plot that a method to recover TDCs should be employed. Sharing and recovering TDCs during the frame is essential to keep available TDCs to catch new incoming events.

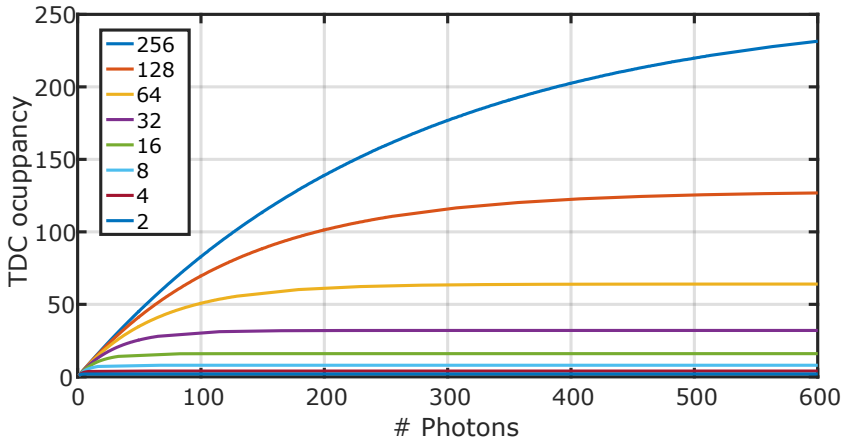


Figure 1.16: Saturation of TDCs with M as a parameter in case the TDCs can take 1 hit per frame. M is the number of TDCs that serve the whole array.

The most interesting part is the beginning of the curves, that are shown in Fig. 1.17. In some applications, specifically PET, the detection of the first photons impinging the detector is crucial for the performance of the system.

A concrete example is given for one of the chips designed in this work. MindHive has a matrix of 64x64 SPADs connected to 64 TDCs by groups of 8x8. In this case, the timing line occupancy will be also accounted. The timing line is shared by 8 SPADs in each group. The total memory the TDCs dump their information to have a depth of 4. Time of transport of timing lines is 1ns and time of conversion of the TDCs is 700ps. Dead time of the SPADs is 25ns. Since the theoretical analysis for this scenario is too complex and out of scope, we are showing only the simulations in Fig. 1.18 for different number of TDCs that are shared along the SPAD matrix.

There are some observations to be made about this simulation. This plot should be analyzed in three portions. For relatively low activity A, (lower than 1 Gcps) all the delay times, recovery times and transport times are negligible so the maximum capacity of the sensor in a given frame is equal to the depth of the memory multiplied by the number of clusters ($4 \cdot 64 = 256$). For this reason, the availability is close to one for cases where the number of photons is much less than 256 and in this portion, the availability almost does not depend on the activity. As the number of photons increases, the availability tends to $256/N$ which is the capacity of the full memory divided the number of photons. For very high activity (A > 1T cps), the electronics is too slow; thus no module can recover on time to get a second hit. For this reason, the electronics becomes one-time-use in a given frame and the availability turns constant. With moderate activity, in the middle portion of the plot, it can be observed that the activity starts to reach the point where the electronics delay begins to play a crucial role. In this case, the more activity there is, the more hits will be missed.

This plot gives a powerful insight. The expected activity can drive the design

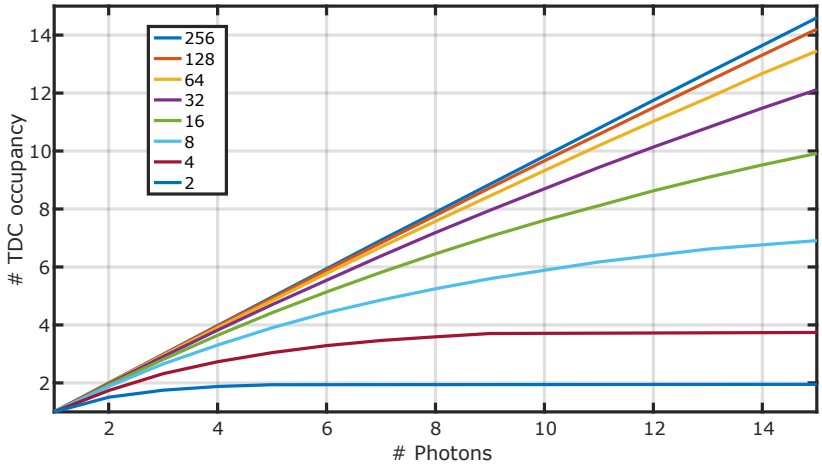


Figure 1.17: Zoom of saturation of TDCs with M as a parameter. In order to avoid large data loss, TDCs must always work in this region.

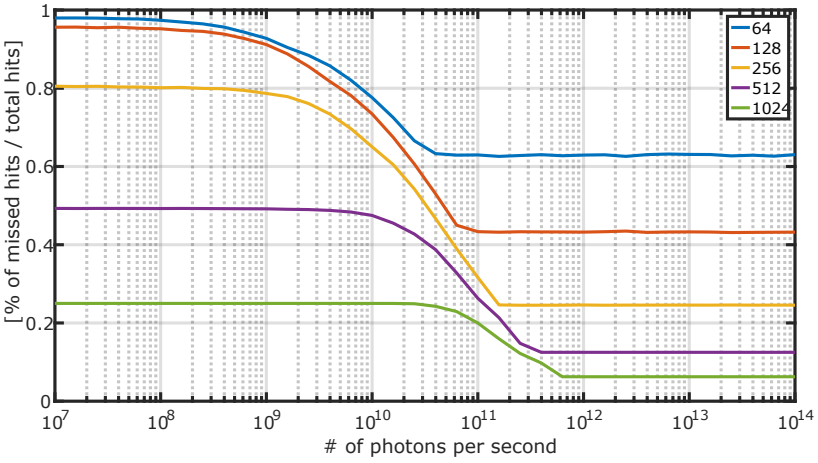


Figure 1.18: Saturation of the TDCs in MindHive as a function of activity and number of photons impinging the sensor as parameter. The number of photons simulated is constant in the considered frame.

in order to make the sensor operate in a proper region. For example, a very fast electronics with low ToC in a sensor that always works in the region 1 can be a waste or resources.

Total jitter of the system: the time-marks of every detected event are generated by the TDCs of the system. The time resolution or jitter of these time-marks is affected not only by the SPAD response to the avalanche and the TDC resolution, but as well by all the components that are in the whole chain from the moment the photon hits the sensitive area and the moment the TDC converts the time to generate the time-stamp. The time chain is composed by:

- SPAD.
- Timing line: all the intermediate modules that bring the signal from the SPAD to the TDC.
- TDC: module to measure the time.

$$J_T \propto \sqrt{J_{SPAD}^2 + J_{TL}^2 + J_{TDC}^2},$$

where J_{TL}^2 is the jitter introduced by the timing line and jitter is defined as the standard deviation of the underlying random variable associated with the measurement.

SPAD Matrix saturation:

The SPAD matrix of the sensor suffers from a very similar problem as TDC banks: high activity can saturate the matrix; resulting then in under-counting. As explained at the beginning of this section, SPADs have a dead time during which they cannot detect another hit. Furthermore, there are many effects that, all combined, can make the system miss photons. The most important causes are listed below.

- SPAD dead time.
- SPAD recharge method: dead times can be longer if global shutter is used.
- Hit counters saturation: counters can be saturated either by short inter-arrival time or maximum count.
- Saturation of hit lines: they have a maximum hits per second of operation that will chop high activity.
- SPADs clustering: shared resources might not be available at all times.
- Read-out speed: the hits might not be read on time to get new hits.

There are systems that recharge the SPADs at the end of the frame with a global recharge signal and other systems that recharge the SPADs as soon as they fire. In the first case, there is a static saturation that tightly depends on the number of SPADs in the matrix; while in the second case the saturation is more related to the instantaneous activity. Both cases are analyzed here.

1

Frame-based systems using global recharge signal To analyze this case, we consider a SPAD matrix of NxN SPADs which are not recharged during the frame; they therefore become one-time-use. The matrix is considered ideal; no DCR, no background light or screamers are then considered. For the first arriving photon, the occupancy of the matrix is $0/N$. For the second photon, the occupancy is $1/N$. For the third photon, the occupancy is $2/N$ or $1/N$ if the second photon hits the same SPAD as the first one. For the fourth photon, the occupancy is either $3/N$, $2/N$ if two of the previous photons hit the same SPAD, or $1/N$ in case all the previous photons fell into the same SPAD. In general, for p photons the occupancy follows a binary tree. The level of occupancy can easily be transformed into a probability tree. Fig. 1.19 shows a portion of a tree for the photon i to the photon i+1.

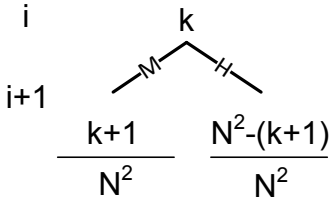


Figure 1.19: Tree that shows the probability of hit (H) and miss (M) for the $(i+1)_{th}$ photon for sensor of NxN SPADs and number of hits equal to k.

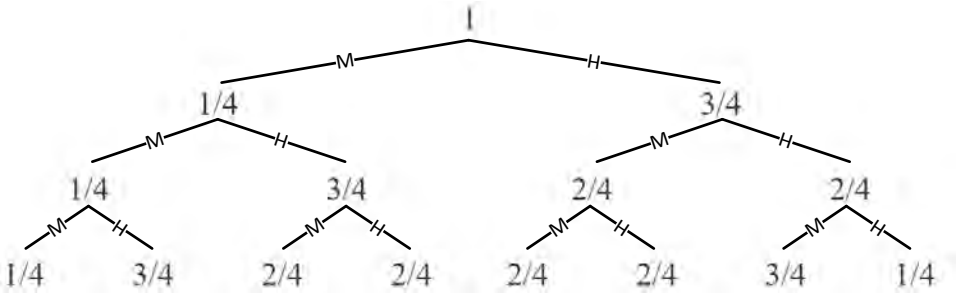


Figure 1.20: Tree that shows all the probabilities of a hit (H) when a photon is detected by an unused SPAD and misses (M) when a photon hit a SPAD that had already triggered. The i_{th} level of the tree shows the conditional probability of the i_{th} photon.

By using this small probability tree, it is possible to find the probability of detection of H hits in a matrix of NxN, provided F photons hit the sensor, described by Eq. (1.7). As an example, a tree for a system $N = 2$ that is shown in Fig. 1.20.

$$P(H,F) = \frac{1}{N^{2(F-1)}} \prod_{i=1}^{H-1} (N^2 - i) \sum_{\forall k_i} (1^{k_1} 2^{k_2} 3^{k_3} \ldots H^{k_I}), \tag{1.7}$$

where the constants k are such that

$$k_1 + k_2 + k_3 + ...k_H = H.$$

Fig. 1.21 shows an example where $N^2 = 9$; F photons randomly impinge the MD-SiPM. Equation 1.7 was implemented in MatLab to see how the occupancy grows as the number of photons increases and how it causes saturation to a NxN sensor. The equation is accompanied with a MonteCarlo simulation, also written in MatLab.

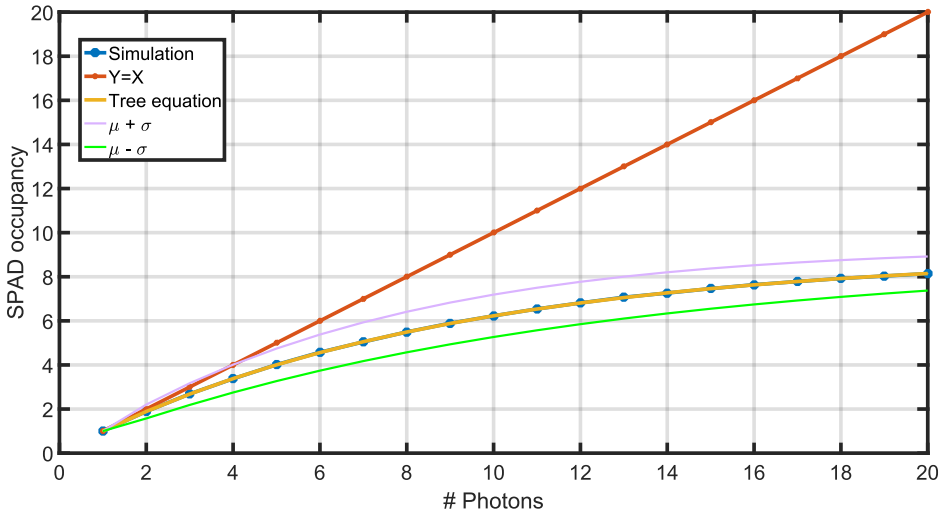


Figure 1.21: The curve Y=X is the detection curve of an ideal sensor with no dead time in its SPADs. The theoretical curve and MonteCarlo simulation give the same result. Two extra curves are plotted, showing one standard deviation above and below.

As a second example, a MonteCarlo simulation has been performed for a case where the sensor has NxN = 1024 SPADs. Results are shown in Fig. 1.22.

Some works use the complete SPAD array as one-time-use [12] and some others keep on resetting the SPAD array in order to alleviate the saturation problem [24]. However, these equations and simulations still remain for the time between two pulses of reset signals. In that case, the frame is considered between two reset pulses. Full dynamic process has been analyzed in the previous section, where there is no frame, and the dead times of the electronics in the sensor have been accounted. These plots can be used to correct saturation for a certain measurement in order to mitigate the loss in linearity. Similar results have been found in [19].

1.4. Implementations' aspects

The aforementioned applications, LiDAR and PET, have many aspects that should be taken into account when searching for concrete implementations. In case of considering MD-SiPMs for these systems, time resolution and intensity detection

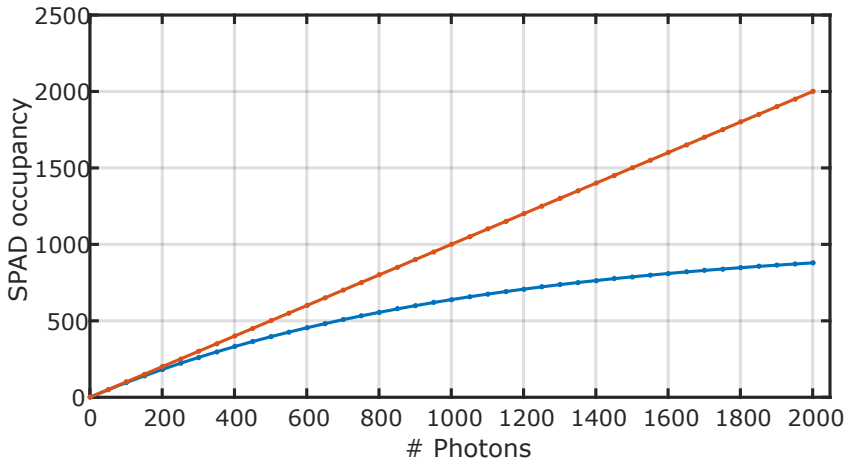


Figure 1.22: Saturation of a SPAD matrix with 1024 SPADs in blue. Ideal system response in red.

play a decisive role in the final outcome. The critical features and methods are discussed in the following.

1.4.1. LiDAR

Map reconstruction

The system captures all the photons coming back from the scene and, as a result, a list of points (x;y;z) is obtained. The ways these points are used to reconstruct the scene are varied. There are many methods to perform the map reconstruction. While some of them are simple and can be done on chip taking reasonable area, other methods that are very complex and might include other type of sensors, are calculated off chip. Here, the main methods are showed and analyzed.

Peak detection: this method is one of the simplest ones and perform well in basic environments. The sensor is split in its natural grid of NxN pixels and for every pixel a histogram of Z is built. As a result, a matrix of NxNxM elements (M being the range of the TDC) is obtained. A function to detect the peak in the histogram for every pixel can be used and an image is reconstructed out of this information, as shown in the following.

function DetectPeak(int $histogram[]$)
  $maxvalue.amp = 0$;
  $maxvalue.z = histogram.length - 1$;
  for $z = 0 \rightarrow histogram.length - 1$ **do**
    if $histogram[z] > maxvalue.amp$ **then**
      $maxvalue.amp = histogram[z]$;
      $maxvalue.z = z$;
    end if

```
    end for
    return maxvalue;
end function
function Reconstruct(int VoxelMatrix[][][])
    for x = 0 → VoxelMatrix.length(0) do
        for y = 0 → VoxelMatrix.length(1) do
            image[x][y] = DetectPeak(VoxelMatrix[x][y][]);
        end for
    end for
    return image;
end function
```

The initial values of the function "Reconstruct()" have to be 0 for the amplitude and it should be the range of the TDCs for the distance. In this thesis, the peak detection technique was used as first approach. In some works [25], the histograms are calculated on chip and only the result is transmitted. It can achieve high speed because only part of the information is transmitted at the cost of large extra area utilization.

Enhanced Peak detection or convolution techniques: there could be some situations where the sole peak detection does not fit well such in borders of object or multiple reflections that might create echo images. In this case, a small window of 3x3 can be used to generate a post-processed image that correct discontinuities in the voxels. For instance, the following matrix can be used to soften edges by convolution.

$$M = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} /16$$

then every pixel of the new image will be:

$$I_{new}(x, y) = I(x, y)$$

```
function SoftenEdges(int image[][])
    n = 3;
    for x = 0 → image.length(0) do
        for y = 0 → image.length(1) do
            newimage[x][y] = conv(SubMatrix(image, x, y, n), M);
        end for
    end for
    return newimage;
end function
```

where SubMatrix() gets a sub-matrix nxn of elements of M centered in (x;y). Sometimes, more than one peak could be detected in the same histogram, and every measurement will alternate between them. In this case, it is possible assign

the minimum of a sub-matrix to that point which corresponds to the closest object. Possible matrices used in this case could be:

$$M_1 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, M_2 = \begin{pmatrix} \infty & 1 & \infty \\ 1 & 1 & 1 \\ \infty & 1 & \infty \end{pmatrix}$$

And the function to generate the post-proccessed image is:

function SoftenEdges(int $image[][]$)
  $n = 3$;
  for $x = 0 \rightarrow image.length(0)$ **do**
    for $y = 0 \rightarrow image.length(1)$ **do**
      $tempmatrix = prod(SubMatrix(image, x, y, n), M)$;
      $newimage[x][y] = Min(tempmatrix)$;
    end for
  end for
  return $newimage$;
end function

Notice that M can have different shapes that will be related to the optics of the system.

Simultaneous Localization And Mapping (SLAM): more complex scenes can generate multiple-reflections paths; thus, creating several peaks that depend on the angle and position of the sensor. In these conditions, the peak-detection technique will more certainly not be sufficient to process the data and generate the image. Using more aggressive convolution methods might only deteriorate the quality of the image and lose details. For these complex situations and, furthermore, when more than one sensor is used for mapping, SLAM is the preferred technique [26]. SLAM uses algorithms to assign a given probability to every point in the image by accounting for the uncertainty of the sensors employed in the system. In this way, every point acquired by the sensors are represented as (x;y;z) + (Δx;Δy;Δz). This probabilities are assigned and computed into a voxel. SLAM techniques could be computationally costly when the uncertainty of the sensors is high; hence, it is very important to achieve good resolution in x, y and time. In chapter 4, this technique is used to generate a 3-D map using a sensor that was not specifically designed for LiDAR but for many applications. Concolor measures (x;y;z) for every hit; however, collisions are likely to happen since TDCs are shared per semi-column. As a result, information of the x-axis and y-axis is kept (the number of triggered TDC) but the information of the z-axis might have large uncertainty as hits are not assured to be in a particular pixel. This uncertainty along with the time resolution are expressed for every point in the following way (position uncertainties in x and y axis are considered negligible):

$$point(x; y; z) = (x_h, y_h, z_h) + (0, 0, \Delta z),$$

where x_h, y_h and z_h are the values of each hit.

All this hit information is stored as probabilities in a voxel matrix in combination with flash technique to get a 3-D image. Fig. 1.23 shows a plausible hit map of a system for a given frame. For this example, the system has 8x8 SPADs that are connected to 16 TDCs, shared by each semi-column. Only one of the connections of the TDCs is shown in the picture for clarity.

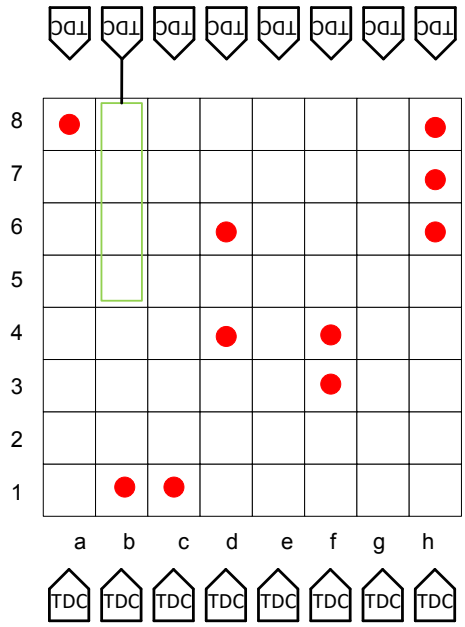


Figure 1.23: Diagram of hits on an imaginary 8x8 image sensor whose SPADs are served by 16 TDCs grouped as shown. Only the connection for the second top TDC is shown for sake of clarity.

For the hit in a8, x, y and z axis are perfectly defined as it is the only hit in the semi-column. The same can be said about the hits b1 and c1 and the hits in d6 and d4. The situation is different for the hits in f3 and f4 since they share the TDC. X and y positions are perfectly defined but the position in the z axis is not. This multiple hit in the TDC, known as collision, can be automatically discarded or, alternatively, can be accounted for both the pixels with a probability $P(x, y, z) = \frac{1}{2}$. In the case of the triplet in h6-8, the algorithm applies the same technique, assigning $P(x, y, z) = \frac{1}{3}$. In general, for a multiple hit the algorithm assigns $P(x; y; z) = \frac{1}{H}$, H being the number of simultaneous hits. These probabilities are added up into a matrix and after a number of predefined frames, a probability estimator is used to obtain the final image. Mean, media, min and max estimators or even the Maximum likelihood estimator can be used. Results are shown in chapter4.

1

1.4.2. PET

Time resolution: if the system held infinite time resolution, x, given by $x = c\frac{t_0 \dot{-} t_1}{2}$, would be the exact location of the annihilation. Since sensors have finite time resolution that responds to Gaussian distribution, the uncertainty of x is given by:

$$\sigma^2(x) = VAR(\frac{c}{2}(t_0 - t_1))$$

and since t_0 and t_1 are both independent Gaussian variables,

$$\sigma^2(x) = \frac{c^2}{4}(\sigma^2(t_0) + \sigma^2(t_1))$$

In systems where all the detectors are of the same kind, the uncertainty of x will then be:

$$\sigma(x) = \frac{c}{\sqrt{2}}\sigma(t)$$

In PET systems, the uncertainty is usually expressed as Full-Width at Half-Maximum (FWHM) that is 2.2 times the standard deviation, then:

$$FWHM_s = \frac{c}{\sqrt{2}}FWHM_t$$

.

This equation is very important, as it explains how the resolution of the detectors is directly translated into spatial resolution. For instance, if the detectors have a time resolution of 200 ps, the spatial resolution will be 4.2 cm. A graphical representation of LoRs is shown in Fig. 1.24.

The higher the time resolution of the detectors is, the better the resolution in the estimation of the position of the annihilation will be; thus, giving more precise data to the reconstruction algorithms to create an image with better quality.

Timing parameters for PET applications: there are parameters of image sensors that are very important when they are used for PET. Single-Photon Time Resolution (SPTR) specifies the time resolution (FWHM) when the sensor is hit in any part of it by only one visible photon. On the other hand, Coincidence Time Resolution (CTR) is the resolution (FWHM) of the system when two gamma photons hit two detectors.

Statistics on time-stamps: although SPTR is the time resolution of the detector, by applying statistics it is possible to get a better time resolution when more photons are used. In chapter 4, four different methods are used to improve the Multi Photon Time Resolution (MPTR). The more photons are used for the estimation, the higher the time resolution will be. In the experiment, a laser triggered by the system was used. For example, the average method simply takes the average of all the photons $\mu = \frac{\sum t_i}{N}$. Then, the resolution $s = \frac{1}{N-1}\sum_{i=0}^{N}(t_i - \mu)$.

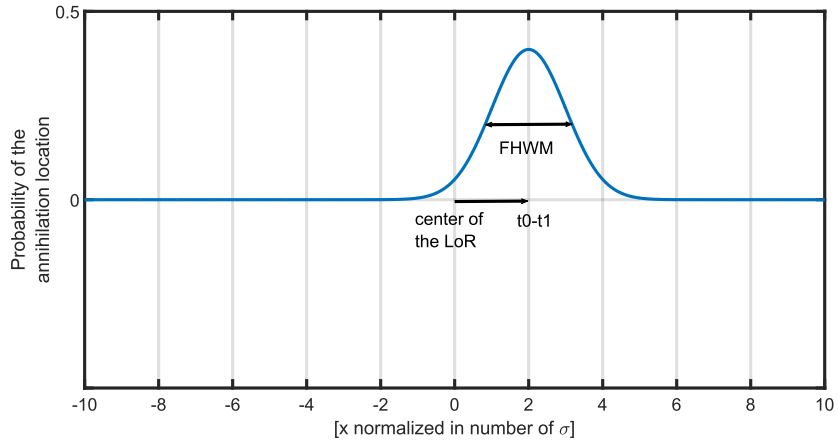


Figure 1.24: Example of a LoR when $t_0 - t_1 = 2$. The x axis is the normalized position in number of standard deviations respects to the center of the LoR. y axis is the probability density function of the annihilation location.

However, the situation when a scintillator is coupled to the sensor is a little bit different. The photons coming from a scintillator do not have all the same variance in their time-of-arrivals. The photons arrive following a Poisson distribution shown in Fig. 1.25.

There are many methods used to estimate the time-of-arrival of the gamma photon [27]. The minimum standard deviation achievable by an unbiased estimator is dictated by Cramér-Rao inequality that is shown as follows:

$$VAR(\hat{\theta}) \geq \frac{1}{nE\left[\left(\frac{\delta L(x;\theta)}{\delta\theta}\right)^2\right]},$$

where $L(x;\theta)$ is the logarithm of the likelihood function and E is the Expectation operator. Fig. 1.26 shows a simulation of time resolution for two different methods (average method and first-photon method). The jitter of the detector considered was 500ps, PDE and coupling factor were combined into a sole optical gain of $G = 0.1$. The PDF for photons of the scintillator is given by:

$$f(t) = \frac{-e^{-t/\tau_r} + e^{-t/\tau_d}}{\tau_d - \tau_r},$$

and the cumulative probability function is

$$F(t) = \frac{-\tau_r e^{-t/\tau_r} + \tau_d e^{-t/\tau_d}}{\tau_d - \tau_r} + K,$$

where τ_r is the rise time of the scintillator, τ_d is the decay time of the scintillator and K is a constant such as $\lim_{t\to\infty} F(t) = 1$.

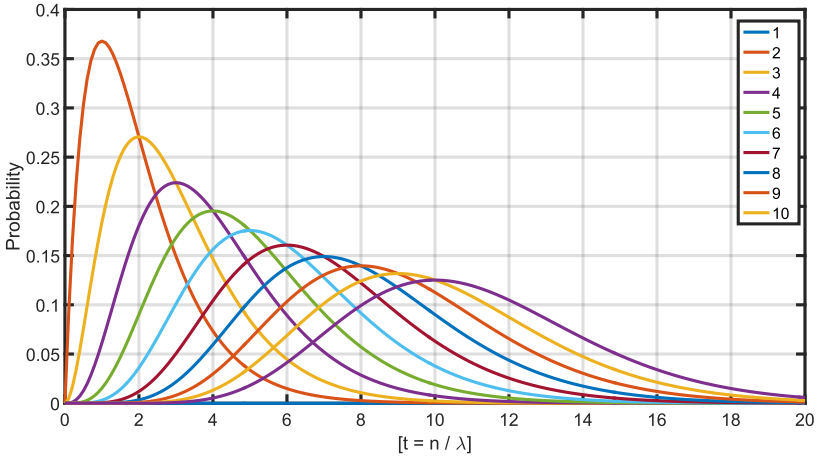


Figure 1.25: Distribution of the first 10 photons that follow Poisson distribution.

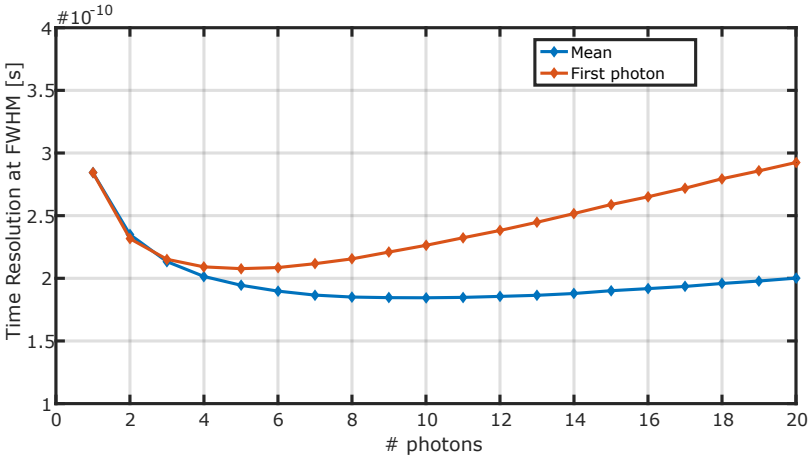


Figure 1.26: Simulation of CTR of a system (FHWM) as function of the number of used photons for two different methods: First-Photon and Mean.

The optical gain (G) of the detector can also have a large impact in the time resolution as the first photons are known to be the ones with the least variance and they therefore lead to a better time resolution as explained in [28].

Two gamma photons hitting the detectors does not ensure the LoR described by them is a straight line that contains the annihilation position. Scatter is a process where a gamma photon partially deposits its energy in the medium it is going through and gets deviated. The equation that governs this process is called Compton scatter and is as follows:

$$\Delta\lambda = \frac{h}{m_e c}\left(1 - cos(\theta)\right),$$

where $\Delta\lambda$ is the change in the wavelength of the gamma photons when it experiences Compton scatter, h is Plank's constant, c is the speed of light, m_e the mass of the electron and θ is the angle of deviation in the trajectory of the gamma photon respects to the trajectory before the collision.

The energy of the gamma photon gets reduced by an amount given by:

$$\frac{E_1 - E_0}{E_0} = \frac{\Delta E}{E_0} = \frac{hc}{\lambda_1} - \frac{hc}{\lambda_0},$$

then

$$\frac{\Delta E}{E_0} = \frac{hc}{m_e}\frac{1 - cos(\theta)}{\lambda_0(\lambda_0 + hc/m_e(1 - cos(\theta))},$$

Fig. 1.27 shows the shape of this equation.

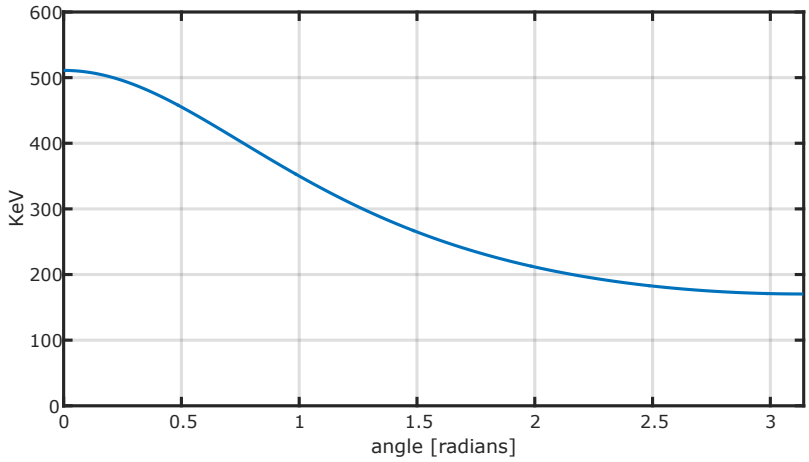


Figure 1.27: Energy of gamma photons after a Compton scattering with an electron. Maximum loss of $2/3$ occurs when the photon is reflected 180 degrees.

It is interesting to observe that the maximum energy transferred to the material where the scatter occurs is 2/3 of the total energy. If a single scatter occurs in the crystal, the maximum energy deposited will then be 337KeV.

A typical energy spectrum of a scintillator, when is used to detect a ^{22}Na source, is shown in Fig. 1.28, where it is possible to distinguish different particular points, explained in the following:

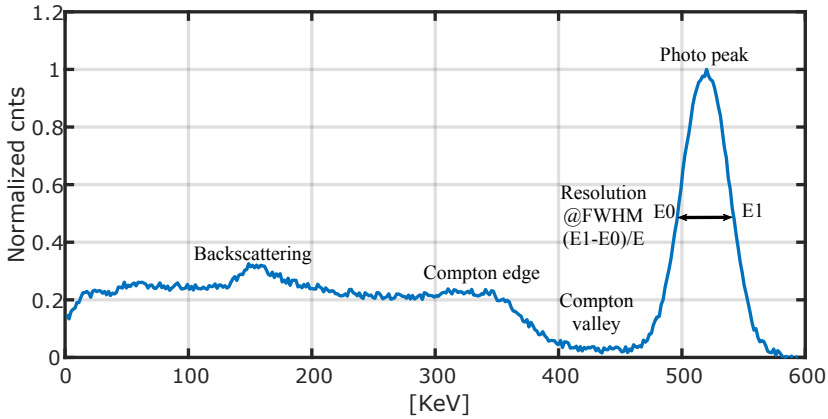
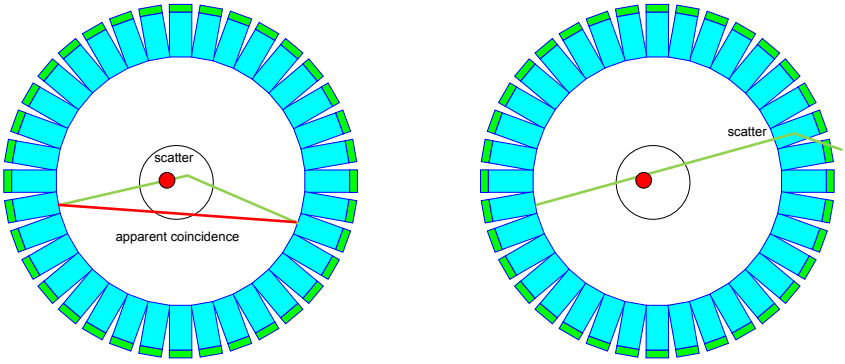


Figure 1.28: Typical energy spectrum of a scintillation NA Iodine crystal for a ^{22}Na source. Figure adapted from [29]

- The photo peak of ^{22}Na is centered at 511 keV. If the detector were ideal, all the counts of the photo peak should be exactly at 511KeV. The spread is caused by the limited resolution of the system. Energy resolution usually is specified in percentage $\frac{\Delta E}{E}$ at FWHM. Typical values are in the range of 10% to 20%.
- Compton interactions in the crystal deposit at most $2/3$ parts of the total energy of the Gamma photon. This creates an edge in the spectrum called "Compton edge", after which, the "Compton valley" is observed .
- Compton interactions in external objects lose energy that goes from 0 to, at most, 2/3 parts of the total energy of the gamma photon. The energy of the Gamma photon that then will be deposited to the crystal is $E_d = 511keV - max(E_L)$ where E_L is the energy lost in the interaction with the material. Therefore, $E_d < \frac{1}{3}511KeV$ and it corresponds to the backscattering point that stands out due to the probability of its occurrence.

The energy resolution of the detector will impact in the final quality of the reconstructed picture. When scattering occurs, either in the patient's tissue or in the crystal, the trajectory of the gamma photon changes and its energy drops. This change in energy can be measured and this "false event" is discarded. Fig. 1.29 shows the aforementioned effects. Hence, the energy resolution of the detector has



(a) One of the gamma photons scatters in the patient's tissue and therefore creates a false coincidence marked in red.

(b) Gamma photons can also scatter inside the crystal, thus creating an undistinguishable situation from the case where it is scattered inside the patient.

Figure 1.29: False coincidences happen when two simultaneous detections mislead to a wrong LoR due to single or multiple interactions that cannot be accounted by the system.

to be enough to be able to distinguish a true event from a false event coming out of scatter interactions. The energy resolution of the detector is the convolution between the resolution of the scintillator and that of MD-SiPMs, explained previously in this introduction.

1.5. Organization of this thesis

This work is divided in three major sections to set the grounds for the next generation of image sensors. The first section explains the main modules of image sensors and all the electronics associated with them. Theory, practice and implementation is shown for each module along with measurements and examples. The section comprises the chapters 2 and 3. Once the framework to understand image sensors has been set, then the section II, comprising chapters 4 and 5, explains, shows and demonstrates specific implementations for the main two applications addressed in the work which are PET and LiDAR. Finally, the section III (chapter 6) introduces the last generation of image sensors and gives insights about how the future of image sensors could be. Results about the implementation of these new ideas and concepts are also shown in the same section.

1.6. Thesis contributions

The main goal of this thesis is to bring new concepts, ideas and specific implementations to pave the way to the next generation of image sensors. This goal is tackled in two different ways.

1.6.1. Quantitative improvement

In chapter 2 and chapter 3, the electronic framework for image sensors is thoroughly explained. The content of chapter 4 and chapter 5 is focused on applications. This

thesis gives new ideas about architecture and components to quantitatively improve the features of image sensors:

TFIFO memories: a new architecture was introduced by this work to overcome the main drawbacks of SRAM memories when they are operated for FIFO applications. This idea has been implemented and proved to work. Results are shown in chapter 2.

In-pixel AC-coupled amplifiers: a novel design and implementation have been done to achieve per-pixel signal amplification. Results are shown in chapter 3.

Sliding-Scale TDCs for image sensors: this thesis shows a new generation of TDCs that, by using sliding-scale technique, can drastically reduce DNL and INL of the entire system.

1.6.2. Qualitative step

This work is also focused on changing the paradigm of the way we design image sensors. New concepts and ideas are shown to enable the next qualitative step for smart SPAD sensors.

Read-out schemes: the idea of generic read-out systems for image sensors was introduced. Different schemes and components to perform the read-out of pixels, TDCs, registers and memories have been designed and proved to work for different types of applications. Reutilization is key to quickly move on to next versions of sensors.

Smart SPAD sensors: new sensors will most likely have aid from artificial intelligence to process information. The last chapter (6) fully explains this concept and shows the first implementation of a neural network in a MD-SiPM that can process data coming from the SPADs in order to solve problems that used to belong only to human domain.

References

[1] G. F. Knoll, *Radiation Detection and Measurement, 3rd ed.*, 3rd ed. (John Wiley and Sons, New York, 2000).

[2] H. P. K. K., *Photomultipliers Tubes, basics and applications* (Hamamatsu Photonics K. K. Electron Tube Division, February 2006).

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

[4] M. Perenzoni, D. Perenzoni, and D. Stoppa, A 64x64-pixel digital silicon photomultiplier direct ToF sensor with 100Mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14spacecraft navigation and landing, in *2016 IEEE International Solid-State Circuits Conference (ISSCC)* (2016) pp. 118–119.

[5] C. Niclass, M. Soga, H. Matsubara, S. Kato, and M. Kagami, A 100-m Range 10-Frame/s 340 $\times$96-Pixel Time-of-Flight Depth Sensor in 0.18-μm CMOS, IEEE Journal of Solid-State Circuits **48**, 559 (2013).

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

[7] Mike1024, Lidar, Available at https://en.wikipedia.org/wiki/Lidar#/media/File:LIDAR-scanned-SICK-LMS-animation.gif (2020/04).

[8] K. Ito, C. Niclass, I. Aoyagi, H. Matsubara, M. Soga, S. Kato, M. Maeda, and M. Kagami, System Design and Performance Characterization of a MEMS-Based Laser Scanning Time-of-Flight Sensor Based on a 256 $\times$ 64-pixel Single-Photon Imager, IEEE Photonics Journal **5**, 6800114 (2013).

[9] M. Beer, O. M. Schrey, C. Nitta, W. Brockherde, B. J. Hosticka, and R. Kokozinski, 1×80 pixel SPAD-based flash LIDAR sensor with background rejection based on photon coincidence, in *2017 IEEE SENSORS* (2017) pp. 1–3.

[10] A. S. Andrew Murphy, Positron emission tomography, Available at `https://radiopaedia.org/articles/positron-emission-tomography` (2020/04).

[11] P. E. V. D. L. Bailey, D. W. Townsend and M. N. Maisey, *Positron Emission tomography* (Springer, 2005).

[12] A. Carimatto, S. Mandai, E. Venialgo, T. Gong, G. Borghi, D. R. Schaart, and E. Charbon, 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.

[13] S. Cova, A. Longoni, and G. Ripamonti, Active-Quenching and Gating Circuits for Single-Photon Avalanche Diodes (SPADs), IEEE Transactions on Nuclear Science **29**, 599 (1982).

[14] S. Pellegrini, B. Rae, A. Pingault, D. Golanski, S. Jouan, C. Lapeyre, and B. Mamdy, Industrialised SPAD in 40 nm technology, in *2017 IEEE International Electron Devices Meeting (IEDM)* (2017) pp. 16.5.1–16.5.4.

[15] F. Nolet, S. Parent, N. Roy, M.-O. Mercier, S. A. Charlebois, R. Fontaine, and J.-F. Pratte, Quenching Circuit and SPAD Integrated in CMOS 65 nm with 7.8 ps FWHM Single Photon Timing Resolution, Instruments **2** (2018), 10.3390/instruments2040019.

[16] F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, and I. Rech, 83-ps Timing Jitter With a Red-Enhanced SPAD and a Fully Integrated Front End Circuit, IEEE Photonics Technology Letters **30**, 1727 (2018).

[17] A. C. Ulku, C. Bruschini, I. M. Antolović, Y. Kuo, R. Ankri, S. Weiss, X. Michalet, and E. Charbon, A 512 × 512 SPAD Image Sensor With Integrated Gating for Widefield FLIM, IEEE Journal of Selected Topics in Quantum Electronics **25**, 1 (2019).

[18] K. Morimoto, A. Ardelean, M.-L. Wu, A. C. Ulku, I. M. Antolovic, C. Bruschini, and E. Charbon, Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications, Optica **7**, 346 (2020).

[19] I. M. Antolovic, C. Bruschini, and E. Charbon, Dynamic range extension for photon counting arrays, Opt. Express **26**, 22234 (2018).

[20] M. Moreno-García, R. del Río, . Guerra, and . Rodríguez-Vázquez, 5×5 SPAD matrices for the study of the trade-offs between fill factor, dark count rate and crosstalk in the design of CMOS image sensors, in *2014 10th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)* (2014) pp. 1–4.

[21] D. P. Palubiak and M. J. Deen, CMOS SPADs: Design Issues and Research Challenges for Detectors, Circuits, and Arrays, IEEE Journal of Selected Topics in Quantum Electronics **20**, 409 (2014).

[22] V. Sesta, F. Villa, E. Conca, and A. Tosi, A novel sub-10 ps resolution TDC for CMOS SPAD array, in *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)* (2018) pp. 5–8.

[23] R. K. Henderson, N. Johnston, S. W. Hutchings, I. Gyongy, T. A. Abbas, N. Dutton, M. Tyler, S. Chan, and J. Leach, 5.7 A 256×256 40nm/90nm CMOS 3D-Stacked 120dB Dynamic-Range Reconfigurable Time-Resolved SPAD Imager, in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)* (2019) pp. 106–108.

[24] L. H. C. Braga, L. Gasparini, L. Grant, R. K. Henderson, N. Massari, M. Perenzoni, D. Stoppa, and R. Walker, An 8×16-pixel 92kSPAD time-resolved sensor with on-pixel 64ps 12b TDC and 100MS/s real-time energy histogramming in 0.13µm CIS technology for PET/MRI applications, in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers* (2013) pp. 486–487.

[25] 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 **54**, 1137 (2019).

[26] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, MonoSLAM: Real-Time Single Camera SLAM, IEEE Transactions on Pattern Analysis and Machine Intelligence **29**, 1052 (2007).

[27] E. Venialgo, S. Mandai, and E. Charbon, Time mark estimators for MD-SiPM and impact of system parameters, in *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC)* (2013) pp. 1–2.

[28] M. Fishburn and E. Charbon, System Tradeoffs in Gamma-Ray Detection Utilizing SPAD Arrays and Scintillators, Nuclear Science, IEEE Transactions on **57**, 2549 (2010).

[29] L. Didactic, Lidar, https://www.ld-didactic.de/software/524221en/Content/Appendix/ComptonSpectrum.htm (2020/04).

2

High Speed electronics used in image sensors I: read-out

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

Marie Curie

"A man may imagine things that are false, but he can only understand things that are true, for if the things be false, the apprehension of them is not understanding."

Isaac Newton

Systems, in particular image sensors for this matter, could be a great design and they might be able to perform many interesting tasks; however, the readout module will ultimately define how much we can get out of them. This chapter discusses the importance of read-out modules and gives insights about aspects that should be accounted for a readout design. Further on, the reader can find concrete new ideas and solutions for the most common problems found in the fascinating world of Image Sensors. Part of the content of this chapter has been published in [1].

In this chapter, different strategies for read-out systems and protocols to tackle particular problems encountered in image sensors will be discussed. Generally, image sensors have massive amount of information to be transmitted to an external system that usually performs the processing according to the needs of the aimed application. This information is usually composed by detected intensity, timing information, addresses of events, reference values and various variable or fixed registers. Here, the specific problems related to read-out systems for image sensors will be presented and explained from a holistic point of view.

2.1. Read-out systems

2.1.1. ASIC vs FPGA controllers

There exist two different approaches to design read-out and processing systems for image sensors. These pure-digital modules can either be included on chip along the rest of the image sensor circuit, or alternatively, can be placed in an external FPGA. There are several pros and cons when comparing FPGA vs ASIC systems; the most notorious being power, flexibility, cost, barrier for entry and performance [2]. Each of those categories has to be considered and evaluated according to the application that the system aims at. Here is a summary for image sensors:

- Power: strictly depends on the type of application and might or might not be a problem.
- Cost: usually should not represent a problem since the readout system is a small part of the whole sensor. However, in same cases, due to the complexity of the read-out, it could be a point to be considered.
- Barrier for entry: not applicable in this case since the sensor is already designed in an ASIC.
- Performance: on-chip communications are much faster than inter-chip communication, such is the case for chip-FPGA systems.

As general rule, FPGAs are the preferable option when the system is still not mature and the design is at its early stages and ASIC solutions are used once the system is mature enough and it is ready for mass production. Flexibility and type of application come into play.

Information delivered by image sensors is large and can be the bottle-neck of the system [3]. Researchers are always trying to find new techniques to tackle this problem that becomes bigger and bigger as image sensors have more resolution and capabilities [4]. Prompt processing with the consequent volume reduction helps to relax the communication systems. In case a given data needs to be processed, it is crucial that the processing unit be as close as possible to the source of that information; this not only helps lowering the requirements for the communication systems but it also reduces power consumption and latency. In this work, both approaches, using FPGAs and read-out on chip, were used and their results compared.

2.1.2. FPGA-based read-out system implemented for [1]

A complete read-out system was designed in FPGA to set all the parameters, to write the masking configuration and to read out the information of the events. In this subsection, the most important blocks and modes of operation of the read-out system are explained. The image sensor 9x18 MD-SiPM [1], was part of the project EndoTOFPET-US [5], hereafter called Endotof chip.

Endotof chip was designed by Dr. Shingo Mandai, with whom I worked in close collaboration to design the read-out and to perform the measurements.

Introduction to 9x18 MD-SiPM Endotof

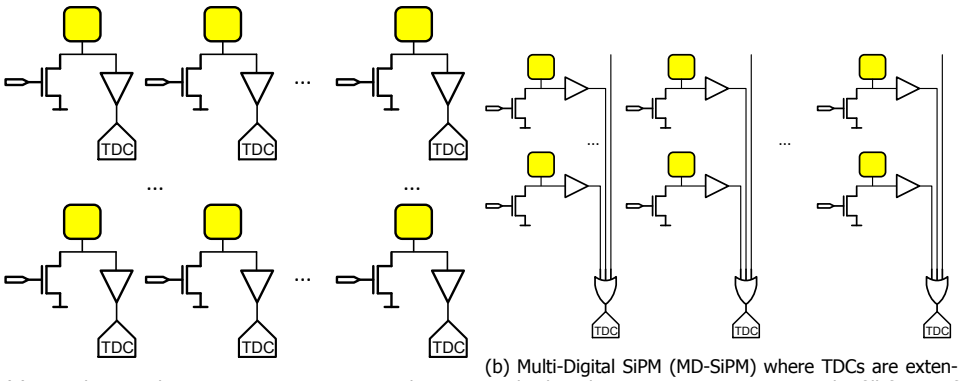
The sensor was fabricated in a standard CMOS process and is fully MRI compatible i.e. it can operate in large static B fields; it is designed to couple with an array of 9x18 0.71 x 0.71 x 15 mm^3 LYSO scintillators. The chip comprises a matrix of digital SiPMs, each composed of 416 pixels. The first 432 pixel responses during a scintillation event are captured and digitized by a bank of column-parallel time-to-digital converters capable of a time bin of 49.5 ps. The chip can produce up to 67.5 million time-stamps per second to adequately capture a gamma-ray event rate of 625 kcps, which is consistent with typical values used in prostate and pancreatic cancer PET diagnostics; it communicates to a central acquisition unit through a 320 Mbps LVDS protocol and it dissipates less than 300 mW.

Multi-channel digital SiPMs (MD-SiPMs) [6] are an emerging family of sensors capable of capturing a large number of individual time-stamps of incoming photons digitally. While the optimal solution would be to associate a time-stamp capability to each pixel (Fig. 2.1a), PDE considerations forced us to use extensive sharing of resources so as to achieve a higher fill factor (Fig. 2.1b). In our case, we used column-shared TDCs, whereas SPADs shared a TDC every three rows similarly to [7]. Higher granularity, both in spatial and temporal domain, leads to fundamental time resolution limits, enabling time-of-flight PET to achieve significant improvements in single-photon time resolution and noise robustness [8].

Architecture

The sensor chip comprises 18x9 MD-SiPMs, each composed of 416 pixels in an array of 26x16, a bank of 16x3x9 TDCs, a smart reset mechanism, noisy pixel masking memory, a digitally controlled 25V voltage generator, and a fast readout bus operating at $320Mbps$. The floorplan of the sensor chip is shown in Fig. 2.3 along with the micrograph of cluster and SPADs. The chip can be operated in two different modes that have very specific proposes. The first mode is mostly meant for calibration, this allows the system to have the complete access to all pixels individually. The second operation mode is used when the chip is working in a PET system. A logic circuitry was designed to sum the SPADs that have been fired in every MD-SiPM. Energy and time-stamps for every MD-SiPM are computed and registered into an internal memory. Every frame of $6.4\mu s$, the sensor chip transfers the energy and time-stamp of each detected gamma event to a data acquisition unit (DAQ) via an FPGA interface through LVDS.

A pixel, the fundamental component of the MD-SiPM, consists of a SPAD, a fast quenching mechanism, a 1-bit memory, masking circuitry and a gate to connect the



(a) Digital SiPM where every SPAD is connected to one TDC.

(b) Multi-Digital SiPM (MD-SiPM) where TDCs are extensively shared among SPADs to preserve the fill factor of the active area.

Figure 2.1: Types of SiPMs.

output of the SPAD to the TDC associated with it. SPADs usually have counts even in complete absence of light; this undesirable effect, called Dark Count Rate (DCR), is handled in two different ways in this sensor chip. A masking circuitry designed within every pixel contains a 1-bit memory that can be set during the configuration phase. In the case of a high DCR SPAD, the memory is set to 0, then a driver forces the digital output to 0 and deprives the SPAD from firing capability. This mechanism is of paramount significance due to the fact that the TDCs are extensively shared by the SPADs and only one noisy SPAD can nullify the associated TDC. Though masking inhibits noisy SPADs from triggering the TDCs when no event is present, it goes to the detriment of PDE and energy resolution. The PDE is reduced in the same proportion as the masking increases and since there are fewer available SPADs to estimate the energy, so energy resolution is worsened. The masking module, discussed in the last paragraph, is designed only to deal with extremely noisy SPADs. Special attention should be paid to the fact that regular DCR (i.e. DCR at or near the median value) can also negatively affect the chip performance by firing the TDCs when no photons have been detected. To mitigate this effect, the chip has a dedicated circuit called 'smart reset'. This mechanism has essentially two main parameters: a threshold (TH_{TDC}) and an interval time. The chip ascertains if a gamma event has occurred in the last interval time (programmable from $50ns$ to $6400ns$) by checking if the number of fired TDCs exceeded the threshold that was defined beforehand. If this threshold was not exceeded, i.e. no event is present, all the TDCs are reset as well as the pixels that are connected to them, otherwise the process continues. Thanks to an on-chip evaluation and decision process, smart reset enables the re-utilization of TDCs and pixels that may have fired due to dark counts, thus allowing a much larger dynamic range, better energy resolution and better time-stamp statistics. As an illustration, we report a simulation of the detection process in Fig. 2.2. The figure illustrates the number of fired TDCs when the chip is in complete darkness; both cases with and without use of smart reset are shown. Five simulations were performed as examples. When smart reset

is not activated, the number of fired TDCs increase exponentially over time and the occupancy rapidly gets closer to the total number of TDCs for that particular MD-SiPM column, consequently the number of TDCs would be very low if an event occurred in the last part of the frame. The quantity of free (not fired) TDCs is, in general, strongly dependent upon the instant when the event might occur. On the contrary, by means of the smart reset, the number of available TDCs is always kept below a certain value defined by the total number of TDCs minus the threshold. As a result, there is always enough headroom for a gamma event, no matters when that event occurs.

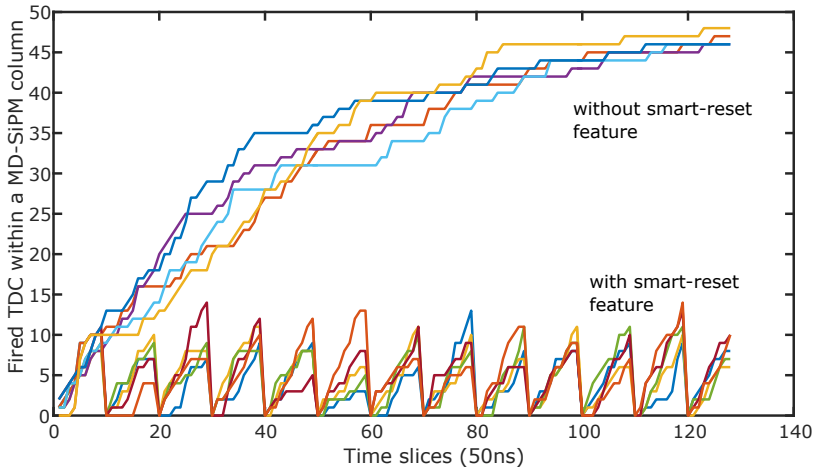


Figure 2.2: Five simulations, with and without SR, of number of fired TDCs on a MD-SiPM column when DCR = 3000 cps, smart-reset time = 500 ns. All the SPADs are considered to have the same DCR and no screamers or masked SPADs are present.

Operation

Right after a photon impinges the SPAD, the generated pulse rapidly propagates and fires the TDC so as to register the time-stamp. The 1-bit memory on pixel latches the value of the SPAD at the end of the frame; thus, permitting the SPAD to operate during the next frame while the read-out system sweeps all the SPADs to send all the information to the DAQ. A decoder addresses the pixels row by row and connects them through a common bus to a register bank that holds the information until is read out from the chip. The pixel contents are copied directly to the output registers in case the chip operates in calibration mode. Then, the register is clocked by the FPGA that is closely attached to the chip and the data is read out; immediately, a new pixel row is copied to the registers and again is read out. The process continues until everything has been transferred. In PET mode, i.e. an event-driven mode where gamma events are detected and Compton/noise events are filtered at the sensor level, the pixel contents are copied and summed for every MD-SiPM and only the result is transferred to the output registers. An

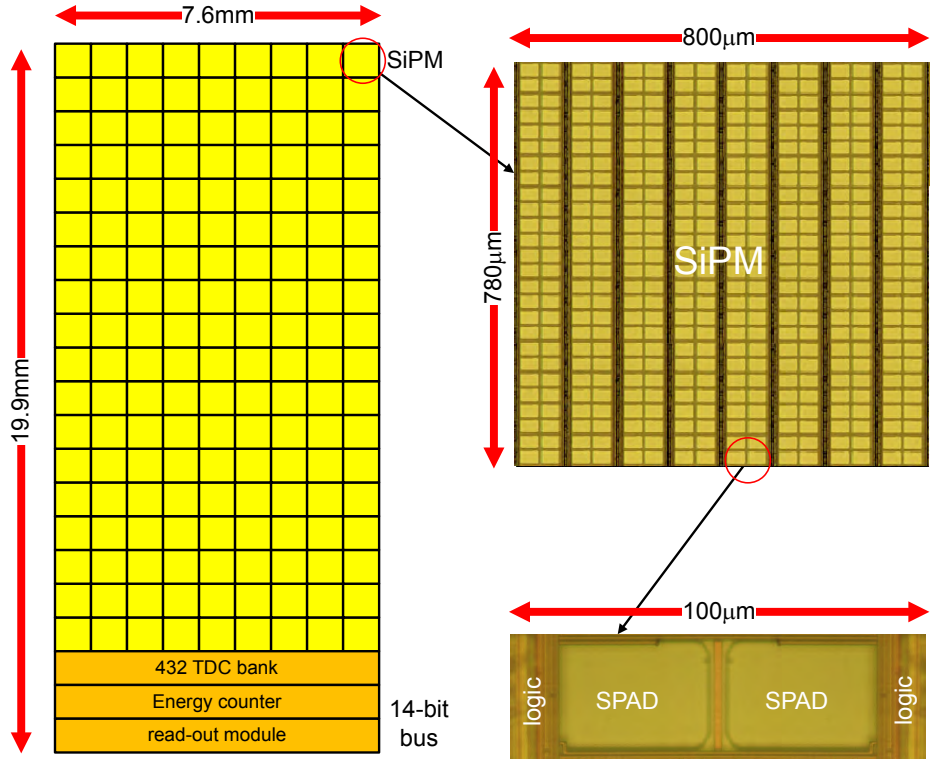
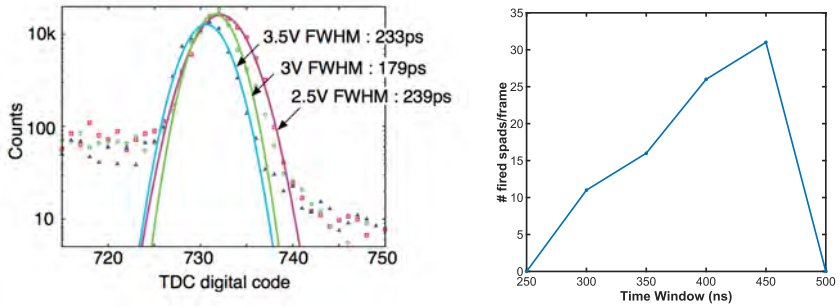


Figure 2.3: On the left, diagram of the sensor, activate area, TDC bank and digital registers are displayed. On the right, micrograph of a cluster and SPADs.



(a) Single Photon Time Resolution (SPTR) of the sensor for different excess bias voltages (2.5 V, 3 V and 3.5 V).

(b) DCR measurement using smart-reset feature set to 500 ns.

Figure 2.4: Measurements results of Endotof chip. Time resolution and smart-reset feature.

integrated module on chip evaluates the overall count and decides which MD-SiPMs that detected a gamma event photon shower. Only the TDCs that are linked to those MD-SiPMs are transferred, the remaining TDCs are right away dismissed, to avoid overloading the communication system. In this way, up to 4 gamma events can be read out in 6.4 μs. Meanwhile, the FPGA provides all the signals that the chip needs to measure the new frame. All these operations are run simultaneously, thus achieving a very high effective time window ratio of 98%, during which the chip is available for detection. The TDCs have a time bin or LSB of 49.5 ps, while they exhibit a worst-case differential nonlinearity (DNL) of 1.79 LSB and integral nonlinearity (INL) of 7.16 LSB. The FPGA estimates time mark and energy of the gamma event from all valid time-stamps collected during a frame using a fast algorithm that ensures an accuracy limited only by the Cramer-Rao bound [8], [9].

Results

A full characterization has been performed on the chip; the most relevant measurements for PET systems are Single Photon Time Resolution (SPTR), shown in Fig. 2.4a and energy resolution shown in Fig. 2.5. The system has an energy resolution of 15.7% when the dector is coupled to a LYSO scintillator. The smart-reset feature explained in the previous section was also measured and analyzed. Fig. 2.4b. The number of fired TDC increases exponentially and gets reset every 5 slices of time in case of no events.

Read-out

In this section, the read-out system will be described and explained.

Block Diagram: the read-out block diagram is presented in Fig. 2.6. The chip and FPGA form the endoscope that is one of the detectors of the system. The model of the small FPGA is Lattice iCE40 LP8K. The firmware is prepared to execute configuration, calibration and operation in both the modes. It delivers all the signals required by the chip and reads out pixel and TDC data. All the details of the firmware

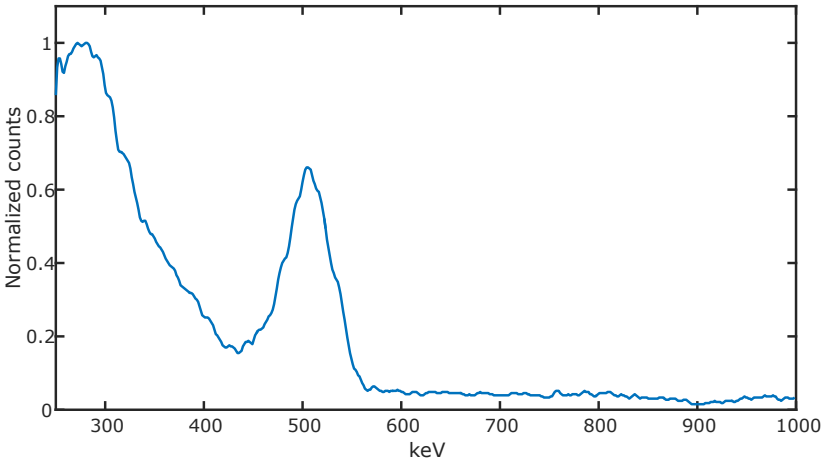


Figure 2.5: Energy resolution of the system is 15.7% when a LYSO scintillator is coupled to the sensor.

are given in the next subsection. The acquisition board collects events data from the endoscope and from the external plate in order to process the coincidences and filter the singles to only transmit the first ones to the PC where it can be used to generate a 3D representation by means of reconstruction algorithms.

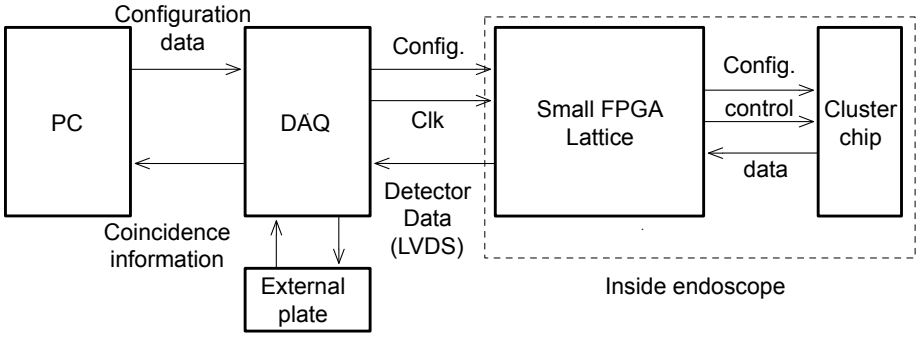


Figure 2.6: Scheme of the whole read-out system. In this work, the focus was on the side of the FPGA.

Firmware architecture: Fig. 2.7 shows a general description of the firmware written for this project. The 'command unit' takes the commands from the DAQ in order to perform configuration and operation of the chip. The 'Interface' module sets all the parameters of the sensor and carries out the read-out. The parameters include: window time, smart-reset threshold, smart-reset period, masking data, etc.. The data captured by the sensor, either gamma events or laser events or just noise if the system is working in the calibration stage, is sent to TDC decoder to

compress the volume of data and then to a FIFO. There is a module called 'Aligner & serializer' that aligns data coming from different frames and serializes it into a 1-bit FIFO that is transmitted via LVDS to the DAQ. As part of this project, the LVDS deserializer was also included in this design.

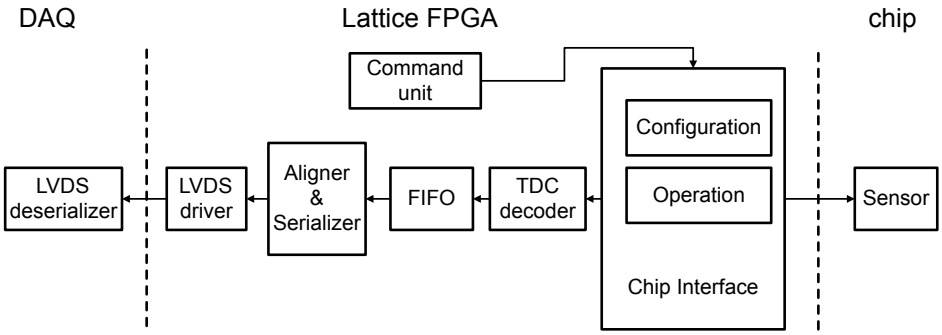


Figure 2.7: Block diagram of the firmware inside Lattice FPGA plus the interconnection points to the rest of the system chain.

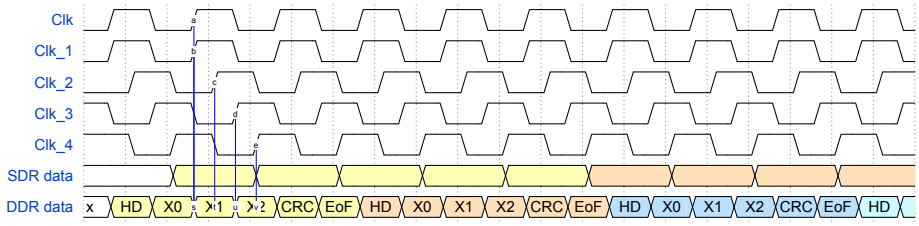


Figure 2.8: Waveforms for SDR and DDR data packets when phase is unknown.

LVDS deserializer: the events information is sent through LVDS using DDR 320 Mbps communication. Although the FPGA takes the clock from the DAQ and the whole system is synchronous, the phase of the differential pair when it arrives to the DAQ depends on many factors as output delays of the FPGA, input delays of the DAQ and even the length of the cable which cannot be accounted in the design. This is the reason, a circuit for clock or phase recovering has to be included in the receiver. The limited space in the prototype made the use of a dedicated lane for the clock impossible to implement. Fig. 2.8 shows the typical case when Single Data Rate (SDR) and Dual Data Rate (DDR) are synchronous with the clock but the their phase is unknown. Were the frames latched by the main clock of the system, the data would be incorrect since the latching process happens at the time of the signal transition. Since the actual phase of the data is unknown, the system uses 4 phases of the main clock that are easily generated by the PLL of the FPGA. Fig. 2.9 shows the LVDS deserializer. Sampled strings are stored in 4 registers to be analized. The structure of data frames is as follows: they start with a head word that includes

a constant value, length of the packet and padding bits; after the header, there are N words with the payload and at the end comes a CRC code and End-of-Frame (EoF) word. This structure is used for the phase detection and data alignment. The content of the registers is compared to the structure of the data frames in order to chose the phase that latches the data correctly. The firmware keeps calculating the statistics of the communication and shows the percentage of hits vs misses of the system. The firmware can analyze the phases again at any moment if it is required due to any change in the system and regain synchronization.

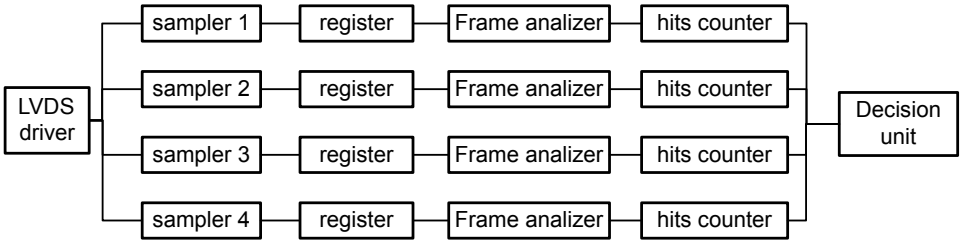


Figure 2.9: Four sampling channels clocked by 4 phases of the PLL are used to find the proper phase of the string.

Configuration: the core of the read-out system is the 'Interface' module because this is the module that operates directly on and with the chip. The operation starts with configuration of the chip. It takes the bitmap from the DAQ and sends all the parameters to the chip. The sensor has a 48-bit register that stores global parameters. The masking information of the chip is stored in in-pixel memory that has to be written with external signals and clocks proceeding from the 'Interface' module.

Operation for PET: the sensor is prepared to work in PET mode, which sets the read-out to send the content of the TDCs and the energy calculated on chip. For PET application, it is not needed to know which pixels have fired, but instead, how many of them did, which is proportional to the energy of the gamma photon. The time of the events is also very important to estimate the LoR. By dismissing the pixel information, and only reading time and the total of fired pixels, the system saves time and power. Internally, the chip is simultaneously performing 3 tasks during a frame:

- Events capture: the chip is sensitive to light and every hit is recorded in the in-pixel memory and its time of arrival is registered by the TDCs.
- Energy counting: an internal module, driven by the FPGA is collecting the pixel information and adding it up into a register.
- Transmission: data packets are sent to the FPGA.

One of the goals of the sensor is to remain operative for the longest time possible. This means that detection, processing and read-out have to be performed

simultaneously. The firmware has 3 modules that assist the chip to perform these 3 tasks at the same time; however, those tasks depend on each other so the only solution is a pipeline-alike processing sequence as shown in Fig. 2.10.

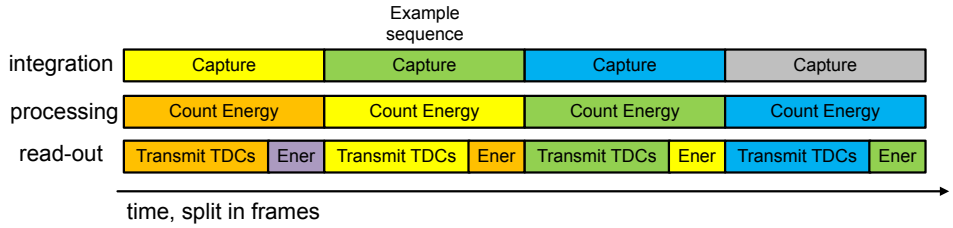


Figure 2.10: Sequence of data frames. Same color corresponds to the same information. Same horizontal position corresponds to the same time slot in which those tasks are performed.

At a given time, in the example sequence, the chip is transmitting the TDC information of the green sequence that was captured on the previous frame and sending the energy of the yellow sequence that was calculated on the previous frame and captured two frames back. The energy of the green sequence is sent with the TDC information of the blue sequence 1 frame later. Realignment of the events and TDC decoding are done on the FPGA. This way of operations ensures that the chip is sensitive and functional for 98.2% of the time.

2.1.3. Specific read-out systems for image sensors

A typical block diagram for a read-out system can be found in Fig. 2.11. The system has several units whose information can be accessed externally through a read-out module that is connected to all those internal components. The first subsystem is the interface between all the components to be read and the core of the read-out block. The core is the central unit that takes commands to read or write and acts as an arbiter to control the flow of the information. The core writes the data into a synchronous output FIFO that is read by a serializer unit. The serializer does not have to work at the same speed as the core or any of the components. By separating write and read clocks by means of a FIFO, it is possible to change the read speed according to the needs of the system. At the end of the chain, a physical driver conveniently transforms the data bits into signals suitable for the system, such as single ended, LVDS, sub-LVDS, etc.

Bus protocols

There exist several types of bus protocols that can be used as interface in digital systems. The application, speed, complexity and robustness will determine which of these is the most appropriate for the system. The most popular bus protocols are listed below along with their typical applications.

- Advanced Microcontroller Bus Architecture (AMBA): is an open bus protocol meant for SoC communication covering large number of applications. It supports single and multibyte transmission and multiplicity of masters and slaves. It is oriented to multi-microprocessor systems.

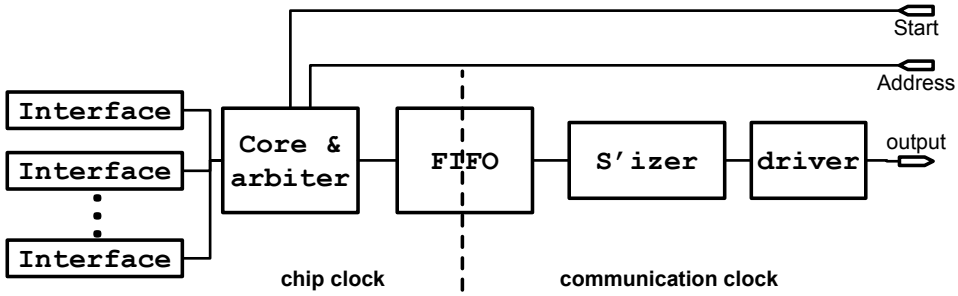


Figure 2.11: Block diagram of a general read-out system.

- WishBone: is an open bus protocol meant for SoC communication. It supports single and multi-byte transmission and multiplicity of masters and slaves and it is lighter than AMBA.
- Advanced eXtensible Interface (AXI): is a revision of AMBA to add many features desirable in micro processors systems, such as support on unaligned data, out-of-order transaction completion, atomic operations, etc.

2.1.4. Image Sensor Communication Protocol (ISCP)

Although these open-core bus protocols are well extended and documented, they are pure digital and not meant to include mixed-signal blocks with very little or non-existent digital interface. In addition, they use only MUX signaling for address selection, which might not be practical when systems are large and highly distributed. ISCP is introduced in this work; it is a protocol is intended to be a communication solution optimized for systems that have the structure of image sensors. The main characteristics of ISCP are listed below:

- Single-master multi-slave.
- Single and variable burst transactions.
- Optional handshake.
- Tri-State data bus for large distributed systems.

The way of operation is as follows: the master is the only device that can initiate a data transaction by asserting the "Start" signal. The length and type of transaction is defined by the master and the slave can accept it or reject it based on the its own state. When several slaves are connected to the master, these can be polled to see which one is ready to send data and the master can start the transaction with anyone. The Fig. 2.12 depicts the algorithm interaction between master and slaves. As explained before, the handshake is not mandatory; in this case, it is a forcing transaction that the master can start at any moment. This mode is designed for modules that have very little digital interface without any kind of handshake (e.g. TDCs), or for very simple modules that are constantly updating values that will be always valid either they are read before or after update (e.g. temperature sensor and auxiliary SPADs that can be used as activity estimators).

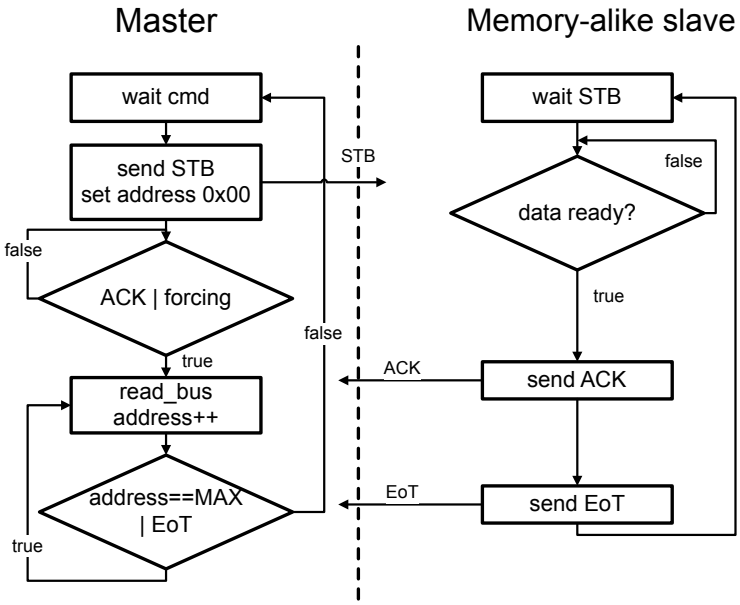


Figure 2.12: Address scanner: depth, width and range to be swept are configurable from the core.

The type of information that needs to be delivered in images sensors usually fall into two categories. The first category being sequences of words of fixed or variable length: this is the case for time-stamps and events. In general, the components in this category have a FIFO or registers to store sequential data. The sequence could be fixed or variable in length. The second category is composed by memory-alike components: this is the case for a fixed number of TDCs, fixed number of pixels, results from on-chip processing, MUXed registers, etc.. The lengths of the components in this category are known and the information is arranged in a vector or matrix that have well-defined addresses. These two different categories of modules lead to the implementation scheme that is explained in the next subsection.

2.1.5. Implementation

Address scanners: an address scanner is a module that can select particular addresses in a component and transfer the information on demand to the next stage. This module can be used to read TDCs, pixel information, registers, etc. A diagram can be found in Fig. 2.13, where the input ”Start” initiates the read-out process. The parameters ”Width” and ”Depth” set the width and the depth of the memory-alike component to be read. In every clock cycle, the address scanner reads the current information and sets the next address to let the component resolve it in one clock cycle. The address is swept in all its range. An end-of-read (EoR) flag is asserted at the end so that the core knows it can continue. The address scanner can be easily modified to accept components of the second category. An extra input ”Valid” is added as a second termination condition for the read process.

The address scanner starts the read process and keeps on reading until either the address has reached the maximum possible or the Valid flag has been pulled down. The parameter ”Depth” can be modified at any time to enable different read-out schemes. The implementation has been done in VHDL, synthesized using Innovus and its complete code can be found in the appendix.

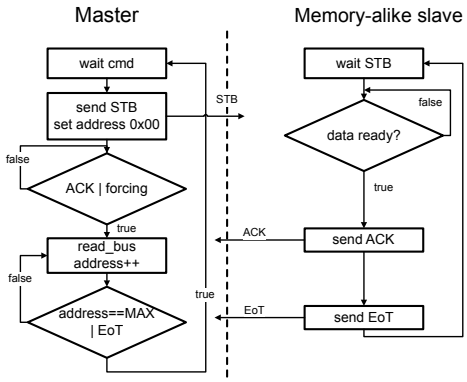


Figure 2.13: Address scanner: depth, width and range to be swept are configurable from the core.

Wrappers: each component might have a different protocol and different type of signals, lengths or information. For image sensors, the most common type of components were introduced along the solution to read them out: the address scanner. However, there might be components that do not fit in either of those two categories. In order to include these components, it would be necessary to change the interface with the core. This is highly dis-encouraged because it would lead to non-regular structures; thus, more verification would be needed, and this is risky and time-costly. The purposed solution is an intermediate module that is inserted in between to make said component be seen as components of the first or second category. As an example, in Fig. 2.14, a 256-bit register has to be included in the read-out system. Fixed-length registers do not fit in either of the two categories, so it needs a wrapper to make it fit to the system. As the main bus of the core is 16 bits in this case, a MUX with Width=16 and Depth=16 (Address bits = 4) can be used as a wrapper. Although it is debatable that wrappers also need to be checked and verified, they are very simple in general and do not interfere with the core or the address scanner that can be standard and general for any type of image sensors. The verification does not have to include the complete interface, but rather can be performed in a simplified scenario.

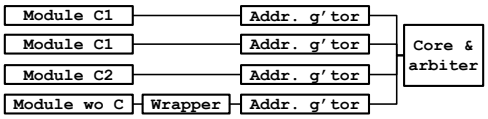


Figure 2.14: Interface between modules and core for different categories.

2.2. True First-In-First-Out (TFIFO) memories for high data throughput

2.2.1. New type of memory for ToF systems

Time-of-Flight (ToF) applications using image sensors discussed in this work, such as LiDAR, 3D cameras and PET systems, extensively use SRAM memories [10] for FIFO functions to store photon arrivals. Since SRAMs are fast and have good data-area density, they have been the preferred option. Nonetheless, SRAMs, due to their Random-Access nature, are not inherently a FIFO. They have address decoders that are meaningless for FIFO operation and additional state machines are required to make SRAMs behave as FIFOs, thus causing undesired overheads and leading to sub-optimal solutions in terms of speed, area and power consumption. In this section, an asynchronous True FIFO (TFIFO) is purposed. The TFIFO is an asynchronous FIFO memory by design; this means that all its components are meant to operate in asynchronous FIFO mode and were optimized for it. Therefore, the TFIFO can achieve better results than its closest competitor: SRAM memories. The design was implemented in 40nm TSMC technology, tested and the result are shown and discussed. Although the system was designed for ToF sensors to time-stamp random photon arrivals, it can be also used as a data bridge in a multiple-clock domain system. Two channels back-to-back could be used to implement a full-duplex TFIFO. In ToF systems, photons and events are originated by random Poisson distribution phenomena [11], random LASER reflections or random DCR. In every case, these events are asynchronous with the read clock of the system. This is the reason whereby the asynchronicity of the memory element is mandatory. For the same reason, the write signal cannot be limited to a periodic clock-like signal.

2.2.2. Discussion: TFIFO vs SRAMs vs Registers

There are different ways to implement the memory element used in ToF systems. The most popular ones are registers and SRAM memories. Registers are used when the amount of data and the size of the system is small. The advantages of registers are very clear: high speed and easy implementation. Although registers can be manually designed, they can, it can also be synthesized by digital tools. Even when they have only one clock domain, it is possible to operate them in ping-pong fashion to enable two different clocks for writes and reads. The dead time for performing the ping-pong switching time is negligible. Area and power do not represent a problem as long as the system remains small. As systems scale up and grow in size, the area and power used by registers quickly become unmanageable. This is the type of scenarios where SRAMs excel. Fig. 2.15 shows how area changes for both types of

memory presented as a function of their capacity. The calculations and estimations were based on TSMC 40nm technology and also under the premises there is no lower bound for the depth of the SRAM memory which might not be the case in off-the-shelf IP cores. The plot shows that while for small systems, SRAMs are not practical due to their size, they become a serious solution when more capacity is needed. However, as explained in the introduction of this topic, SRAM memories, used for FIFO operations, carry with several disadvantages that are listed here.

- SRAM memories have address decoders that do not have a meaning for FIFO operations.
- Decoders have to be replicated in case of double port SRAM memories.
- An external FSM circuit is needed in order to make SRAM memories operate as FIFO memories.
- Often times, SRAMs require PLL for clock generation and periodic signals.

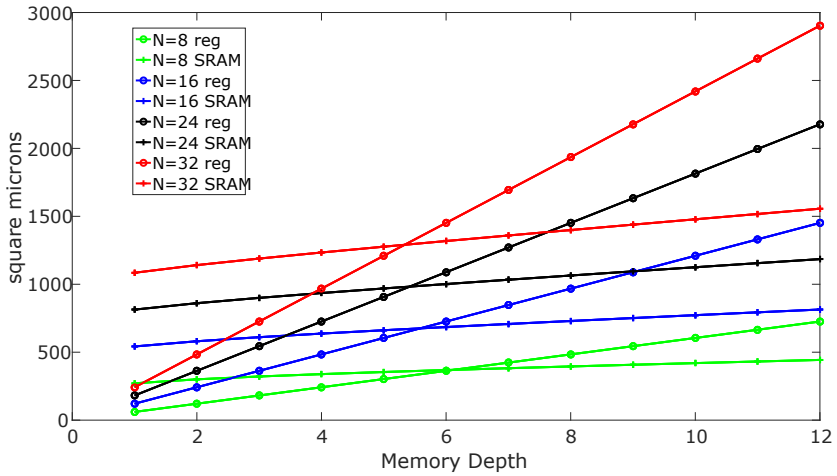


Figure 2.15: Comparison between SRAM memories and register-based memories.

The purposed TFIFO makes use of static bit cells to store information, and all the electronics is designed to cope with the disadvantages of SRAM memories when FIFO operation is needed. The memory classification for TFIFO is: volatile, static, FIFO, asynchronous.

2.2.3. Architecture

Block diagram of TFIFO: the TFIFO is organized in 6 blocks depicted in the Fig. 2.16. The memory block has 4-kbit storage capacity arranged in 128 32-bit words. At the top, 32 buffers are placed to drive the differential data to the bit array. At the bottom, 32 sense amplifiers read and latch the differential outputs from the bit cells. On the left, 128 word drivers handle the read and write requests. Almost-full logic is included at the top left next to the inputs to easy the communication

with the writing module. Similarly, the almost-empty logic is placed at the bottom left next to the outputs so as to easy the communication with the reading module. The complete circuit of the TFIFO is shown in Fig. 2.20 and Fig. 2.34 shows a micrograph of the chip along with its layout.

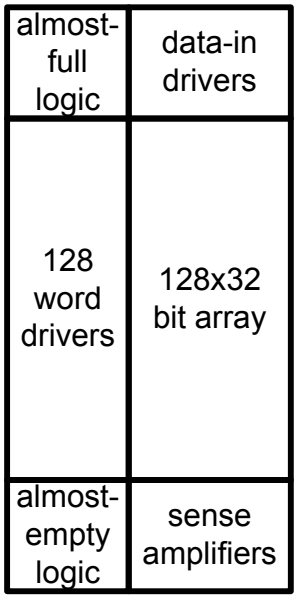


Figure 2.16: Block diagram of the TFIFO.

The bit cell: the schematic of the bit cell used in the design is shown in Fig. 2.17. Every bit cell has two back-to-back inverters to store 1-bit datum and 4 transistors to perform independent read and write operations, controlled by the signals Write Enable (WE) and Read Enable (RE). The layout, exhibited in Fig. 2.18, was carefully designed to share as many signals as possible among neighboring cells. The achieved area is $0.82\mu m^2$ when abutted. Static Noise Margin (SNM) post-layout simulations of the cell that account for process variation are shown are Fig. 2.19.

Word drivers (WD): words drivers determine the action to be performed (read or write). The drivers are interconnected to the previous two drivers and the next two drivers, thus forming a circular FIFO. The physical distribution of the positions, depicted in the inset of Fig. 2.21a, guarantees equal timing between any two consecutive drivers by folding the structure and interleaving the word drivers. The circuit of the WDs is shown in Fig. 2.22. The core of the word driver is a T flip-flop that stores its state. There are two logic branches that attend the read and write requests and send the pulses to the bit array according to the aforementioned

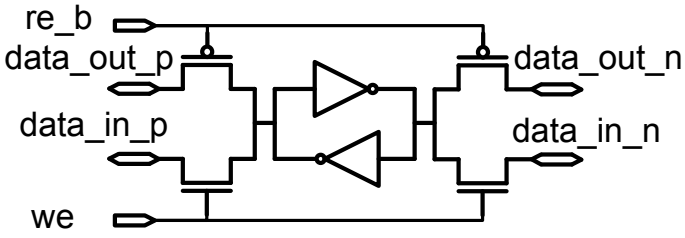


Figure 2.17: Schematic of the bit cell.

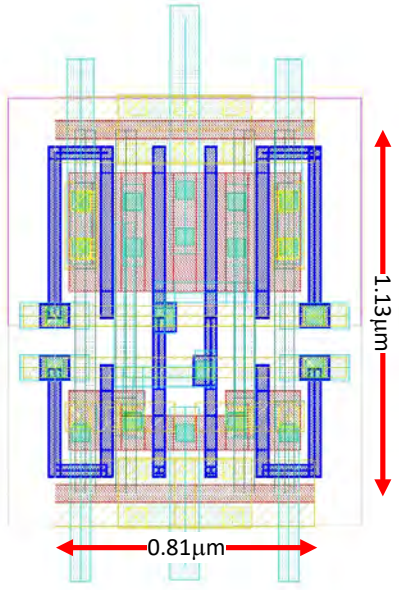


Figure 2.18: Layout of the bit cell.

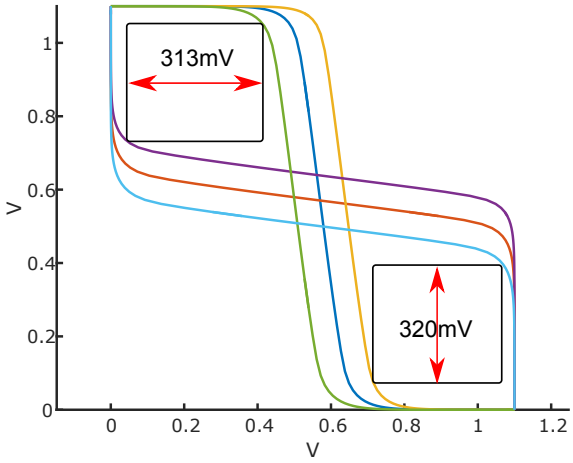


Figure 2.19: Static Noise Margin (SNM) 313 mV. It represents the maximum tolerable noise that can occur inside the bit cell without modifying its content.

conditions. After any of these requests, a pulse is sent to the flip flop and its state switches.

Fig. 2.21b shows a section of the TFIFO; the WD 32 is connected to the positions 31 and 33 to attend the two basic FIFO operations: pop and push, and it is also connected to the positions 30 and 34 to provide information about the status of the TFIFO: almost full and almost empty. Connections for the WDs 32 and 95 are shown in the same figure in order to understand the interconnection pattern that is repeated along the whole memory for all the word drivers.

Details on the layout: Fig. 2.23 shows the actual layout of the same middle section reported in the previous paragraph. Now it can be observed how the layout of a single bit cell can be clustered in an array shape as it was anticipated. In a digital design, every row is flipped with respects to the next and previous row in such a way that the NWELL is at the top in one row and it is at the bottom in the next row. This means the bit cell has to be designed to abut perfectly with itself when the second instance that is flipped up-side down. The same criterion is needed along the X axis because the bit cell has to be able to be abutted with a second instance that is flipped horizontally. In this way, a macrocell of 2x2 bit cells can be created and repeated throughout the whole design. Further improvement could be done if especial DCR rules are used in SRAM-aware technologies. In this case, minimal width and clearance are smaller than standard values. For instance, in TSMC 40nm technology, very small SRAM bit cells can be made by using these rules; the smallest that has ever been reported is 0.242 μm^2.

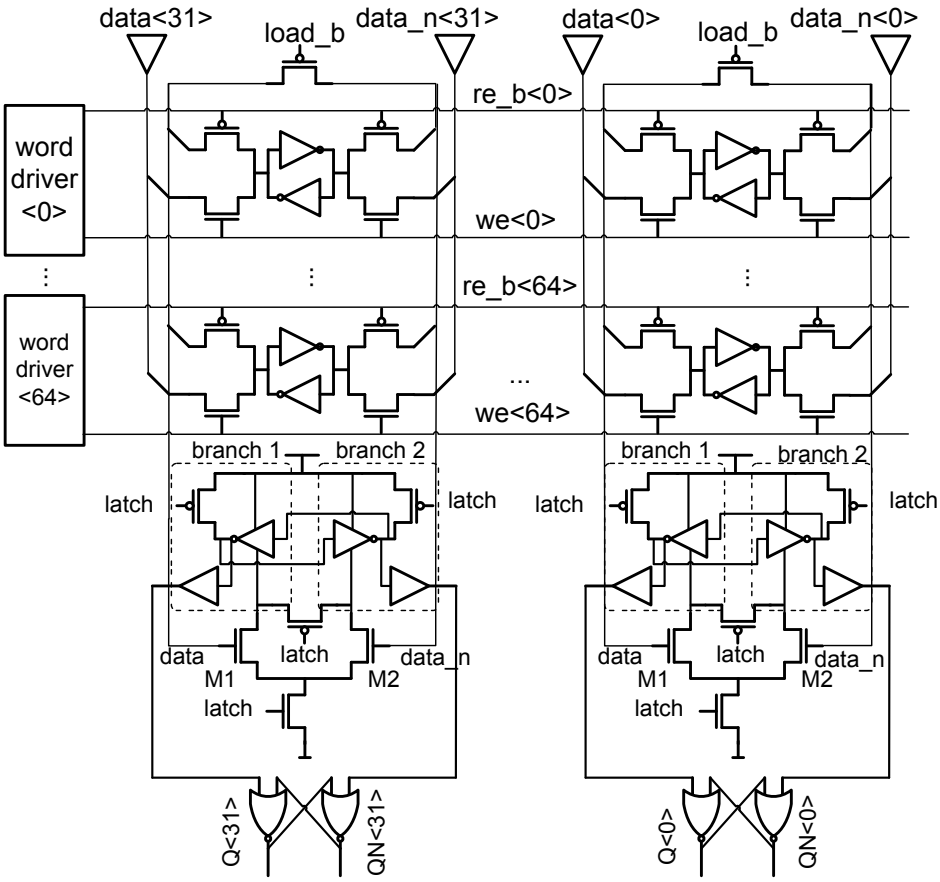
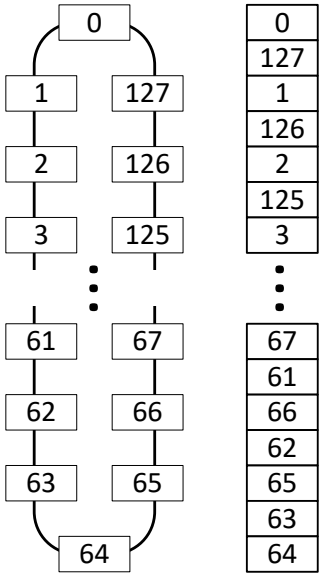
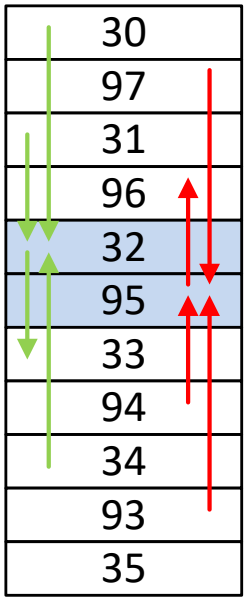


Figure 2.20: Complete circuit of the TFIFO: memory array with 128x32 cells, buffers at the top, sense amplifiers at the bottom and word drivers on the left.



(a) Logical distribution of the word drivers vs physical distribution.



(b) Zoom of a middle section with all the connections of the highlighted word positions 32 and 95.

Figure 2.21: Middle section of TFIFO.

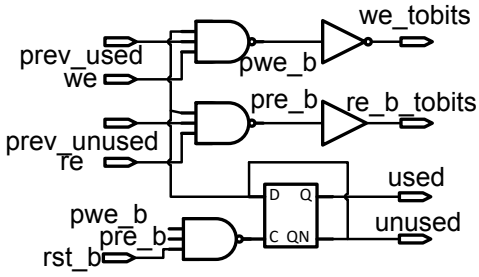


Figure 2.22: Circuit of word drivers: 1 flip flop keeps the state of the memory word and the electronics attends the read or read requests.

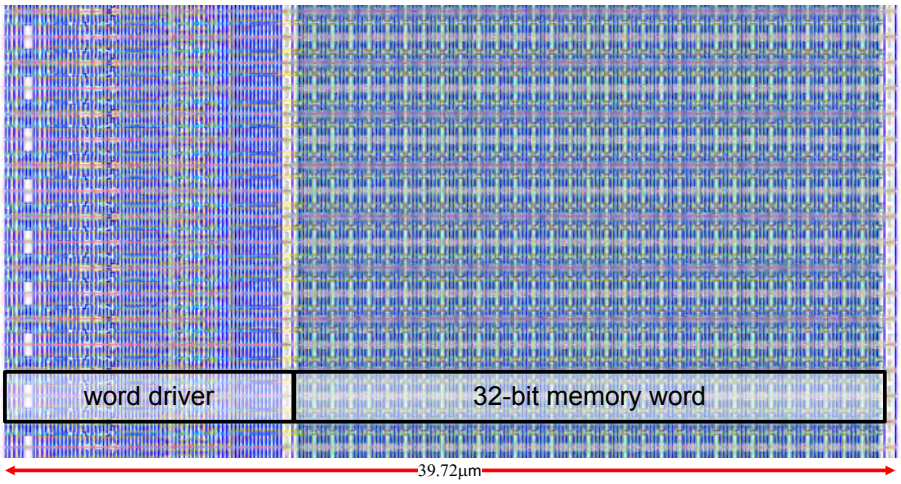


Figure 2.23: Layout of the same middle section of TFIFO shown in 2.21.

Sense amplifiers: sense amplifiers are comparators that evaluate a small differential input at the moment the "latch" signal is asserted and provide a full-swing differential signal at their outputs [12]. After a the decision is taken, the outputs remain in their value until the sense amplifier gets reset by the "reset" signal. The schematic is shown in Fig. 2.24. The basic structure comprises a low-threshold NMOS transistor pair that pulls down two inverters when the "latch" signal is asserted. The branch that gets a higher voltage at its input pulls stronger and makes the node go to 0. The branches are connected to a NOR-based active-low SR latch that changes its output once the winner branch has been pulled down. The design was optimized for voltages that are in the range from Vdd/2 to Vdd. The layout of the sense amplifiers is shown in Fig. 2.25. The traces are equalized in terms of skew and capacitance. In order to alleviate the effects of process variations, the sense amplifiers have been laid out considering the effects of mismatch and asymmetry.

Pointer logic: The FIFO pointer logic was implemented in a distributed way by means of the WDs that perform read and write operations and set the almost-full and almost-empty flags. The WDs have two possible states: *used* (S=1) and *unused* (S=0). Every WD independently decides when to react to a read or a write request based on its own state and the state of the previous WD: $D_i = (Si-1, Si)$. A write request must be attended at the i_th position if $Di = (1;0)$ while a read request must be attended if $Di = (0;1)$. WDs assert WE or RE according to the request and change their own state $S_i^{next} = not(S_i)$ (after a write, $S_i = used$, after a read, $S_i = unused$). Almost-empty and almost-full flags are calculated independently by the WDs based on their own state, the state of the WD two positions back and the state of the WD two positions forward $T_i = (S_i - 2, Si, S_i + 2)$. The TFIFO is almost full when exists at least one trio of word drivers T_i such as $T_i = (1;0;1)$. In case

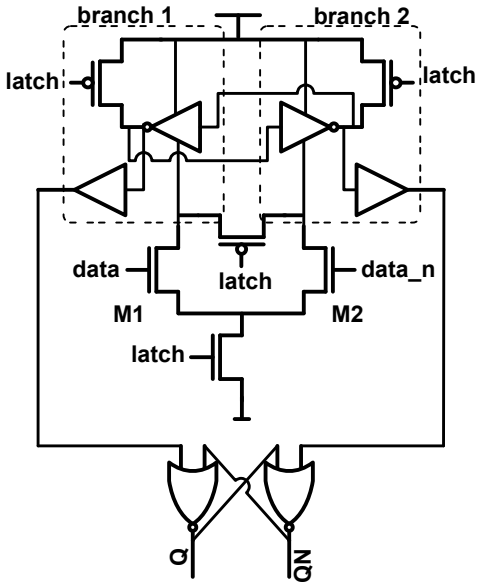


Figure 2.24: Schematic of the sense amplifiers showing the two main branches that collapse into a stable state that depends on the differential input when the signal "latch" is asserted.

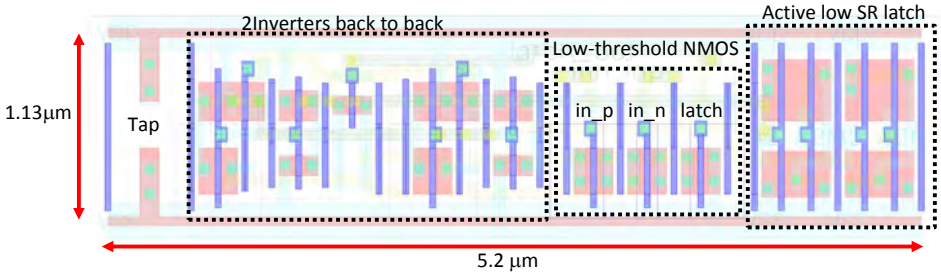


Figure 2.25: Sense amplifier layout. By means of extraction tools and simulations, the capacitances of every differential lines and differential devices were laboriously equalized.

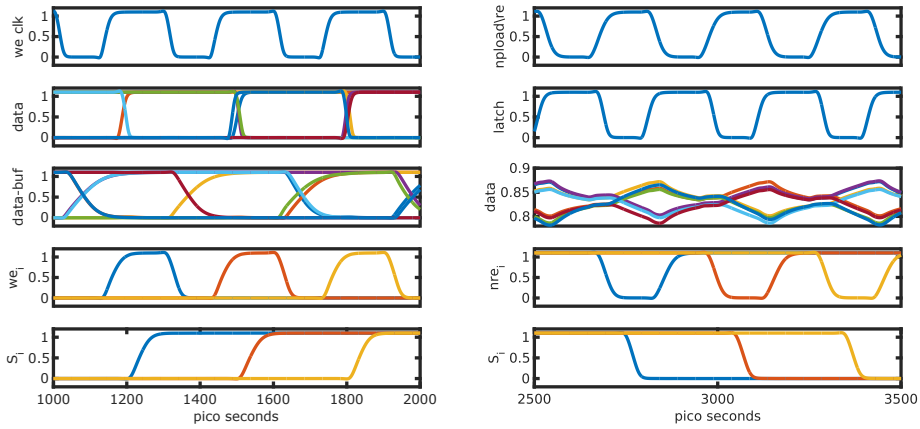
this condition is not reached, it guarantees that the TFIFO has at most 125 data and can be written. The TFIFO is almost empty in case exists at least one trio T_i such as $T_i = (0; 1; 0)$. If this condition is not met, it is enough to ensure that the TFIFO has at least 2 positions filled and it can be read. This non-centralized system, along with the circular arrangement of the drivers, is compact, thus lowering power and propagation times. Moreover, since there are no addresses, it becomes trivial to change the size of the bit array to fit any need. The uniformity of the layout enables the usage of scripts to perform place and route operations for any desired size, tested up to 128 positions.

2.2.4. Operation

Push operations: push operations require 2 steps: charge and store. In the first step, the input DL are set by the input drivers. In the second, the WE is asserted according to the earlier conditions and selected bit cells store the values. As opposed to a pop, the push request can be either a clock or any number of non-periodical pulses because the TFIFO was designed to store external random events from ToF sensors where the events obey the Poisson distribution. Thus, one pulse is enough to perform the two steps. The WD state gets updated at the end. Waveforms extracted from a post-layout simulation are shown in Fig. 2.26a.

Pop operations: pop operations, shown in Fig. 2.26b, are carried out in 3 steps: load, retrieve and latch. In the first step, the differential lines (DL) are connected to each other through PMOS transistors that equalize the voltages of the parasitic capacitances to Vdd - Vth/2. Then, RE is asserted and the enabled bit cells start charging the DL with their stored voltage. Lastly, the sense amplifiers are enabled and they assess the polarity of the DLs, and latch their values. It takes half clock cycle for each step, thus 1.5 cycles for a pop. However, when several pops are requested, they can be chained in a pipeline, thus reducing the pop to 1 clock cycle with half a clock latency.

Almost full and Almost empty circuitries: Circuitry to perform almost-full and almost-empty logic is included in the figure. Fig. 2.27 shows the circuitry for Almost-full (AF) and Almost-empty (AE) functions. Every WD has a NOR gate to check the conditions already mentioned $T_i = (1; 0; 1)$. The outputs of the NOR gates are ORed with a weak pull-up to speed up the response. If the condition is no longer met, the NMOS switches off and the PMOS pulls up the line; this might take several clock cycles. When the TFIFO is working at very high speeds (input fast=1), a complementary circuit strengthens the pull-up only after the AF flag was set for a fast recovery and keeps a weak pull-up the rest of the time. This lowers the time for the flag to less than 1 clock cycle at the price of a higher power consumption. Alternatively, it is possible to suppress the either AF or AE functionality in order to save power for systems where the length of data is known beforehand. In image sensors the AF flag is usually dismissed since the information will be lost in any case as it means the read-out block is saturated and cannot take any further data at the moment. This is achieved by tying the pull-up voltage to VDD, then the shared line



(a) Post-layout simulation of a typical waveform for push operations running at 3.33GHz, write signal can be either a clock or random pulses with a space of half a clock in between.

(b) Post-layout simulation of a typical waveform for pop operations running at 3.33GHz, read signal should be periodic like a clock or random pulses with a space of half riodic.

Figure 2.26: Simulations of write and read operations.

(almost-full-b) remains in low state all the time and no power is taken. This can be done independently for AE flag.

The waveforms for the almost-full circuitry are shown in Fig. 2.28. The shared line (almost-full-b) has a pull-up to VDD that keeps it at high state. Any word driver can pull it down once the almost-full condition is met. If the input fast = '1', the pulse goes to the flip-flop clock input and sets Q to '0'; as a consequence, the strong pull-up is activated and the system is ready to react when the transistor that originated the sequence goes to high impedance. The line is quickly recovered and the negative-edge detector sends a short pulse to the flip-flop, thus asserting Q and the strong pull-up is released. This system ensures fast speed in the line for both the flanks of the almost-full and almost-empty flags. This circuit enables the flag circuits to react in a time of 225ps, that corresponds to about one clock cycle when the memory operates at its maximum frequency. The flag remains up for at least one more clock, that is also needed by an external circuit to resynchronize the flag signal and avoid metastability.

Reset function: after power-on, the states of the WDs are unknown; therefore initialization is needed. WDs can be switched from *used* state to *unused* or vice-versa but cannot be set to a specific state. This is because the set-reset circuitry was avoided to increase performance and reduce area. For initialization, a burst of at least 128 pop pulses must be sent. By design, the system is a FIFO memory and as such, pop operations are destructive; thus, all the WDs will switch to *unused* state after 128 pop pulses. If the TFIFO becomes empty before 128 pulses, which is the most likely scenario, no further read operations will be attended as all states

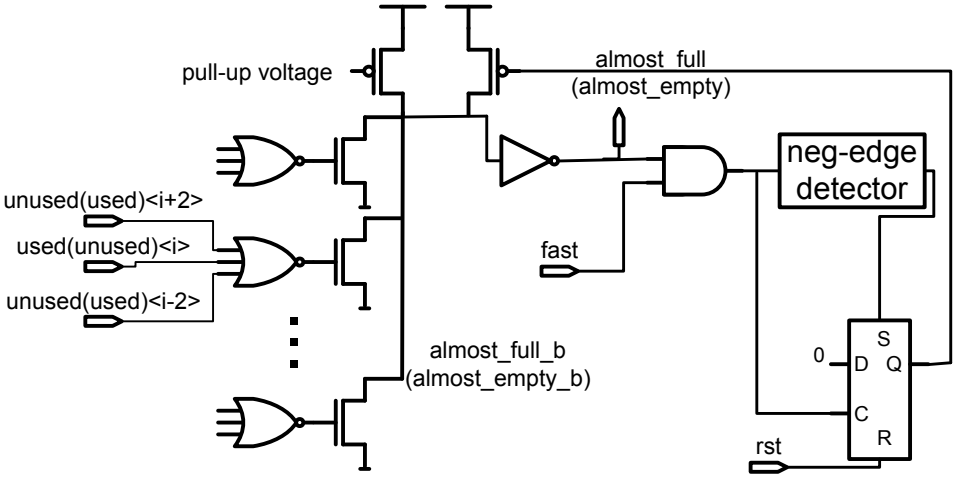


Figure 2.27: Circuit for almost-full (almost-empty) flags.

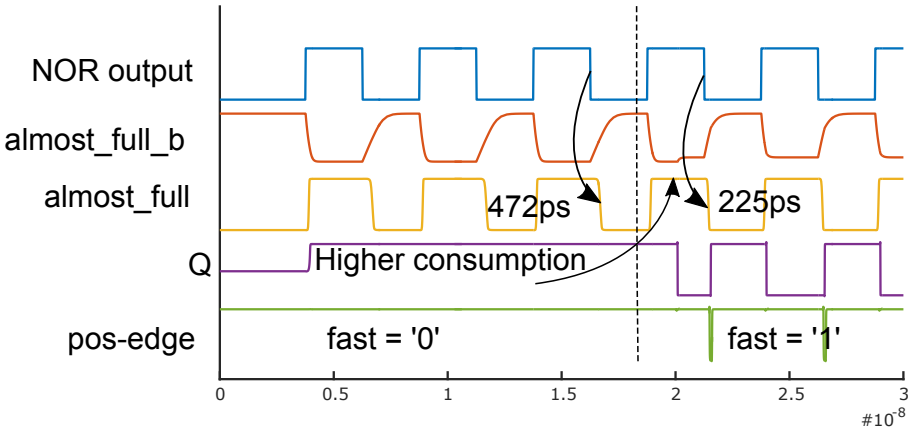


Figure 2.28: Post-layout simulation for almost-full circuit.

are *unused*; $D_i = (0;0)\forall i$. Finally, the reset signal only changes the state of the position 0 to *used*. As a consequence $D_0 = (0;1)$ and $T_0 = (0;1;0)$; the TFIFO is ready to operate. Fig. 2.29 shows a typical reset sequence, followed by push and pop sequences. In the picture, only the first 16 positions are shown for clarity. The first stage is a random waking-up occupation scenario of the TFIFO after power-on. This distribution of *unused* and *used* states is corrupted and cannot occur during normal operation. After every read operation request that the memory receives, every malformed sub-FIFO (every consecutive group of word drivers whose states are *used*) will change their most-top word driver state to *unused*. The number of read requests that it takes to make the TFIFO empty equals the number of memory positions of the longest malformed sub-FIFO. The module that performs this reset sequence can read the almost-empty flag to stop the operation or repeat this operation 128 times so that any malformed sub-FIFO will be erased. No matter the method chosen, after all those read requests, all the word drivers will be in *unused* state. Therefore one pulse to change the state of the first word driver is required and the memory is ready for operation. In the same picture, the progress of occupancy of the TFIFO is shown when 10 write requests are performed, followed by 10 read requests.

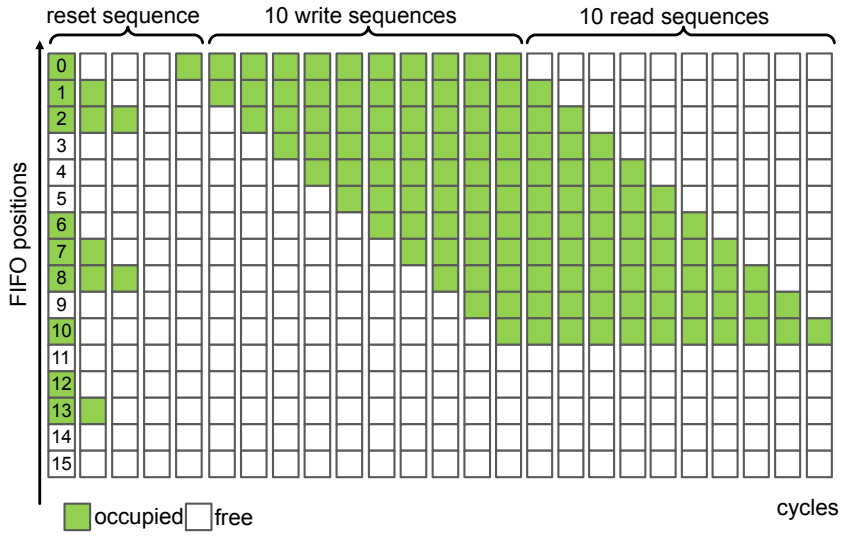
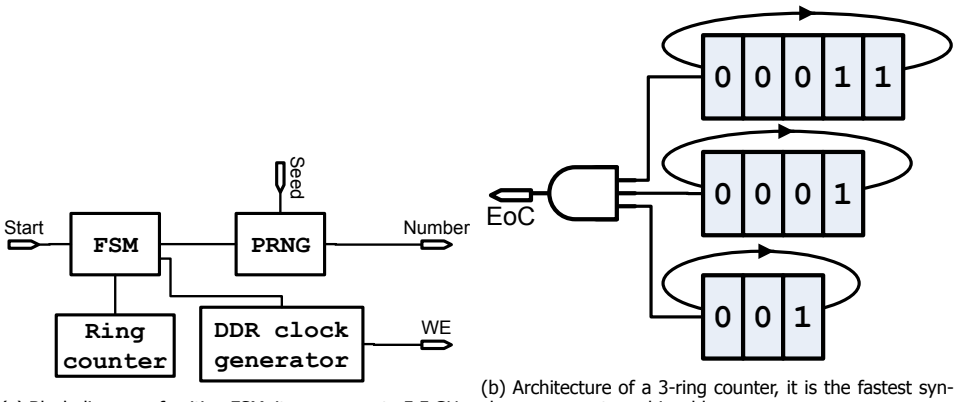


Figure 2.29: Reset, write and read sequences for a TFIFO.

2.2.5. Results

Testing conditions: the TFIFO was tested on chip through 2 finite state machines (FSM) to perform push and pop operations. For test purposes, their voltages can be set independently from the TFIFO. The write FSM is a pseudorandom number generator (PRNG) externally configured with an initial seed and a programmable



(a) Block diagram of writing FSM, it can run up to 5.5 GHz.

(b) Architecture of a 3-ring counter, it is the fastest synchronous counter achievable.

Figure 2.30: Ad-hoc testing electronics used in the measurements.

number of pushes. The read FSM performs the pop operations and saves the data to an external FPGA. It can be programmed in run-time to set the number of pops to perform. The block diagram of the write FSM is shown in Fig. 2.30a. It has two parameters: "Seed" that is the first number where the PRNG will start from and "Count" that is the number of words that will be pushed into the TFIFO. When "Start" is asserted the FSM sets the output to the first value and sends the WE signal. At every clock cycle, a new value from the pseudo-random sequence is generated according to the equation: $b_0^{next} = xor(b_6, b_1)$. The process continues until the counter has overflown. The counter is a special 3-ring fast counter, also purposed in this work, to achieve higher speeds than LFSR counters. The architecture is shown in Fig. 2.30b. Three shift registers of different length (3, 4 and 5) can be set to any number and run at the clock speed. Their MSBs are connected to an AND gate to check for End-of-Count. The speed of this type of counter is the maximum achievable for a synchronous system since the slowest component is simply a shift register which is the fastest synchronous component feasible in a digital system. The overflow flag will rise when all the MSBs are equal to '1'. A function is needed to provide the combinations of the three shift registers that will result in possible final count values. Some combinations are shown in table 2.1.

Both the FSMs were put together in the same block synthesized using Cadence tools. The layout shown in Fig. 2.31. Post-layout simulations confirmed that the FSMs can run up to 5.5GHz.

Frequency and core voltages were swept over their full range in order to check proper operation and to measure power. The results for both the modes fast and normal are shown in Fig. 2.32a and Fig. 2.32b respectively. The maximum frequency of operation and power consumption were measured for a lot of 10 chips. The results are shown in Fig. 2.33.

Comparison: for push operations, the TFIFO achieves a maximum speed of 4.3 GHz or alternatively 116 ps one-shot pulses can be used. For pop operations, the

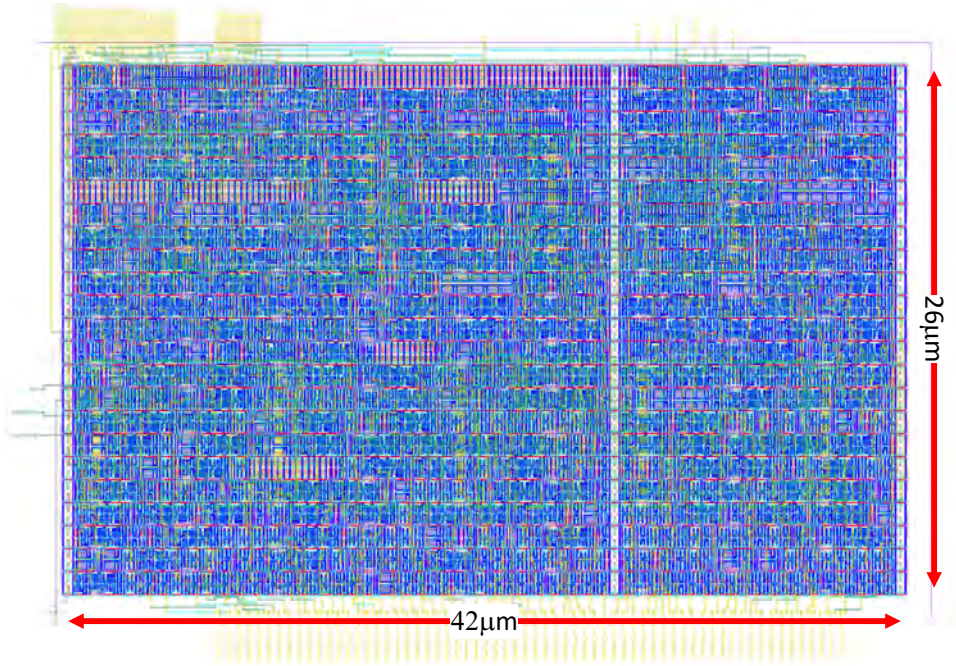
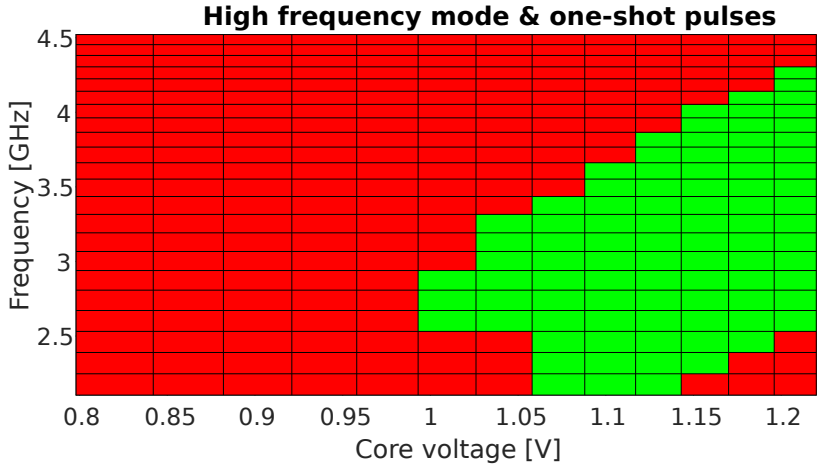
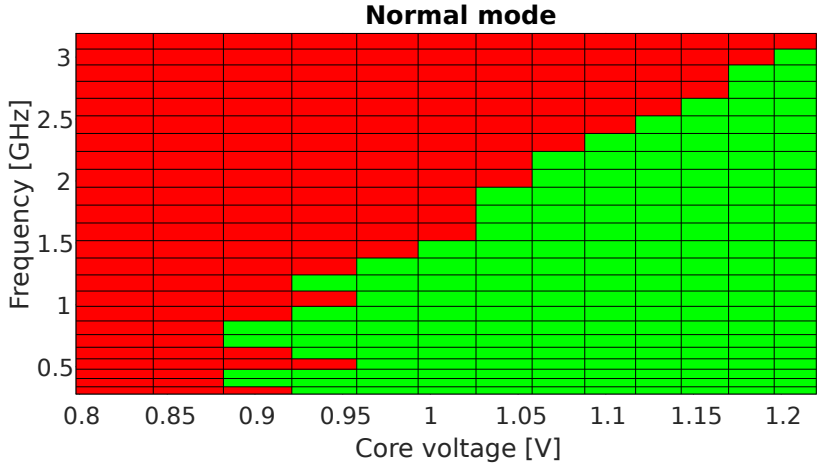


Figure 2.31: Layout of the FSMs. Maximum frequency of operation is 5.5GHz and skew for outputs is below 20 ps.



(a) Results for high-speed mode. Green: passed, red: failed.



(b) Results for low-speed mode. Green: passed, red: failed.

Figure 2.32: Measurements of TFIFO for both the modes low-speed and high-speed.

Number	5-bit	4-bit	3-bit
1	11111	1111	111
5	00001	1111	111
12	11111	0001	001
15	00001	1111	001
20	00001	0001	111
60	00001	0001	001
∞	00000	0000	000

Table 2.1: table with some examples of final count

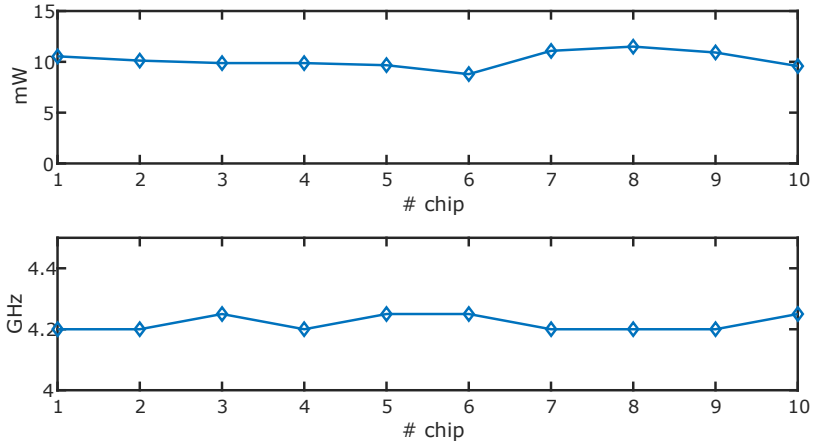


Figure 2.33: Nominal power (top) and Maximum frequency (bottom) for 10 chips.

maximum speed tops 4.2 GHz. The total data throughput achieved is 270 Gbit/sec per channel. These results include the jitter of the ring oscillator used during the tests. Further improvement can be done if a low-jitter PLL is used. The chip has a wide operating voltage of 0.85 V to 1.21 V. At nominal voltage, the TFIFO can reach 3.51 GHz for both read and write operations dissipating a power of 11.34 mW. The results passed/failed as a function of frequency and core voltage are shown in Fig. 2.32a and 2.32b. The comparison made on the TFIFO with the state of the art is shown below in Table 2.2. The table is split into two categories: low-power memories and high-speed memories. The TFIFO displays a speed of operation in the range of the fastest memories in even smaller technologies, yet maintaining low power consumption comparable to ultra low-power memories running at substantially lower frequencies.

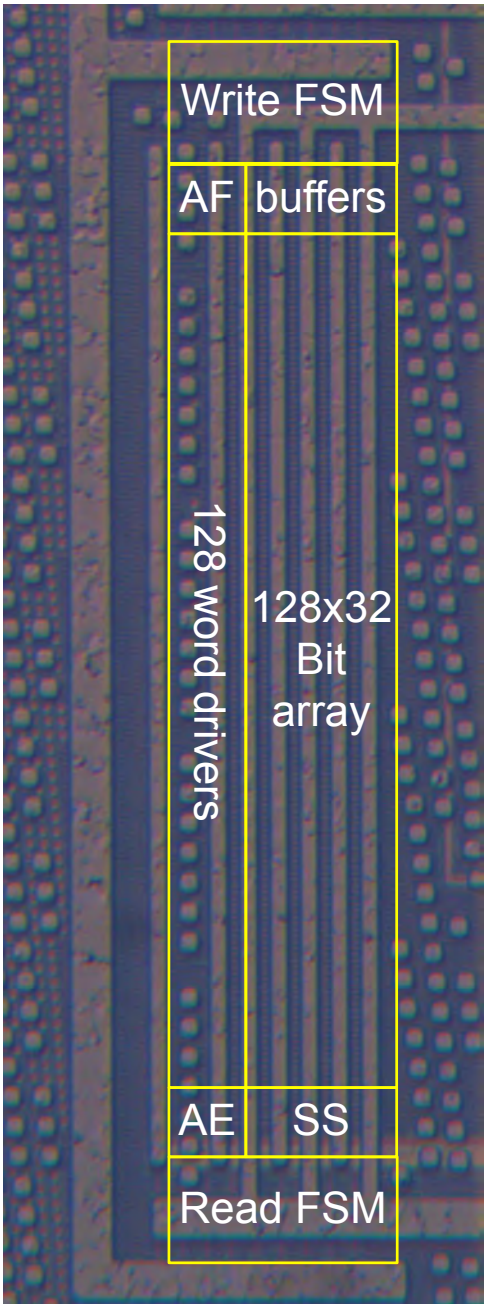


Figure 2.34: Micrograph of the memory. The modules are located in the same direction of the flow of the data.

Type	Ultra low-power FIFOs				High-speed memories				TFIFO
Ref.	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]	this work
Techn.	65nm	40nm	28nm	28nm	7nm	22nm	65nm	32nm	40nm
Topology	9T	9T	9T	12T	6T	6T	10T	24T	8T
Operation	1W or 1R	1W or 1R	1W or 1R	2W or 2R	1W or 1R	1W or 1R	1W + 1R	6R + 2W	1W + 1R
Min (V)	0.35	0.325	0.4	0.4	(3)	0.6	0.7	0.4	0.85
M. size(kb)	72	72	4	4	18	128	1	4	4
Freq. (Hz)	230K	600K	32M	10M	5.3G*	4.6G	1.9G*	4G	**4.2G***
N. Pwr(1)	0.24	0.13	0.09	0.19	(3)	(3)	27	2.34	**0.67***
N. Area(2)	3111	2731	7816	2736	(3)	152	8600	12318	**1701**
Domains(4)	1	1	1	1	1	1	2	1	2

Table 2.2: Comparison between TFIFO and state-of-the-art memories for low-power and high-speed performance. (1) Normalized power: $[\mu W/(kb.MHz)]$ (2) Normalized Area: $[\mu m^2/kbit]$ (3) not reported (4) number of clock domains, very important for asynchronous operation. (*) over voltage

2.3. Conclusion:

On the read-out systems: the chips designed in this work make use of the read-out systems and communication protocol that have been introduced and explained in this chapter. The general scheme and protocol are technology and even module independent as they can be fully reconfigurable and set for a particular application. In this way, Concolor (2D 40nm ST node for PET), Panther (3D 40nm ST node for LiDAR) and MindHive (2D 40nm TSMC for vision) have a share platform for their read-out systems. FPGA-based read-out systems were also proved to work for Endotof. This was the preferred option at the experimental stage of any chip design.

On the TFIFO: the TFIFO concept, as a new type of memory, was designed to cope with the usual trade-off that exists between speed and power in SRAM memories, specially for image sensors. The TFIFO was successfully tested and the achieved results demonstrate the system outperforms SRAM memories for FIFO operations.

The advantages attained by this design are listed below:

- Higher frequency of operation than SRAM memories for a given technology.
- Lower power consumption.
- No need of PLL or periodic write signals.
- Resizeable by-design: it can be sized for any number of words, even not-power-of-2 or odd numbers.

References

[1] A. Carimatto, S. Mandai, E. Venialgo, T. Gong, G. Borghi, D. R. Schaart, and E. Charbon, 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.

[2] A. Amara, F. Amiel, and T. Ea, FPGA vs. ASIC for low power applications, Microelectronics Journal **37**, 669 (2006).

[3] H. M. Sayed, S. A. Taie, and R. A. El-Khoribi, An improved technique for LIDAR data reduction, in *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)* (2016) pp. 1–4.

[4] V. Cao, K. Chu, N. Le-Khac, M. Kechadi, D. Laefer, and L. Truong-Hong, Toward a new approach for massive LiDAR data processing, in *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)* (2015) pp. 135–140.

[5] N. Aubry, E. Auffray, F. B. Mimoun, N. Brillouet, R. Bugalho, E. Charbon, O. Charles, D. Cortinovis, P. Courday, A. Cserkaszky, C. Damon, K. Doroud, J. M. Fischer, G. Fornaro, J. M. Fourmigue, B. Frisch, B. Fürst, J. Gardiazabal, K. Gadow, E. Garutti, C. Gaston, A. Gil-Ortiz, E. Guedj, T. Harion, P. Jarron, J. Kabadanian, T. Lasser, R. Laugier, P. Lecoq, D. Lombardo, S. Mandai, E. Mas, T. Meyer, O. Mundler, N. Navab, C. Ortigão, M. Paganoni, D. Perrodin, M. Pizzichemi, J. O. Prior, T. Reichl, M. Reinecke, M. Rolo, H. C. Schultz-Coulon, M. Schwaiger, W. Shen, A. Silenzi, J. C. Silva, R. Silva, I. S. Schweiger, R. Stamen, J. Traub, J. Varela, V. Veckalns, V. Vidal, J. Vishwas, T. Wendler, C. Xu, S. Ziegler, and M. Zvolsky, EndoTOFPET-US: a novel multimodal tool for endoscopy and positron emission tomography, Journal of Instrumentation **8**, C04002 (2013).

[6] S. Mandai and E. Charbon, Multi-channel digital SiPMs: Concept, analysis and implementation, in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)* (2012) pp. 1840–1844.

[7] S. Mandai and E. Charbon, A 4x4x416 digital SiPM array with 192 TDCs for multiple high-resolution timestamp acquisition, Journal of Instrumentation **8**, 1 (2013).

[8] M. Fishburn and E. Charbon, System Tradeoffs in Gamma-Ray Detection Utilizing SPAD Arrays and Scintillators, Nuclear Science, IEEE Transactions on **57**, 2549 (2010).

[9] 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 **57 7**, 1797 (2012).

[10] 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 **54**, 1137 (2019).

[11] M. Ren, E. Wu, Y. Liang, G. Wu, and H. Zeng, A quantum random number generator based on photon number resolving detection of successive photon pulses, in *2012 Conference on Lasers and Electro-Optics (CLEO)* (2012) pp. 1–2.

[12] S. L. M. Hassan, I. Dayah, and I. S. A. Halim, Comparative study on 8T SRAM with different type of sense amplifier, in *2014 IEEE International Conference on Semiconductor Electronics (ICSE2014)* (2014) pp. 321–324.

[13] I. J. Chang, J.-J. Kim, S. P. Park, and K. Roy, A 32 kb 10T Sub-Threshold SRAM Array With Bit-Interleaving and Differential Read Scheme in 90 nm CMOS, IEEE Journal of Solid-State Circuits **44**, 650 (2009).

[14] M. Tu, J. Lin, M. Tsai, S. Jou, and C. Chuang, Single-Ended Subthreshold SRAM With Asymmetrical Write/Read-Assist, IEEE Transactions on Circuits and Systems I: Regular Papers **57**, 3039 (2010).

[15] Mu-Tien Chang, Po-Tsang Huang, and Wei Hwang, A robust ultra-low power asynchronous FIFO memory with self-adaptive power control, in *2008 IEEE International SOC Conference* (2008) pp. 175–178.

[16] W. Du, Po-Tsang Huang, Ming-Hung Chang, and Wei Hwang, A 2kb built-in row-controlled dynamic voltage scaling near-/sub-threshold FIFO memory for WBANs, in *Proceedings of Technical Program of 2012 VLSI Design, Automation and Test* (2012) pp. 1–4.

[17] M. Clinton, R. Singh, M. Tsai, S. Zhang, B. Sheffield, and J. Chang, A 5GHz 7nm L1 cache memory compiler for high-speed computing and mobile applications, in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)* (2018) pp. 200–201.

[18] E. Karl, Y. Wang, Y. Ng, Z. Guo, F. Hamzaoglu, U. Bhattacharya, K. Zhang, K. Mistry, and M. Bohr, A 4.6GHz 162Mb SRAM design in 22nm tri-gate CMOS technology with integrated active VMIN-enhancing assist circuitry, in *2012 IEEE International Solid-State Circuits Conference* (2012) pp. 230–232.

[19] D. Lee and J. Kim, 5 GHz all-digital delay-locked loop for future memory systems beyond double data rate 4 synchronous dynamic random access memory, Electronics Letters **51**, 1973 (2015).

[20] S. Ataei, M. Gaalswyk, and J. E. Stine, A high performance multi-port SRAM for low voltage shared memory systems in 32 nm CMOS, in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)* (2017) pp. 1236–1239.

3

High Speed electronics used in image sensors II: intensity and timing

"Our virtues and our failings are inseparable, like force and matter. When they separate, man is no more"

Nikola Tesla

"I could trust a fact and always cross-question an assertion."

Michael Faraday

"A chain is as weak as the weakest of its links". Systems, in particular image sensors are definitely not an exception of this rule. SPADs, used as sensitive devices, deliver information about the intensity and timing of the arriving photons. The back-end electronics that handle all those data must be fast and accurate enough in order to not degrade the resolution of the whole system. In this chapter, the main problems of big image sensors regarding intensity and timing will be discussed and analyzed.

In this chapter, different strategies for read-out systems and protocols to tackle particular problems encountered in image sensors will be discussed. Generally, image sensors have massive amount of information to be transmitted to and collected by an external system that usually performs the processing according to the needs of the aimed application. This information is usually composed by detected intensity, timing information, addresses of events, reference values and various variable or fixed registers. In this chapter, the specific problems related to read-out systems for image sensors will be presented and explained from holistic point of view.

3

3.1. Timing signals - clock distribution

3.1.1. Important aspects: Skew-Power-Frequency

It is common knowledge that the performance of large digital systems heavily depends on the clock distribution network. Several aspects, such as skew, power and frequency have to be considered since they have direct impact in the final overall performance. The maximum frequency that a digital circuit can operate at is expressed by Eq. 3.1.

$$f_{MAX} = 1/(t_{CQ} + t_{COMB} + t_{WIRE} + t_{SETUP}) \tag{3.1}$$

where t_{CQ} is the delay of a flip-flop (clock-to-Q), t_{COMB} is the delay of combinational circuits between registers, t_{WIRE} is the delay on the wires, t_{SETUP} is the set-up time of the flip-flops.

Given two sequential elements connected in pipeline mode, the skew is defined as the difference in the arrival-time of the clock for each element $tSKEW = t_f - t_i$. The skew can be positive or negative. For the first case, the maximum frequency of operation is diminished since the second element is getting the clock edge before than the first element, thus that time is wasted. If the skew is negative, the maximum frequency is not affected. However, the data cannot change during the setup or hold times of the second sequential element. Therefore, the condition $t_{CQ} + t_{COMB} + t_{WIRE} > t - t_{SKEW} + t_{HOLD}$ needs to be fulfilled. Mismatch and PVT variations can change the value of the skew in either direction. While positive skews will impact only in the maximum frequency of the system, negative skews could cause fatal errors and make the system stop working under certain conditions. Hence the necessity of designing clock networks that account for this.

3.1.2. Distribution methods

As mentioned in the previous section, skew and power are crucial for any digital system; large efforts have been put on in order to reduce them. Authors have examined several aspects of the trees to go towards skew-free and low-power clock networks.

3.1.3. Aspects of clock trees

skew: the approaches to reduce skew can essentially be split into two categories: physical arrangement of the tree and physical arrangement of the components

(namely buffers and repeaters) that transfer the clock signals to the lower layers of the tree.

For the first category, some works focus on the shape of the network, as it is explained in [1]. Other authors emphasize the importance of the routing method of these networks as shown in [2].

For the second category, it is known that buffers play a very important role in the clock network; different methods have been studied to mitigate their contribution to the total skew. As the authors explain in [3], buffers with adjustable delay can be employed for skew reduction. Other works, like [4], sinusoidal buffers are used in order to reduce power consumption by 20%.

In many systems, due to gating method (method to reduce power), various types of gates are needed and this goes in detriment of the skew of the clock tree as it is shown in [5] along with a method to overcome this issue.

power: power consumption (PC) is another very important feature of any clock distribution network. In any digital system, it can be expressed as $P = afCV^2$ where a is the activity coefficient (from 0 to 1), f the frequency of operation, C the capacitance that is charged and discharged in every cycle and V is the voltage swing. Therefore, all the methods that can be found in the literature approach the solution by aiming at reducing one of those variables involved. The reader can notice that some approaches can be more effective than others according to which target variable they aim at and by how much amount.

Gating is a well-known method to reduce the PC; it reduces the activity coefficient a and the PC will reduce proportionally. As a follow-up of this general idea, the authors of [6] explain that interleaving AND and OR gates based on the activity likelihood of the modules the clock is feeding to, it is possible to reduce the PC by a 11%.

As can be noticed, the variable V has a quadratic influence on the PC. The authors of [7] show that PC is reduced by 30% when standing wave signals are employed for clock distribution by means of inductive terminations on the lines. In [8], it is shown that low-threshold repeaters can reduce the PC by 10% in average. Another interesting approach is shown in [9], where they use a 2D network based on mutual-injection-locked ring oscillators. This method was used in one of the designs, thoroughly explained in chapter 4.

3.1.4. Clock distribution in image sensors: Self-generated clock

As shown in the previous subsection, there are many methods to reduce the PC. Nonetheless, when the system under design is an image sensor, the constraints of image sensors make the methods not completely suitable or feasible to be implemented. These constraints usually include: utilized area, shape factor, fill factor, the interaction between analog and digital domains, the desirable uniformity in heat dissipation and the intermittent activity of the subsystems.

In this thesis, a new method for clock distribution for image sensors is presented. Coherently to what has been explained, this method tries to lower the PC by reducing the variables that affect it while meeting the aforementioned inherit

constraints of image sensors. Activity coefficient (a) and frequency (f) of operation were of particular interest for the design of the purposed method.

In this work, the concept of Selfclock Generation Network (SCGN) is introduced. As an alternative of a clock network that functions at the desired frequency, SCGNs provide an efficient way to deliver high-frequency clocks to digital systems by using global low-frequency clocks. This is achieved my means of local and independent ring oscillators that are activated only when needed. In addition, since every self-generated unit is programmable, it can be adjusted according to the need of the submodule. The features of this clock distribution network are listed below.

- The maximum frequency is not propagated along the chip; only a much slower submultiple of it.
- Digital systems are clocked by the exact number of cycles they need to operate. No need of enable signals or gating clock.
- Different frequencies can be used simultaneously in the chip, thus enabling critical circuits to work at much higher speeds than the rest of the circuits.
- Frequencies, number of clocks and synchronization can be changed on-the-fly by adjusting configuration parameters.
- Big modules can be itemized into smaller components that receive the clock signal independently; thus reducing the intramodular activity.

Because the frequencies of the ring oscillators are programmable, the phases are not meant to be aligned and the system (from intermodule point of view) is never synchronous. This is true even when those frequencies are set to the same value because there is no physical connection between them or phase locking module. Full synchronization is achieved at expenses of one extra clock. It should be remarked that this extra clock behaves as a wait state and does not activate any subsystem or logic, thus affecting execution time but PC is kept.

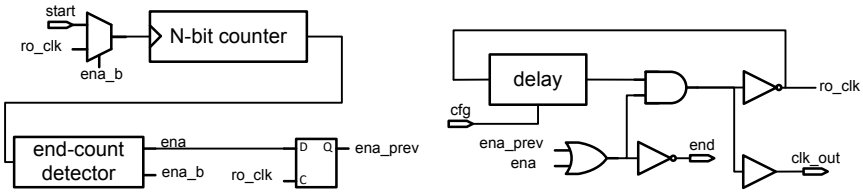


Figure 3.1: Complete self-generated clock diagram. Driver of the generator (left) and VCO (right).

General scheme and single operation

A complete block diagram is shown in Fig. 3.1. The block comprises two main subsystems: a programmable ring oscillator (RO) and a control driver (CD). The RO frequency can be configured by the parameter 'cfg'. The CD activates the RO in order to generate the number of clocks defined by the parameter 'Count'. The waveforms for single operation are shown in Fig. 3.3. The RO has a programmable delay component and a normally-off switch; therefore the RO does not oscillate

and the output is grounded. The CD has a counter whose clock source can be either the 'Start' signal when the counter has reached the maximum count, or the RO in any other case. The 'Start' signal initiates the process by resetting the value of counter. In this condition, the final-count detector asserts the signal 'Ena' with two purposes: it closes the switch of the RO, and sets the RO as the source of the clock for the counter. The oscillation starts and the output of the RO delivers the clock to the digital modules and the counter of the CD. This process continues until the counter has reached the maximum value. The final-count detector opens the switch of the RO and changes the source of the clock of the counter back to the 'Start' signal. As a consequence, the oscillation stops and the 'End' signal is asserted. This is the general block and the simplest mode of operation. The particular implementation for a given system under design should consider 3 different aspects discussed below:

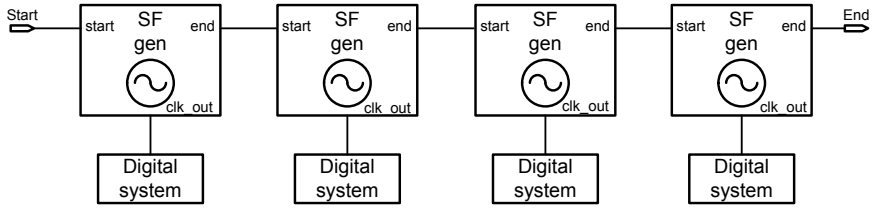


Figure 3.2: Self-generated clock serial connection.

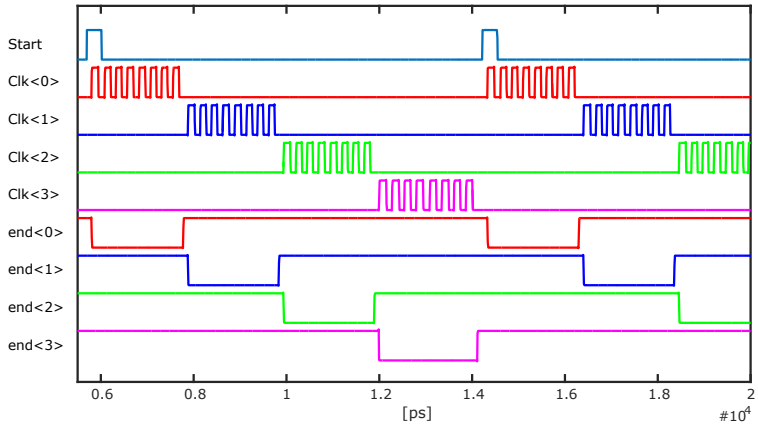


Figure 3.3: Self-generated clock: post-layout simulation for serial operation.

Frequency range : ring oscillators usually have two ways to control the frequency of oscillation. As shown in Fig. 3.4, the analog delay can be set by means of a control voltage. Alternatively, it is possible to set the rail-to-rail voltage to modify the delay time of that stage.

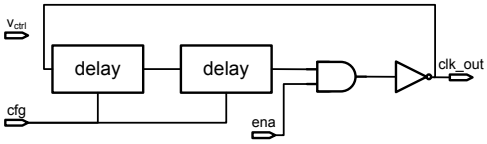
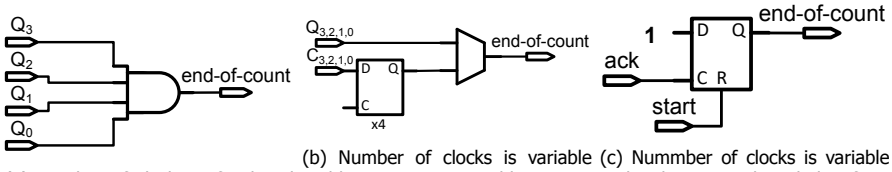


Figure 3.4: Scheme of the VCO. The delay is programmable and the start-stop function is controlled by an AND gate.



(a) Number of clocks is fixed and known, combinational logic can detect the end of count.

(b) Number of clocks is variable and known, programmable register and comparator detects the end of count

(c) Nummber of clocks is variable and unknown; acknowledge from the module is required to provide the end-of-count signal.

Figure 3.5: Different implementation of end-of-count detector according to the system the VCO is serving.

Number of clocks: every system has specific needs that make the number of clocks required dependent on arbitrary conditions. This number can be fixed, so the counter is connected to a fixed logic that detects the end-count or simply the overflow flag of the counter can be used. In case the number of clocks required is variable and known at run time, it is needed to place a programmable end-count detector that is composed by a register and comparator. Lastly, the system might need a variable and unknown number of clock cycles at run-time. In this situation the end-count detector should be replaced by a simple acknowledgement circuit operated by the design. Fig. 3.5 shows the three cases of end-count detector presented.

Serial and Pipeline operation

In serial structures as depicted in Fig. 3.2, where several components are cascaded, the 'End' signal should be used to start the next selfclock generator. The ROs of the serial cascade are not synchronous and the data might violate setup or hold constraints. For this reason, the internal delays of the module ensure that the results of the first digital subsystem be ready at the time the first clock arrives to the next subsystem. Consequently, synchronization is guaranteed at the expenses of half a clock cycle. The waveforms of this mode of operation are shown in Fig. 3.3. For pipeline mode, several 'Start' pulses are sent to the serial system with one clock cycle of slack. This can be very beneficial for systems that do not have high activity, yet they need to operate at full speed when it is required; a very good example for this being Neural Networks in image sensors, which will be discussed in the last chapter. The outputs of this mode of operation are shown in Fig. 3.6.

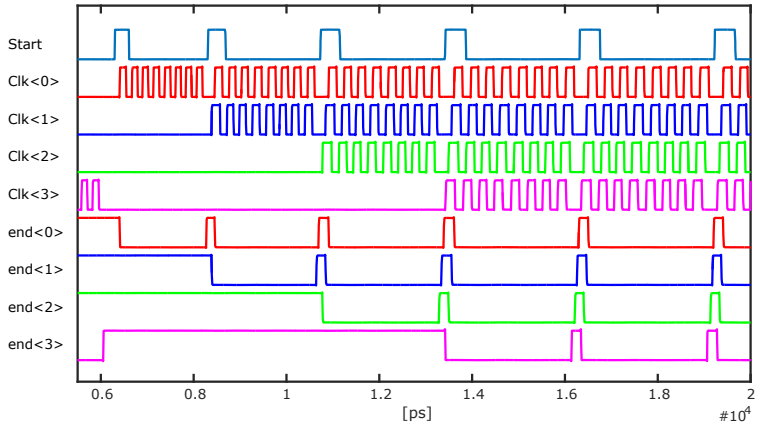


Figure 3.6: Self-generated clock post-layout simulation for pipeline operation.

Combined serial-parallel operation

For more complex systems, serial or pipeline operations schemes might not be enough since the interconnection of the modules can be a combination of serial and parallel circuits. For this case, a synchronizer needs to be interposed between one module and the next one in order to propagate the 'End' signal. A diagram of the serial-parallel system is shown in Fig. 3.7.

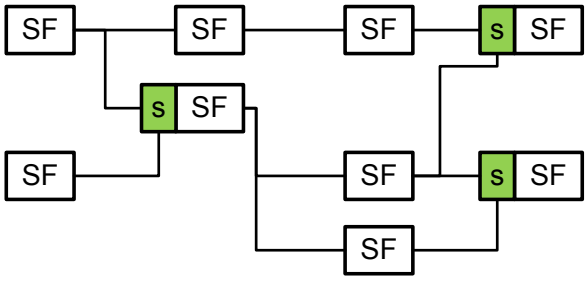


Figure 3.7: Example of serial, pipeline and parallel connections all combined.

Implementation for Neural Networks, methods and results

The implementation for the testing circuits comprises 4 self-generated clock modules used in full pipeline mode that feed 4 8-bit counters equally distributed along half millimeter distance on the chip. For testing purposes, the design has an end-of-count that falls into the first category (fixed and known number of cycles equal to 8). The layout is depicted in Fig. 3.8. The ring oscillator implemented has 4 stages; a configuration input can choose the number of delay elements are used and the control voltage can vary the frequency of operation. The measurements of the frequency as function of the input and voltage control is shown in Fig. 3.9.

Multiple start signal pulses in pipe-line mode were sent to the first of self-generated clock module of the chain, the output clocks feeds 8-bit counters that were read after the operation. For every start pulse, the counters increment by 8 since the number of clocks chosen was 8 because the neurons that have been implemented in MindHive require 8 cycles to operate. Results are shown in the table 3.1.

Start pulse	8-bit counter	Start Pulse	8-bit counter
1	8	30	240
2	16	31	248
3	24	32	0
4	32	33	8
5	40	34	16
6	48	35	24

Table 3.1: Some counter values for given number of start pulses sent to the system.

The system was tested in a wide voltage range and it is proven to work from $650mV$ to $1.21V$, considering $+10\%$ of over voltage. It also was tested for the two different speed modes.

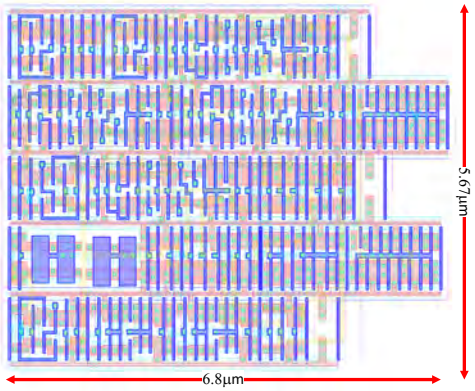


Figure 3.8: Self-generated clock layout. The compact design takes less than 35 μm^2.

3.2. Intensity

3.2.1. Discussion on hit Counting

One of the important features of image sensors is photon counting as it was explained before in the case of PET systems, where the photon counting is directly related to the energy resolution of the system. There are many methods that researchers have been using to tackle this specific problem. In [10], the hits are counted by 1-bit memory unit per pixel that retains the value until the end of the frame. Although the implementation is straight forward, the hit counting is lim-

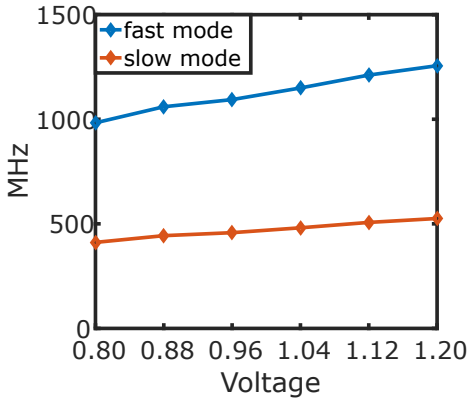


Figure 3.9: Measurements on frequency as a function of control voltage and mode of operation.

ited by the number of pixels or SPADs present in the sensor. The resolution of the sensor defines its hit counting capabilities. In other works, as in [11], the hits are counted by means of a distributed adder that is sampled by a clock running at 100MHz. The results are saved into a memory that works as a histogram. The hit counting capability of this type of systems is higher than in the previous case; however, the counting is synchronous with an arbitrary clock that can create artefacts when is fired by photons that are uncorrelated with the system. Alternatively, it is possible to use a counter that gets incremented as the hits arrive. This counter can be global or local according to the pixel clustering method employed. In either case, the counter has to be fed by a unique line that is fired by any of the several pixels that form the hit-counting cluster. There are basically two ways to carry this out: an OR tree and XOR tree; both analysed in [12] and [**?**]. In order to utilize OR trees, the dead time of the pixel has to be as small as possible not to pile up several hits resulting in undercounting. Optionally, a monostable can be added at the output of the pixel to lower the deadtime down to as much as hundreds of pico seconds. XOR trees do not have this problem as they change in every transition so that multiple fires do not get packed into only one pulse. An important difference is that rise and fall edges of XOR trees are equally important as they carry hit and timing information; this is a big disadvantage for XOR trees as designing a tree that propagates rise and fall edges with the same accuracy can be very challenging, if not impossible. Additionally, the counter at the end of the line has to be sensitive to both the flanks. In this work, it was decided to approach the problem in a segregated way, by using two different trees for hit counting and for event time-stamping. In the next section alternative gates for timing propagation are discussed.

3.2.2. Implementation of a XOR-based tree plus counter for hit counting

As a follow-up of what was explained in the previous section, if the tasks of hit counting and event time-stamping are split, XOR trees have great potential for the first case. The dead time of the XOR tree is as big as the propagation delay of the XOR gate. The SPADs were clustered by 64, forming a macro-pixel that feeds a 20-bit counter that is only rise-edge sensitive. This means that the actual number of photons is $N = 2n$, where n is the count reached by the counter. In case the n is and odd number, N will have an error of 1 hit. The diagram of the XOR tree is shown in Fig. 3.10. The propagation delay of one XOR gate is 40 ps.

Fig. 3.10

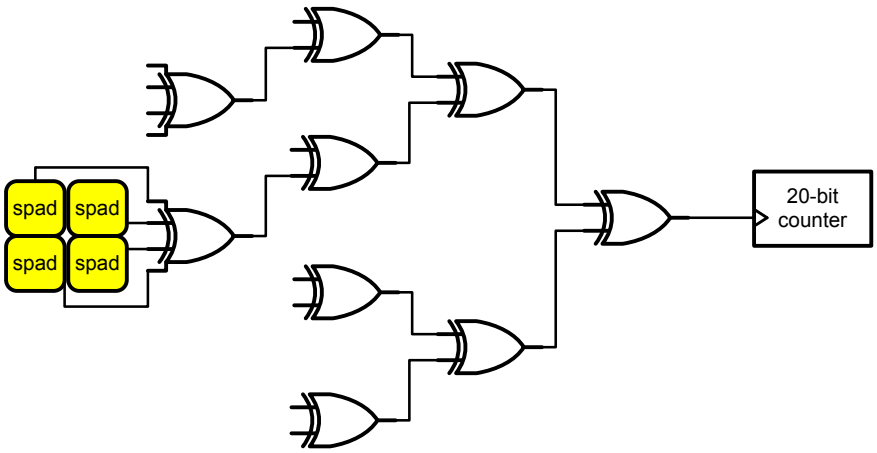


Figure 3.10: Scheme of XOR tree fed by multiple SPADs.

The deadtime of the SPADs are typically in the range of 10ns. This means that, under maximum light exposure conditions, each SPAD can generate up to 100 Mcps. If it is considered that they are clustered by 64 and the counter is sensitive only to the rise edge, there could be 3.2 Gcps at the input of the counter at most. Fig. 3.11 shows the circuit of the counter chosen for this design. It comprises 20 flip-flops that operate in asynchronous mode. For it, it was used the smallest flip-flop available in the technology. The counter attains a counting speed of 7.6 Gcps. In order to make it compatible with a read-out bus, the outputs are only available when OE is asserted; the outputs go to high-impedance otherwise. The layout is shown in Fig. 3.12.

3.3. Time resolution

The time chain on image sensors typically comprises the SPADs (the physical transducer device), the amplifier (first stage in contact with the SPADs), the time propagation lines (the mean the time pulses travel through) and finally the TDCs (the modules that time-stamp the events). The total time resolution or jitter will be given

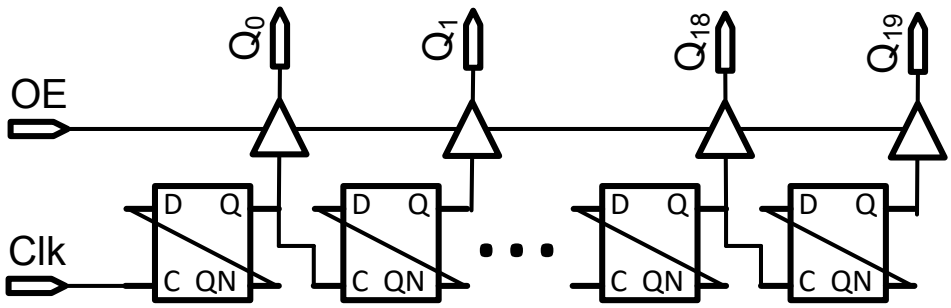


Figure 3.11: Schematic of 20-bit counter used in the design. Maximum number of counted photons is $2^20 * 2 = 2M$.

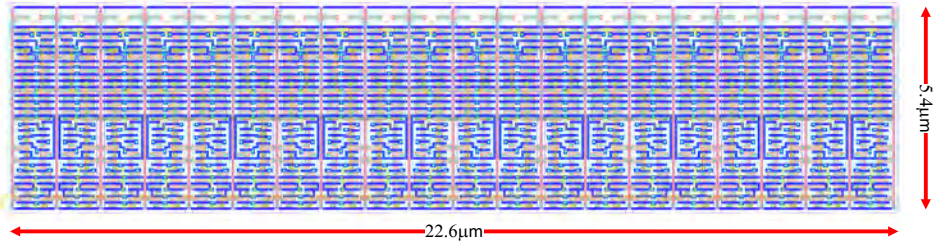


Figure 3.12: Layout of 20-bit counter used in the design.

by $J_T = \sqrt{J_{SPAD}^2 + J_{ampl}^2 + J_{line}^2 + J_{TDC}^2}$. Historically, the jitter of the SPAD has been dominant over the rest of the electronics; however, SPADs are being improved by the day by technology development teams and a jitter as low as 83ps have been recently reported [13]. In addition, due to the physics of the SPADs, it is possible to lower the jitter even further by lowering the threshold in the first amplifier connected to the SPAD, as explained in [14]. On the other hand, TDCs can achieve sub-10ps resolution [15]; as consequence, more than ever, the jitter of the time line plays a big role in the total time resolution. Furthermore, the fact that image sensors are increasing in size and pixel resolution [16] puts more and more pressure on the design of the time lines; this is not the case for the rest of the chain: SPADs, amplifiers and TDCs. A careful design of the employed time lines is needed to ensure the total jitter will not be dominated by other than the SPADs that are the ultimate physical limitation. In the following subsections, the amplifier and the timing lines will be discussed and different options of implementation shown. TDCs are explained in several chapters with particular implementations for the chips designed in this work.

3.3.1. Amplifier

Considerations for in-pixel amplifiers

There are big challenges for this design that need special care to be taken. The main characteristic of an amplifier that is used to generate a pulse when the SPAD fires is the activation threshold; as described, low-threshold amplifiers are the key to improve the SPAD time resolution. This feature is absolutely necessary to justify the use of amplifiers, which otherwise can be replaced by a simple inverter. Other important factors relevant to discuss are: the area, the shape factor and the power consumption. Since one amplifier per pixel is required, the area taken by the design is of big importance so as not to reduce the relative active area of the sensor. There are many SPAD sizes used in image sensors and they also vary according to the application they are used for. Typical values are $20x20\mu m^2$,$10x10\mu m^2$, etc.. In this thesis, the chosen size of the SPAD have been $9x18\mu m^2$ for Concolor, $18x18\mu m^2$ and $9x9\mu m^2$ for Panther, and $10\mu m$ diameter for MindHive. The shape of the amplifier plays also an important role since all the modules and components have to perfectly abut in order to not create dead or inactive areas; moreover for 2D technologies as the ones used for Concolor and MindHive. Last characteristic, but equally important, is the power consumption of the amplifier. Amplifiers are not digital cells that usually take only leakage power when not being used, but they also consume power for biasing and dynamic power during the switching event. Considering that the number of SPADs in the designs are 4096 for MindHive and Panther and 16384 for Concolor, the power consumption of the amplifiers can take big part of the total power budget. To sum up, in-pixel amplifiers design should pursue low-threshold activation, low power, low area and appropriate shape factor.

Implementation of in-pixel AC-coupled amplifier

Design: The schematic of the amplifier purposed, designed and tested in this work for Mindhive is shown in Fig. 3.13 and a simplified operation is shown in Fig. 3.14. The amplifier comprises 3 stages:

a. The first stage is a linear amplifier based on self-biased inverter that is AC coupled to the SPAD. The purpose of this first stage is the amplification of the signal coming from the SPAD. The input and the output of the first stage have the same DC voltage which is half of vdd-inv. The variable resistor, created by a NMOS-PMOS duo can be used to set the strength of the bias point. The optimal value of the resistance is a trade-off between the gain of the amplifier (the bigger resistance the better) and the recuperation time once a pulse is extinguished (the lower the better). The capacitor is a linear metallic capacitor of $10fF$. After a pulse is generated by a SPAD, the first stage generates an amplified inverted pulse with vdd-inv/2 as baseline. The gain of the first stage can be expressed by Eq. 3.2.

$$A = \frac{V_o}{V_i} = \frac{g_m^{NMOS} + g_m^{PMOS}}{g_o^{NMOS} + g_o^{NMOS} + g_L} \tag{3.2}$$

where g_L is the output load which corresponds to the input impedance of the first stage ($Z_{2i} = \frac{1}{SC_i}$, with $C_i = \propto C_{ox}WL$).

Since the N/PMOS of the amplifiers are working in strong inversion, the transconductance equations for NMOS and PMOS are as follows: $g_m^{NMOS} = \frac{2I_D}{V_{GS}-V_{TH}}$ and $g_m^{PMOS} = \frac{2I_D}{V_{SG}-V_{TH}}$

b. The second stage is composed by a low-threshold inverter that works in switching mode, fed by a different power supply. This inverter has its threshold below the DC point of the first stage and its output stays low when there is no input. As soon as output of the first stage crosses its threshold, the inverter switches to high level.

c. The third stage is a chain of 3 inverters that strengthens the signal by 8 times. The sizes of the transistors of the two first stages are very small and the signal needs a boost in order to go the first gate of the time tree, discussed in the next section.

In order to modify the threshold that the amplifier is sensitive to, there are three parameters that can be adjusted. The threshold of the second stage, since is built by a lvt-NMOS and hvt-PMOS, is around $1/3vdd$. The difference between the baseline of the first stage and the threshold of the second stage can be adjusted by changing vdd and vdd-inv. Similarly, the gain of the first-stage linear amplifier depends on vdd-inv and the feedback resistor, which can be easily adjusted by external voltages. Waveforms are shown in Fig. 3.14.

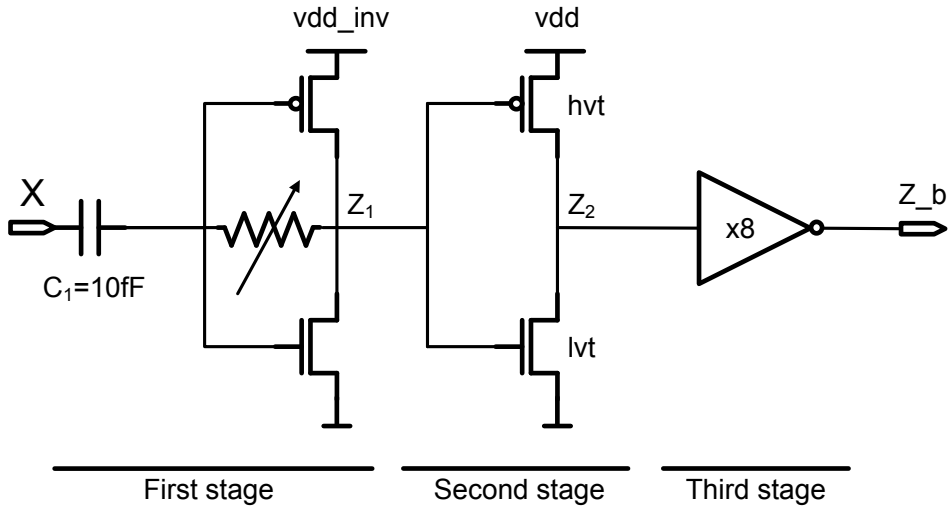


Figure 3.13: AC-coupled amplifier schematic.

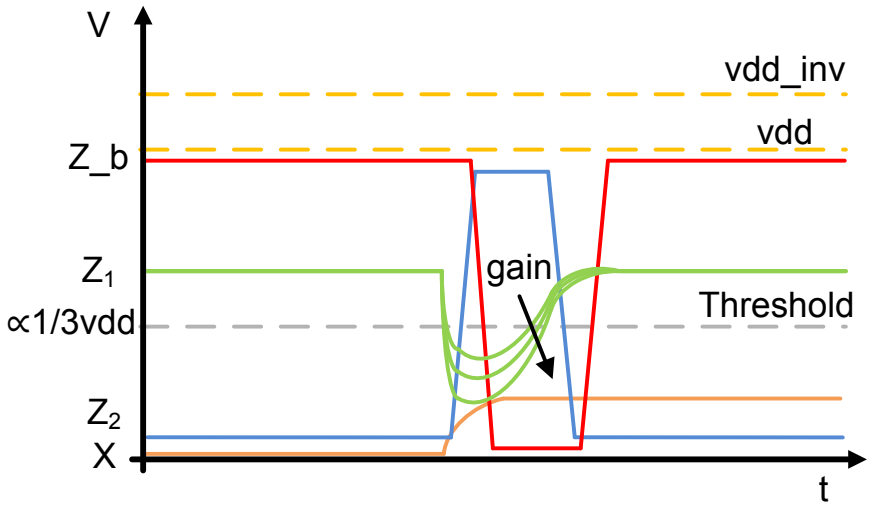


Figure 3.14: Simplified waveforms for the amplifier. X is the input signal, Z1 is the output of the first amplifier stage, Z2 is the output of the second amplifier stage and Z bar is the output of the third stage.

Layout: The layout of the amplifier is shown in Fig. 3.15. The size is $19.44\mu m^2$ and corresponds to 25% of the area of the SPAD that is serving. The layout, comprises two extra elements that are not shown in the schematic. A decoupling capacitor (C_2) was used for vdd to soften the effects of the swift switching behavior of the second stage. A standard inverter was added to provide pulse information to the intensity circuits or any other circuit where timing accuracy is not important.

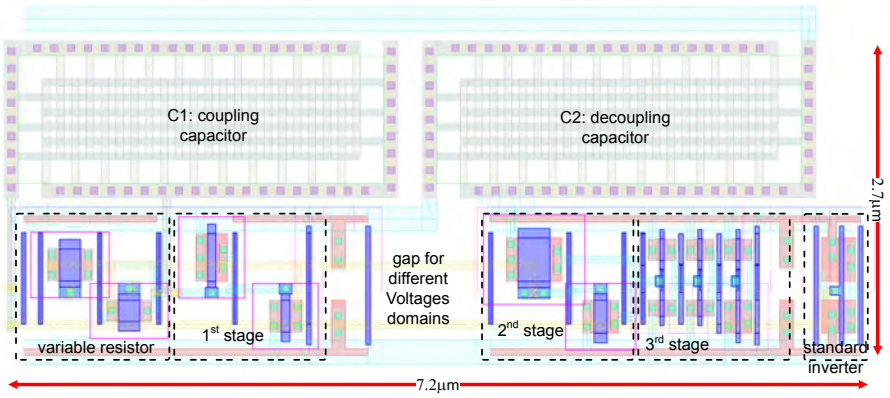


Figure 3.15: Layout of the AC-coupled amplifier.

The design was checked with a post-layout simulation to evaluate its response when it is excited with an input signal with jitter that is in the range of 0 to 400 ps. Since the amplifier was designed for input levels around 200 mV, the simulated input is constructed as follows

$$x(t) = x_p(t) + x_j(t), \tag{3.3}$$

where x_p is the pure step from 0 to 200 mV without any jitter, x_j is the part of the signal that is fully responsible for the jitter generated. This signal goes from 0 to 0.9V and has a jitter that goes from 0 to 400 ps.

The total amplitude of the input signal which jitter is added to is 1.1V, that corresponds to the excess bias voltage planned to use. The simulation, shown in Fig. 3.16, exhibits how much jitter the amplifier adds when it is stimulated with 200mV input signal.

Measurements: one standalone AC-coupled amplifier was measured and characterized. The input signal has 97 mV amplitude with rising time of 2.9ns. The shape of the output signal is deviated from the expected response; however the inflexion and the change in the slope can be explained by the very large off-chip inductance and large voltage swing of the digital buffers. Input and output have 10 nH of inductance when bonding wires, package and traces are accounted. These effects do not exist in the real situation since the amplifier is shortly connected to the SPAD and its output is connected to the timing line. The total jitter measured at

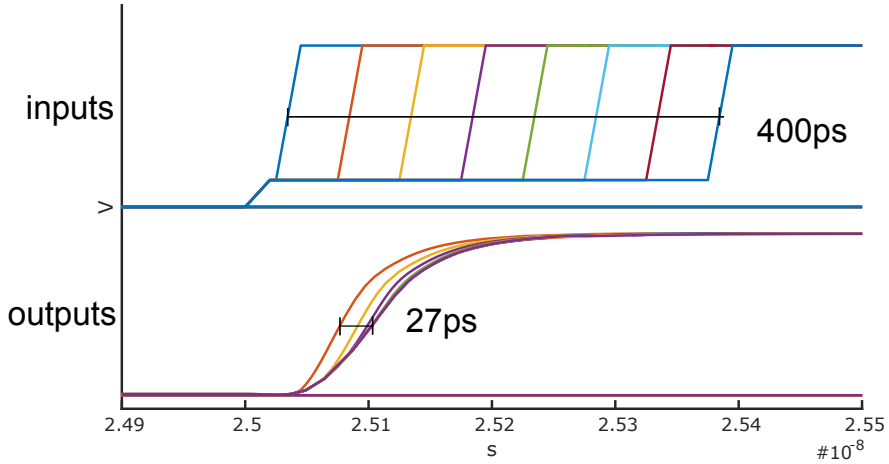


Figure 3.16: Post-layout simulation that considers an amplitude of 200 mV and jitter in the range of 0 to 400 ps.

the output signal was 14.17 ps. In the set-up used for this measurement, the total jitter can be expressed as follows: $J_T^2 = J_{BUFIN}^2 + J_{amp}^2 + J_{BUFOUT}^2$. Unfortunately, it is not possible to distinguish how the components of the chain contribute to the jitter in this set-up; hence, all the jitter was attributed to the amplifier in order to consider the worst-case scenario. Even when this consideration is made, the jitter is negligible respects to the jitter of the SPAD, $J_{amp} = 14.17ps << J_{SPAD}$. One of the measurements, for nominal voltages, is shown in Fig. 3.17.

3.3.2. Time lines

3.3.3. Time propagation trees

We said that XOR trees, though might be very convenient, are not suitable for timing lines because of the challenge that using both the rising and falling flanks might imply. But why is this?, when does it matter? To answer that question, it is required to understand the order of magnitude of time resolution of the gates used in these kind of propagation trees and their contribution in the total jitter.

Analysis of gates

Double-edge sensitive gates: gates usually do not respond to both the edges in the same way. This is because the internal parasitic capacitances of the nodes and the strength of the transistors vary with the voltage levels, thus leading to different dynamic circuits, consequently, leading to different time propagation. In Fig. 3.19, a standard schematic for XOR is shown. Simulations show that the difference between rising and falling edge caused by inputs A and B can be as large as $30ps$ which would be catastrophic for the desired time resolution. Despite the fact that this hypothesis tightly depends on the concrete circuit for a given gate, it is not doable in practice to achieve the same response for both the flanks. This is

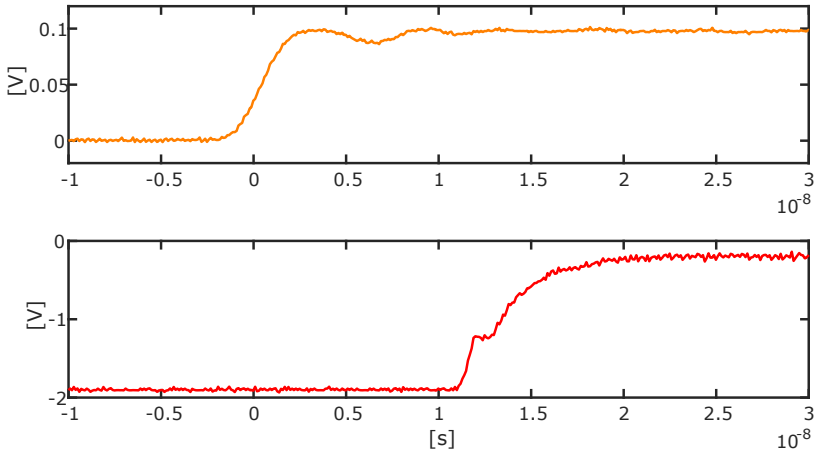


Figure 3.17: Step response of the amplifier. Input signal (top) and Output signal (bottom) of the amplifier.

because a fundamental problem that XOR gates experience. The equation of XOR gates is not linearly separable, therefore the inputs not always pull the output to the same potential (vdd or gnd) unlike other gates, as shown in Fig. 3.18.

Many XOR gates with different topologies have been analysed in this work and the results have been quite similar. Therefore, in implementations where the time resolution is very important, gates that respond to both the flanks are to be avoided. The obvious alternate solution is the rise-edge sensitive gates. More topologies tested in this work are described in [17].

Single-edge sensitive gates: balanced vs unbalanced: in this case, the input signal has always the same edge, but it is alternatively applied in a different input of the gates. Gates usually, unless specifically designed for this, do not respond in the same way for all of their inputs. fig. 3.20. shows the schematic of the OR gate, used in this work as single-edge sensitive gate. The schematic is not the most important thing from the point of view of parasitics, but the layout is. After performing a post-layout simulation, it was observed for this technology, that the OR gate has a skew of 5ps between its inputs. Even when 6ps does not seem to be much, the reader should remember that TDCs employed in image sensors can easily have sub-10ps resolution. The layout of the standard cell was modified in this work in order to match the parasitic capacitances at the highest level that the process allows, and to enhance the rising edge over the falling edge. The same post-layout simulation was performed, achieving 2ps skew, that actually is the resolution of the simulator. The layout of the modified version of the standard OR is shown in Fig. 3.21 and the results of the post-layout simulation is shown in Fig. 3.22.

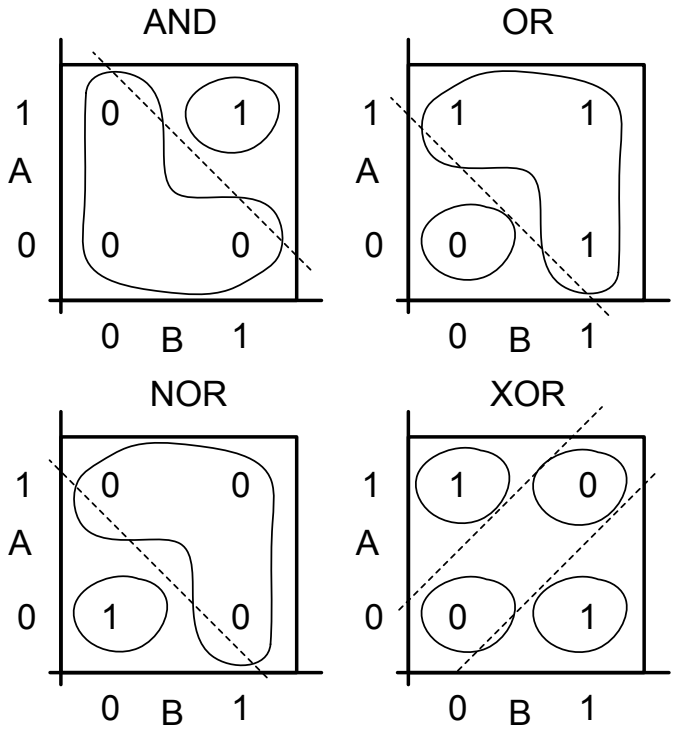


Figure 3.18: Linear gates vs non-linear gates.

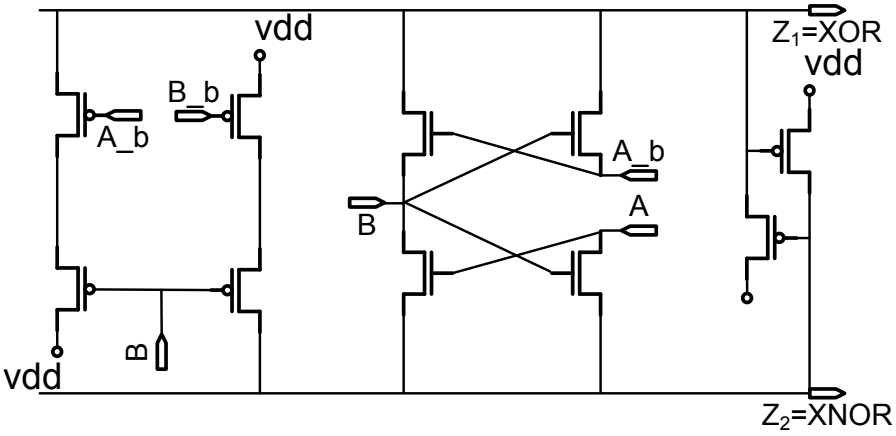


Figure 3.19: XOR general schematic.

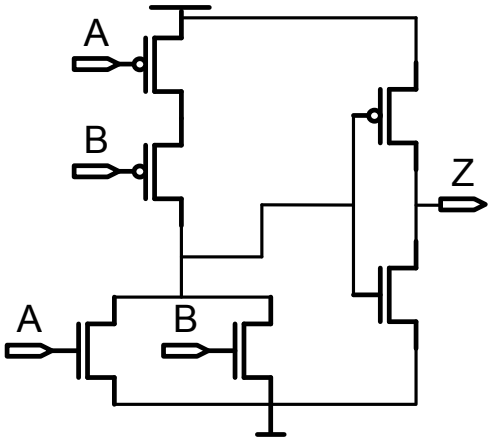


Figure 3.20: OR schematic.

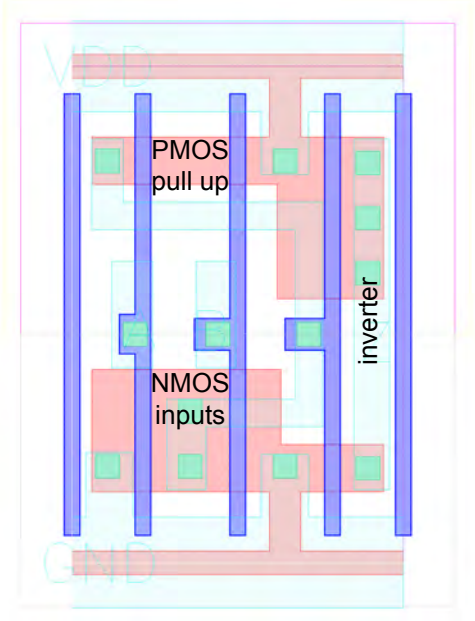


Figure 3.21: Balanced layout for OR gate: the parasitics have been equalized, the NMOSes have been strengthened and the inverter has been sized to favor the rising edge.

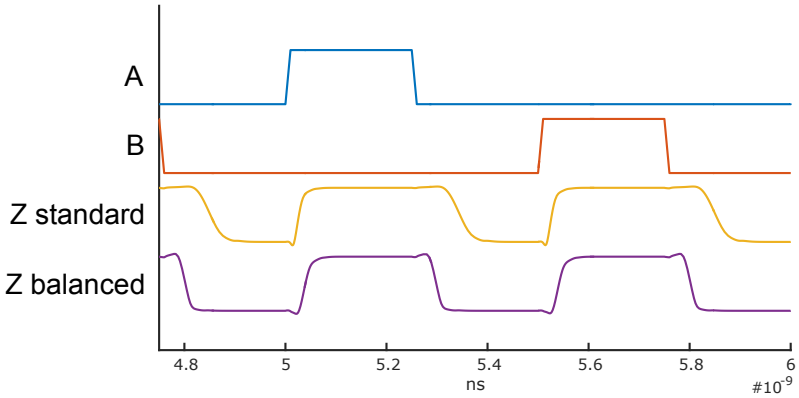
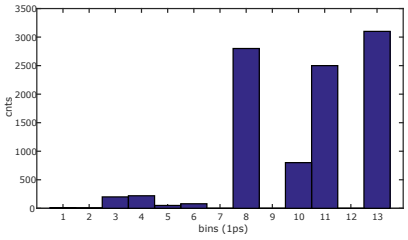
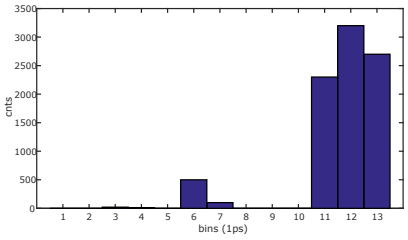


Figure 3.22: Post-layout simulation of standard cell vs balanced cell. The modified cell has a faster response and the difference between its inputs is smaller. As a side effect, the response to the negative edge has a much lower slew rate.



(a) Histogram of the skew of standard OR gate in post-layout simulations. FWHM = 5ps.



(b) Histogram of the skew of balanced OR gate in post-layout simulations. FWHM = 2ps.

Figure 3.23: Post-layout simulation to compare skew between standard OR and balanced OR.

Jitter vs skew: while jitter is the random variability of the response of a module over time when exactly the same signal is applied at its inputs due to non determinable factors, the skew is the repetitive deviation of the response caused by known determinable factors. In general terms, skew can be corrected or accounted for when its causes are known, or after a calibration is used. On the other hand, jitter cannot be corrected since its causes are unclear or they are related to noise or physical processes that might not be even observable. Under this definition, the difference in time response of double-edge gates and unbalanced gates falls into the skew category. For double-edge sensitive gates, if it were possible to know which direction every gate in the XOR tree is switching to, a calibration LUT can be built in order to correct the results given by the TDCs. Seemingly, for unbalanced gates, if it were possible to know which input is asserted for every gate in the XOR tree, another calibration LUT could be built. Unfortunately, all this information is lost in the propagation process; thus, deteriorating the final time jitter. Hence, this at-first-glance skew has to be treated as jitter. The corollary is that any skew whose source is unknown by the system turns into jitter. In applications where the exact position where the photon hit, like in LiDAR, the address of the SPAD is kept and this information can be used for calibration purposes.

Implementation of time lines

In order to define a strategy to design the timing lines, the arrangement of the pixels has to be known. The way the pixels are grouped and connected will define the area the timing line has to cover; the most common way of clustering being: straight clusters ($Nx1$), square clusters (NxN) and rectangular clusters with a preferred direction (NxM with $N >> M$). Another aspect to be considered is the importance of the address (location) of the pixels. In case the system is designed to store the address of the events, this information can be used for later calibration. If this information is lost, the skew that there might be among the pixels that fire the same timing line will increase with N. In Fig. 3.24 and 3.25 are shown several schemes for timing lines that give a visual interpretation of what has been said.

ORed time line: Many image sensors as [10] and [18] implement a distributed OR timing line as the one shown in Fig. 3.26a. This scheme for timing lines is very easy to implement and it does not deteriorate the timing information in relatively small SPAD clusters. The skew can be corrected only in case the sensor keeps the address of the event and the jitter depends on the activity as the line might not be completely recovered by the time a new event occurs. Another important issue is the deterioration of the edge along the line, thus causing closer SPADs to have a sharper edge than afar SPADs, consequently affecting the jitter of the line.

This method was used in one of the designs in this thesis: Concolor, explained in a later chapter. Although the address information is not stored since the main application of the chip is PET imaging, there two factors that conceals the skew: the size of the timing line is short enough and the jitter of the SPADs is 170ps. However, the activity might affect the jitter in some scenarios. High activity swings the timing lines and disturbs the ground and power lines; consequently affecting

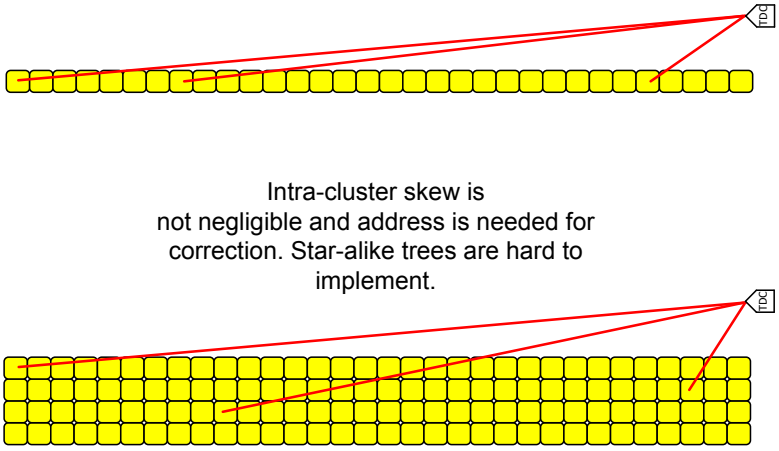


Figure 3.24: Timing-line connection for linear clusters.

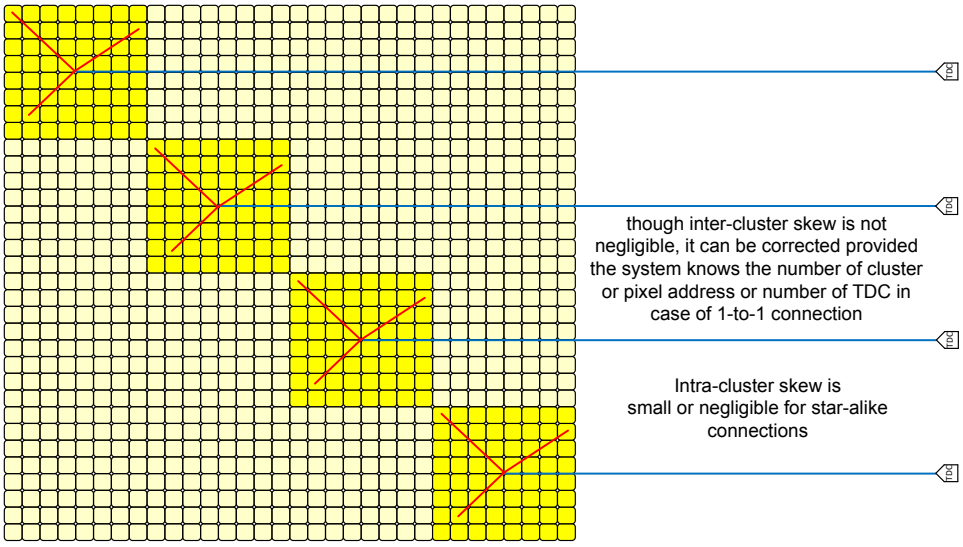
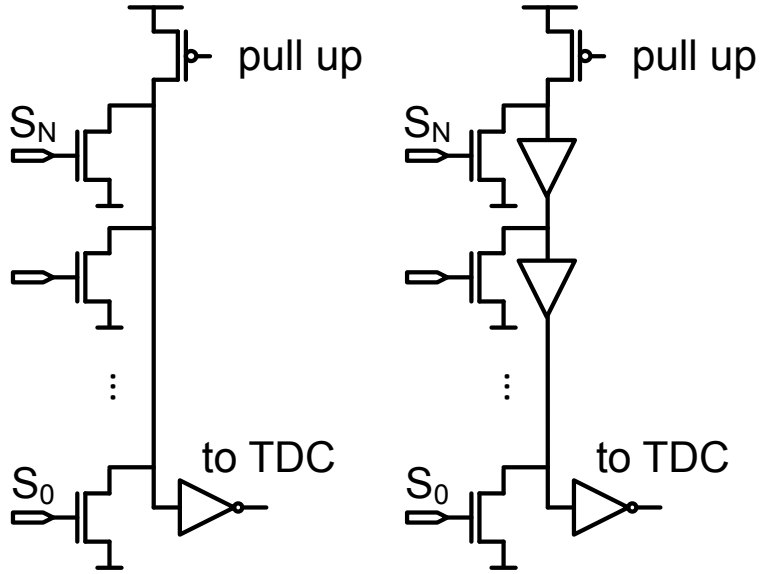


Figure 3.25: Timing-line connection for squared clusters.



(a) ORed timing line achieved by means of open-drain NMOS and PMOS pull-up.

(b) ORed timing line with buffers in between to mitigate the poor edge sharpness.

Figure 3.26: Different cluster schemes for time lines connection.

those signals whose edges are less sharp because a small variation in voltage will be translated into a larger variation in time than for those signals whose edges are sharper.

Buffered ORed time line: The authors of [19] tackled the problem of the fading propagation edges by placing buffers along the line as shown in Fig. 3.26b. This prevents the timing line jitter from worsening. The buffers add an undesirable delay that make the calibration a mandatory process. As discussed, the address information is used to create such calibration.

clusterized OR Square-shaped SPAD clusters enable shorter local connections for timing lines. This means that the parasitic capacitances of the lines are lowered. In this scheme, depicted in fig. 3.25, the intra-cluster skew is negligible while the inter-cluster skew can be fixed only if the address of the pixel, or alternatively the address of the cluster is known.

3.4. Cross-domain signal integrity for analog vs digital domains

The last aspect that will be covered in this chapter is integrity signal especially when cross-domain data transfer is involved. Multiple digital and analog circuits coexist in image sensors. It is well known they must be in different ground schemes well

isolated from each other. Noise and interference immunity of digital circuits is higher than the that of analog circuits. As the time resolution is expected to be on a par with the applications demands ($\propto 10ps$), thus it is required to analyze inter-domain data transfer. The Fig. 3.27 shows a simplified diagram of the analog and digital domains and the interface between them. It assumes off-chip star-alike connections fro VDD and GND. The bonding wires are modelled by RL circuits.

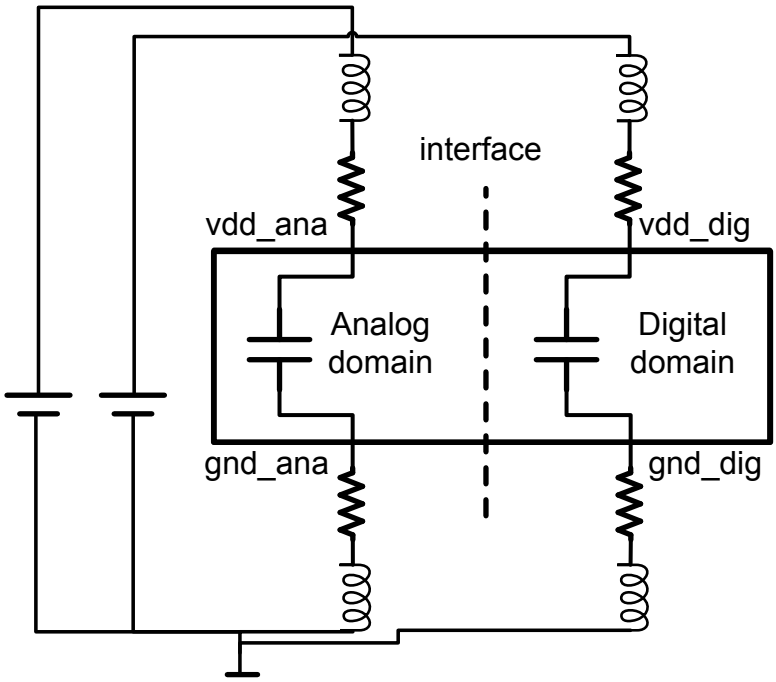


Figure 3.27: General circuit for a chip with multiple power domains. All the connections are considered to be star-shaped to minimize voltage drops. Decoupling capacitors for each domain are also included.

The reader might notice the decoupling capacitors placed in every domain to quickly respond to high frequency current consumption in such a way that those current peaks are not taken from the external power supply, which would lead to a drop voltage on chip. This works well in intra-domain operation. However, decoupling caps do not help when data is transmitted from one voltage domain to the another as it will be explained in the next section.

3.4.1. Inter-domain operation: single ended

In the Fig. 3.28 is shown an inter-domain operation where a buffer from the analog domain is interacting with a buffer from the digital domain, which typically is the case of the outputs of the analog phases of the TDC connected to the inputs of a counter or a register that will store the phase values. The blue currents correspond to the case A is asserted, then the NMOS of the analog domain pulls down the current that goes through the gate capacitance of the PMOS of the digital domain.

This cross-domain current necessarily has to be taken from the power suply of the digital domain and sunk into the ground of the analog domain. Consequently, the current goes through both bonding wires creating two large voltage drops that might not affect the intra domain behavior but it is disastrous for the inter-domain interface. Seemingly, when A is set to 0, the current goes through the other pair of bonding wires, creating the same effect. Fig. 3.29 shows the simulation for typical values of C, L and R for bonding wires, connectors and PADs. In the example, the single-ended signal at the analog domain is generated by a ring oscillator. Analog and digital power and ground voltages fluctuate as the signal changes. Since many decoupling capacitors were placed, the voltage difference between power and ground in each domain remains constant as can be seen in the plot at the bottom (intra-domain voltage ripple). However, the uneven fluctuation of the domains caused by the signal makes the inter-domain voltage difference swing almost 64mV rail-to-rail in this example where only one small buffer (3X minimum transistor) was used. This difference impacts directly into the threshold of the digital buffers with devastating effects on the jitter.

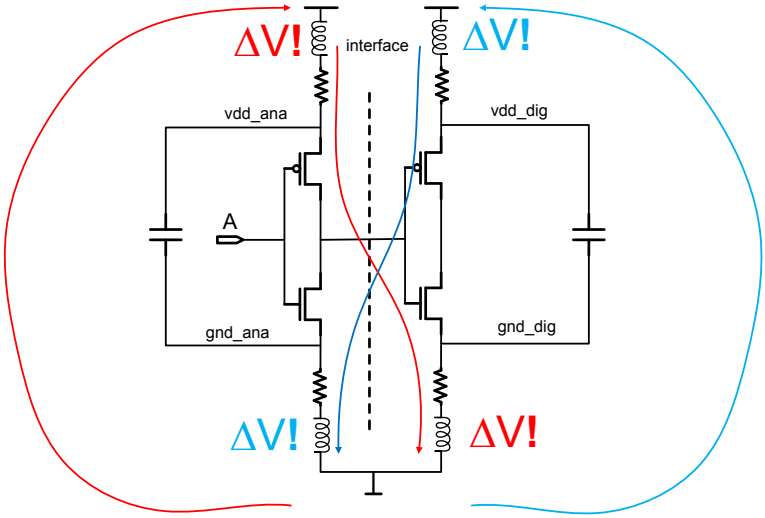


Figure 3.28: Interaction between different voltage domains for single-ended operation.

3.4.2. Inter-domain operation: differential

The main problem observed is that the currents need to ”loop” through the power supplies and the bonding wires. To avoid this situation, it is very important to use differential buffers along with domain isolation so as to create equal and opposite currents. Fig. 3.33 shows the currents in case of operating differential buffers. Black-colored buffers are the same as the previous case, while green buffers are buffers that work in counter-phase. When A is asserted, as in the previous example,

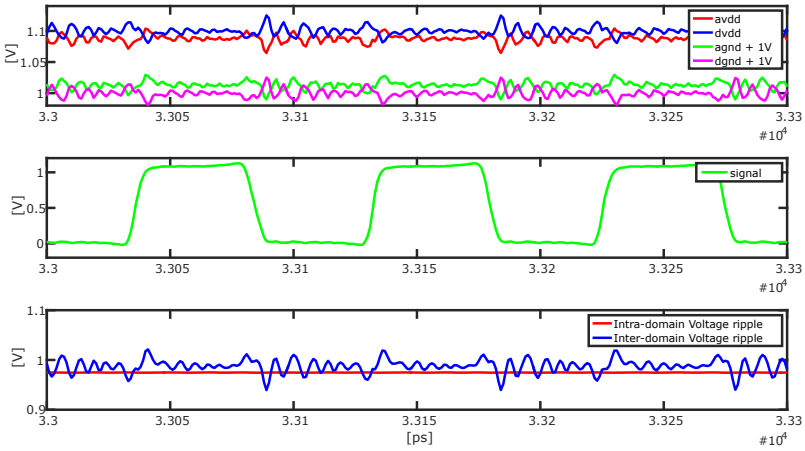


Figure 3.29: Post-layout simulation of cross-domain interaction in single-ended one-substrate case. Analog power (avdd) and ground (agnd) at the top along with digital power (dvdd) and digital ground (dgnd). The signal that crosses from analog domain to digital domain (middle). Inter and intra domain ripples (bottom).

the black NMOS of the analog domain pulls the current that goes through the gate of the PMOS of the digital domain. However, the green PMOS of the analog domain is pushing the current through the gate of the gate of the green NMOS of the digital domain. If the transistors are sized properly to be comparable in speed and amount of current, these currents will be very similar and will be taken from the decoupling capacitors that were meant for intra-domain operation at first. This is crucial since the difference of the pulling and pushing currents will be taken from the external power supply with the effects that have been discussed. Fig. 3.30 shows the simulation for the current compensation. The signal is generated by a ring oscillator of 3 phases. The positive and negative phases are transmitted to the digital domain by means of isolated-substrate differential buffers shown in Fig. 3.31. The current pulled by the buffers are compensated. The result is that the inter-domain ripple is much lower than that one in the single-ended single-substrate version. The rail-to-rail swing in this case is $13mV$ for 3 phases where each buffer is 2.3 times bigger than the buffers in single-ended case. Thus, totalizing an improvement of $\frac{S_{single}W_{single}}{S_{diff}W_{diff}} = \frac{64mV*2.16\mu m}{1/3*13mV*0.930\mu m} \approx 34.30$ times. In both configurations, the intra-domain voltage remains invariable. The buffer has two stages supplied by the voltages and grounds of the domains 1 and 2 respectively. The currents must be perfectly compensated to diminish as much as possible the undesired effects explained in this section. Hence two properly-sized inverters are placed back to back in order to compensate for any process variation and to match the currents. The layout of the buffer is shown in Fig. 3.32. The distance between the stages were made following the recommendations of the technology. Guard rings isolate

the two parts of the substrate.

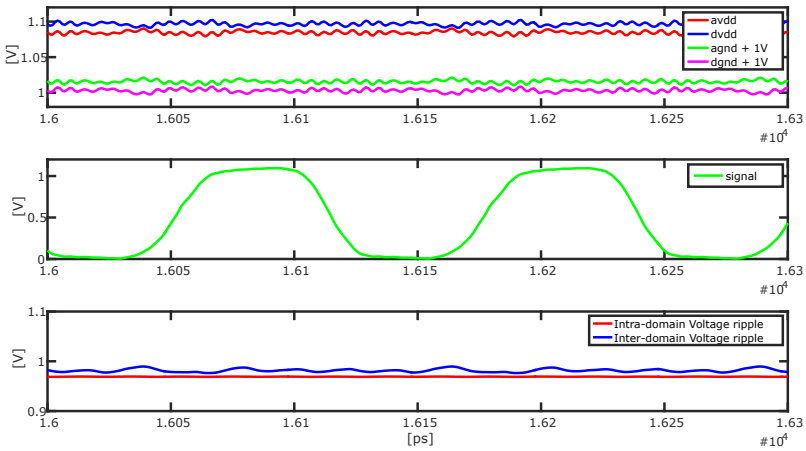


Figure 3.30: Post-layout simulation of cross-domain interaction in differential isolated-substrate case. Analog power (avdd) and ground (agnd) at the top along with digital power (dvdd) and digital ground (dgnd). The signal that crosses from analog domain to digital domain (middle). Inter and intra domain ripples (bottom).

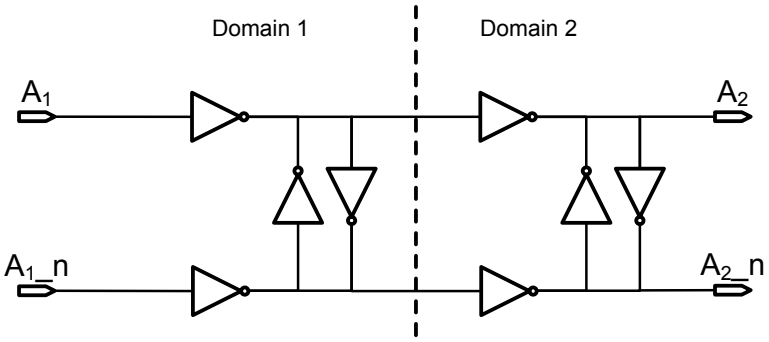


Figure 3.31: Inter-domain isolated-substrate pseudo-differential buffer.

3.4.3. Layout considerations

As explained, having different domains in image sensors and differential communication between domains is crucial for proper operation for analog and digital modules; this reduces drop voltages and power supply bounces. As will be discussed in further chapters, TDCs are very sensitive to any variation of voltage. Any current peak from the digital domain would alter the performance of TDCs. Hence, the domains need to be well separated on the substrate in order to increase the resistance in between, and they must be surrounded by guard rings to enclose noise

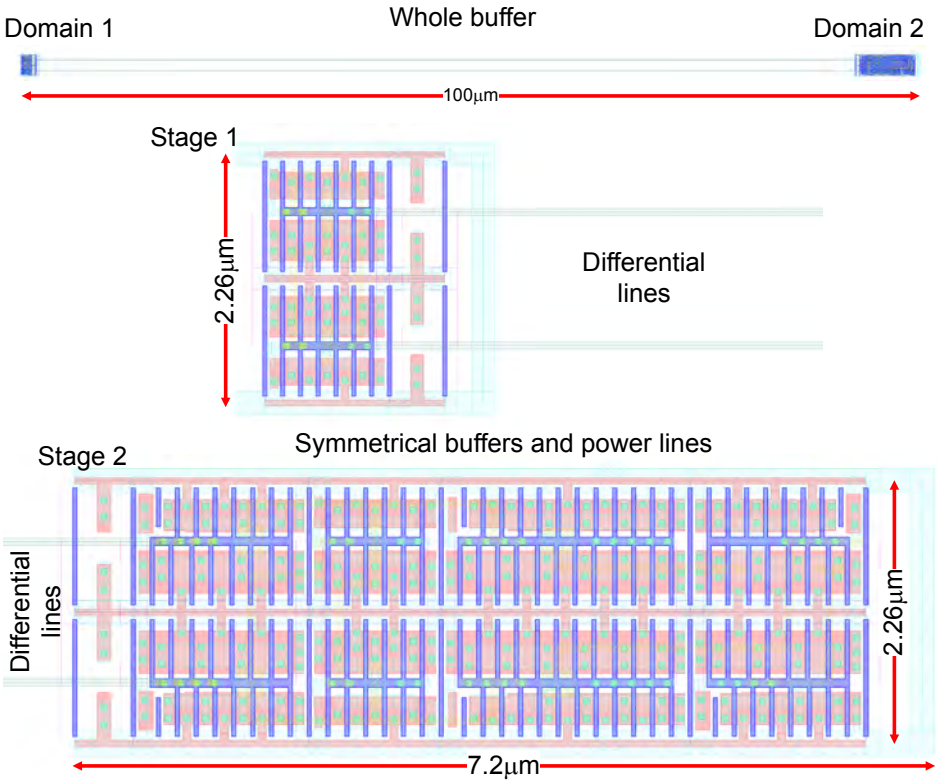


Figure 3.32: Inter-domain isolated-substrate differential buffer.

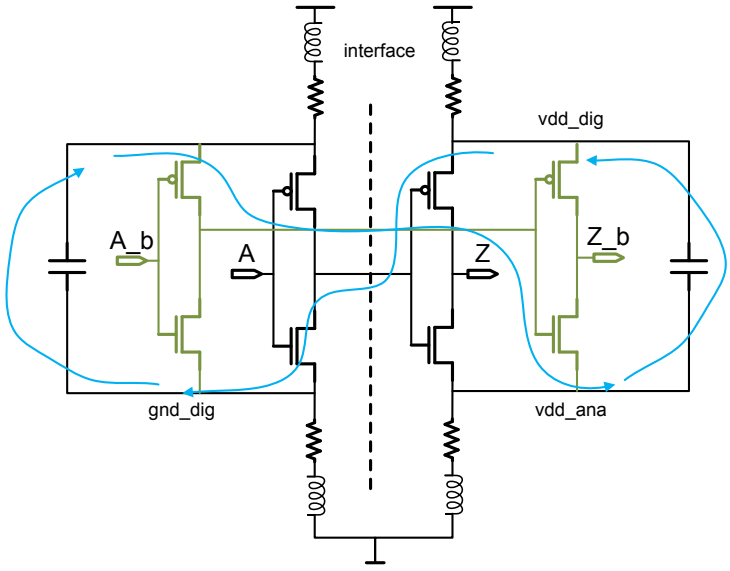


Figure 3.33: Interaction between different voltage domains for differential-ended operation which along with decoupling capacitors make current compensation.

currents. However, very long distance in the interface might lead to poor edges in the signal that might make it more susceptible to noise creating additional jitter. The resistance of the substrate depends on the given technology and it should be considered to estimate the right distance of the interface. In some technologies, it is possible to use intrinsic silicon to create high resistive rings, highly recommended for high frequency circuits.

3.5. Conclusion

In this chapter, we explained the basis for intensity counting and time signals for image sensors. At first, in the introduction, image sensors were explained in a general and conceptual way, introducing all the main modules and their functionalities. In this chapter, the specific main challenges for intensity counting and time signals were discussed and thoroughly explained; followed by specific solutions and implementations that have been included in the chips designed and presented in this work. The solutions are focused to achieving the maximum counting capabilities and the best time resolution, while taking care of the parameters that affect image sensors the most; which are area, shape factor and power consumption.

On the intensity: for this case, granularity and proportion of active-inactive (E) area are the key factors of the design. It would be desirable for a given sensor to have the maximum granularity while maintaining the active area high which in 2D technologies one goes in detriment of the other. The total area taken by the intensity counting circuit in a sensor of N SPADs and granularity G_C (number of SPADs shared by the B-bits counters) can be expressed as follows:

$$A_{counting} = GBK_1A_{flip-flop} + \frac{N}{G_C}K_2A_{gate},$$

where K_1 is the overhead area ratio for the counters, K_2 is the overhead area of the gates used in the intensity counting tree for G counters, $A_{flip-flop}$ is the area of a single flip-flop, and A_{gate} is the area of a single OR or XOR gate.

On the time resolution: like in the case of intensity, it would be desirable to get as many samplers as possible to catch all the time-of-arrival of the impinging photons. Whether a FIFO memory or a sampling register is implemented, the more sampling circuits there are, the more area will be taken by non-active circuits. The following formula helps as a guide to analyze architectures:

$$A_{timing} = TA_{TDC} + N/G_TKA_{gate},$$

where K is the overhead area ratio of the gates used in the timing tree for TDCs, G_T is the TDC granularity (number of SPAD that share a single TDC), A_{TDC} is the area of a single TDC, and A_{gate} is the area of a single OR or XOR gate.

Numerical results: The aforementioned two formulas of area for intensity counting and timing are combined to calculate the active-inactive area ratio of a sensor, the values of the physical components were considered for 40nm technology.

$$E = N * \pi r^2/(N * \pi r^2 + A_{counting} + A_{timing}),$$

where r is the radius of the SPAD.

This table shows how the granularity of TDCs and intensity counters greatly affects the available sensing area. The application will determine which E should be aimed.

G_T / G_C	1	8	16	64	128	256
1	0.11	0.11	0.11	0.11	0.11	0.11
8	0.49	0.50	0.50	0.50	0.50	0.50
16	0.65	0.66	0.66	0.66	0.66	0.66
64	0.86	0.88	0.88	0.88	0.88	0.88
128	0.91	0.93	0.93	0.94	0.94	0.94
256	0.93	0.96	0.96	0.96	0.96	0.96

Table 3.2: Ratio between active area vs. inactive area for different values of G_T and G_C and $r = 20\mu m$.

References

[1] D. R. Anita and M. Jayasanthi, Review of clock distribution networks, in *2015 Online International Conference on Green Engineering and Technologies (IC-GET)* (2015) pp. 1–4.

[2] T. Figliolia and A. G. Andreou, The Conical-Fishbone Clock Tree: A Clock-Distribution Network for a Heterogeneous Chip Multiprocessor AI Chiplet, in *2019 22nd Euromicro Conference on Digital System Design (DSD)* (2019) pp. 160–165.

[3] D. Joo and T. Kim, Managing clock skews in clock trees with local clock skew requirements using adjustable delay buffers, in *2015 International SoC Design Conference (ISOCC)* (2015) pp. 137–138.

[4] S. E. Esmaeili and A. J. Al-Kahlili, Integrated Power and Clock Distribution Network, IEEE Transactions on Very Large Scale Integration (VLSI) Systems **21**, 1941 (2013).

[5] Chia-Ming Chang, Shih-Hsu Huang, Yuan-Kai Ho, Jia-Zong Lin, Hsin-Po Wang, and Yu-Sheng Lu, Type-matching clock tree for zero skew clock gating, in *2008 45th ACM/IEEE Design Automation Conference* (2008) pp. 714–719.

[6] C. Lin, S. Huang, and W. Cheng, An Effective Approach for Building Low-Power General Activity-Driven Clock Trees, in *2018 International SoC Design Conference (ISOCC)* (2018) pp. 13–14.

[7] W. Zhang, Y. Hu, K. Cui, D. Bao, D. Pan, L. Wang, and L. Zheng, Standing wave oscillator based clock distribution, in *2016 International SoC Design Conference (ISOCC)* (2016) pp. 301–302.

[8] Zhe Ge, Juan Fu, Peidong Wang, and Lei Wang, Improve clock tree efficiency for low power clock tree design, in *2016 13th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)* (2016) pp. 840–842.

[9] Y. Tomita, K. Suzuki, T. Matsumoto, T. Yamamoto, H. Yamaguchi, and H. Tamura, An 8-to-16GHz 28nm CMOS clock distribution circuit based on

mutual-injection-locked ring oscillators, in *2013 Symposium on VLSI Circuits* (2013) pp. C238–C239.

[10] A. Carimatto, S. Mandai, E. Venialgo, T. Gong, G. Borghi, D. R. Schaart, and E. Charbon, 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.

[11] L. H. C. Braga, L. Gasparini, L. Grant, R. K. Henderson, N. Massari, M. Perenzoni, D. Stoppa, and R. Walker, An 8×16-pixel 92kSPAD time-resolved sensor with on-pixel 64ps 12b TDC and 100MS/s real-time energy histogramming in 0.13µm CIS technology for PET/MRI applications, in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers* (2013) pp. 486–487.

[12] S. Gnecchi, N. A. W. Dutton, L. Parmesan, B. R. Rae, S. Pellegrini, S. J. McLeod, L. A. Grant, and R. K. Henderson, Digital Silicon Photomultipliers With OR/XOR Pulse Combining Techniques, IEEE Transactions on Electron Devices **63**, 1105 (2016).

[13] F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, and I. Rech, 83-ps Timing Jitter With a Red-Enhanced SPAD and a Fully Integrated Front End Circuit, IEEE Photonics Technology Letters **30**, 1727 (2018).

[14] F. Nolet, S. Parent, N. Roy, M.-O. Mercier, S. A. Charlebois, R. Fontaine, and J.-F. Pratte, Quenching Circuit and SPAD Integrated in CMOS 65 nm with 7.8 ps FWHM Single Photon Timing Resolution, Instruments **2** (2018), 10.3390/instruments2040019.

[15] H. Park, Z. Yu, J. Kim, and J. Burm, Resolution tunable ring oscillator type TDC, in *2016 International SoC Design Conference (ISOCC)* (2016) pp. 241–242.

[16] R. K. Henderson, N. Johnston, S. W. Hutchings, I. Gyongy, T. A. Abbas, N. Dutton, M. Tyler, S. Chan, and J. Leach, 5.7 A 256×256 40nm/90nm CMOS 3D-Stacked 120dB Dynamic-Range Reconfigurable Time-Resolved SPAD Imager, in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)* (2019) pp. 106–108.

[17] T. Nikoubin, F. Eslami, A. Baniasadi, and K. Navi, A new cell design methodology for balanced XOR-XNOR circuits for hybrid-CMOS logic, J. Low Power Electronics **5**, 474 (2009).

[18] S. Mandai, V. Jain, and E. Charbon, A $780 \times 800\ \mu\text{m}^2$ Multichannel Digital Silicon Photomultiplier With Column-Parallel Time-to-Digital Converter and Basic Characterization, IEEE Transactions on Nuclear Science **61**, 44 (2014).

[19] 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 **54**, 1137 (2019).

4

Concolor: Multi-purpose design, focused on Positron Emission Tomography in 40nm ST technology

"Ampere was the Newton of Electricity;"
"Faraday is, and must always remain, the father of that enlarged science of electromagnetism."

James Maxwell

To become good at anything you have to know how to apply basic principles. To become great at it, you have to know when to violate those principles.

Gary Kasparov

A concrete implementation of an image sensor for Positron Emission Tomography presented in this chapter is the first-ever MD-SiPM fabricated in 40-nm technology. Synchronous and asynchronous applications based on the chip were proven to work. Details of the architecture and design are shown in the chapter, along with measurements. Part of the content of this chapter has been published in [1].

This chapter is dedicated to a concrete implementation of a system that was fabricated in ST 40-nm technology [2] mainly focused on Positron Emission Tomography. Other photon applications, such as LiDAR have been also considered for the design. The reader will be walked through the important aspects of the design and the thinking process. The structure, circuits and modes of operation are itemized and analyzed thoroughly. Then, results for both the applications are presented and discussed.

4.1. Sensor architecture

A multipurpose monolithic array of 2 × 2 multichannel digital silicon photomultipliers (MD-SiPMs) fabricated in 40-nm CMOS technology has been designed in this work [1]. Each MD-SiPM comprises 64 × 64 smart pixels connected to 128 low-power 45-ps sliding-scale time-to-digital converters (TDCs). The sliding-scale technique enables the TDCs to use a random segment of their scales for each measurement to mitigate the DNL caused by irregularities in the layout. The system can operate in two different modes: 1) event-driven and 2) frame-based. The first is suited for positron emission tomography (PET), where the events are uncorrelated with the system clock, and the second for synchronous application like LiDAR, where the events are synchronous with the clock of the system. The design includes electronics to indirectly capture gamma events by means of a scintillator optically coupled to it. The digital readout is fully embedded in the sensor and it is reconfigurable by SPI. Data packets are sent following a simple protocol that makes it compatible with an external FIFO, therefore making use of an FPGA optional. Every MD-SiPM can deliver up to 64M time-stamps/s. The sensor can be arranged in any type of configuration through a dedicated synchronization input and can be used to operate jointly with an event generator, such as a pulsed laser, which is useful in many applications. Inherently compatible with STMicroelectronics 3-D-stacking technology, the sensor can serve as front-end electronics when it is used with a different SPAD silicon tier.

4.2. Decription of the system

The system is composed by four MD-SiPMs [3] organized in independent quadrants as depicted in Fig. 4.1. Every MD-SiPM has three main components: 1) a 64 × 64 dual SPAD pixel array; 2) a bank of 128 TDCs; and 3) a digital core that exchanges commands and data from and to an external system. Four global skew-free signals (shutter, clock, main reset and self-reset) are routed along the whole sensor. The SPAD array is divided into 32 panels of 4 × 128 dual SPADs along with the electronics to operate it. Columns in the panel are split in two semicolumns whose SPADs share one TDC to register the arrival time of the photons, as shown in Fig. 4.3.

The electronics that serves the SPAD has five subcircuits plotted in Fig.4.4 and explained below, along with circuitry for global signals.

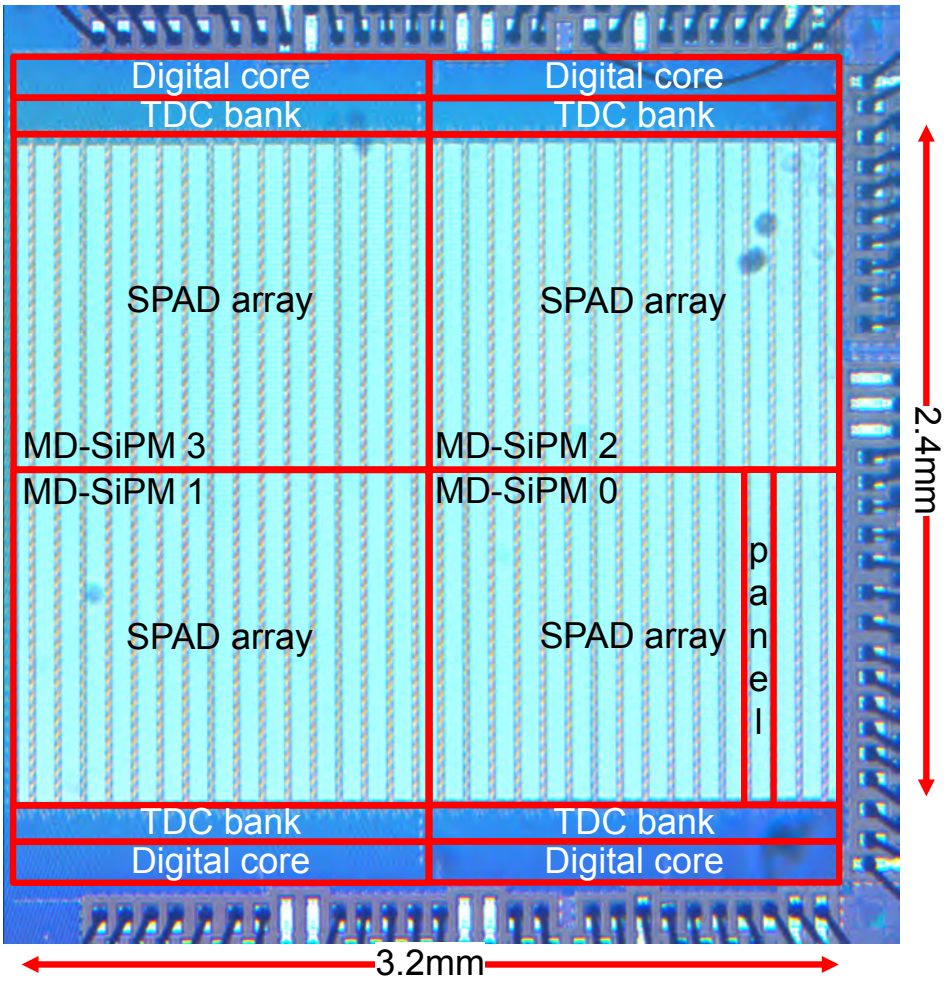


Figure 4.1: Micrograph of the sensor. TDC bank, digital core, SPAD array and one of its panels are shown for MD-SiPM 0. The 3 remaining quadrants are a replica of the same module laid out in central symmetry.

4.2.1. Pixel electronics

Quenching and reset: the first component represents passive quenching and active recharge circuits, implemented with the transistors Q_1 and Q_2. The output resistance of the transistor Q_1 can be set by the voltage V_Q. The active reset transistor Q_2 can be activated either by the self-reset circuit or by the global reset. Both Q_1 and Q_2 are thick oxide transistors to stand excess bias voltages beyond $1.1V$ upto 3.3V.

Memory element: a 1-bit memory, implemented with a flip flop, stores the SPAD state (fired or not fired) that can be read out at the end of the measurement process. This memory element also is reused for other tasks as well, so as to reduce the area requirements for the pixel circuitry. The in-pixel memory is daisy-chained to the previous and next in-pixel memory and to the digital core at both ends, then forming a sort of circular register. The purpose is to enable its use in the read-out and configuration processes. Both SPAD firing information and masking patterns go through this circular register that is present in every electrical column.

Self-reset module: when the system is operating in event-driven mode, the time of exposure begins with a synchronization signal and runs indefinitely until an event occurs. During this time, the SPADs get fired due to DCR, thus reducing the pixel availability of the sensor. In order to recover those fired SPADs, the self-reset module gets a command from a digital core (explained in the ”Operation” section) and decides based on its own state whether the pixel should be reset or not. This mechanism improves DCR rejection and intensity resolution of the system. It is an improvement of [3].

Masking memory: every pixel circuitry has two 1-bit static memory to store the masking information for the SPADs. This memory has the ability to disable the SPAD in case its DCR is too high for the given application. The output of this memory acts directly on the reset circuit. A high-level value does not have any effect, and a low-level value shuts down the reset circuit, thus preventing the SPAD from getting reset. As a consequence, the SPAD does not get recharged once fired, and can never trigger an avalanche again. This is called ”optical masking”. The same signal is applied to the electronics so the SPAD output is disabled. This is known as ”electrical masking”. Both the masking memories are connected to the flip-flop of the pixel circuit for reutilization. Two signals, ”write1” and ”write2”, performs the write of the value of the flip flop into the corresponding masking memory.

Timing transistor: Q_3 is a transistor used to register the time of arrival of events through the timing line. The timing line remains in the high state thanks to a pull-up transistor to Vdd until any SPAD in the semicolumn pulls it down after an event occurs. The circuit is presented in section 4.5.

Global signals: some of the signals of Concolor have to synchronously reach every pixel to meet the time constraints of the digital circuits and to ensure timing

accuracy in the gating process; therefore, these signals have been promoted to global signals. The digital clock that makes the information travel through the flip-flop needs to meet hold and set-up times and transition time required by the technology. The signal SPAD-enable is the global shutter of the system that starts the measurements in the sensor. Global shutter was preferred over rolling shutter to enable gating operation of the sensor. It successfully worked with a gate of 20 ns. An H-tree scheme was chosen for these signals due to known performance and jitter of this type of clock distribution technique. A diagram of buffers and connections is shown in Fig. 4.2.

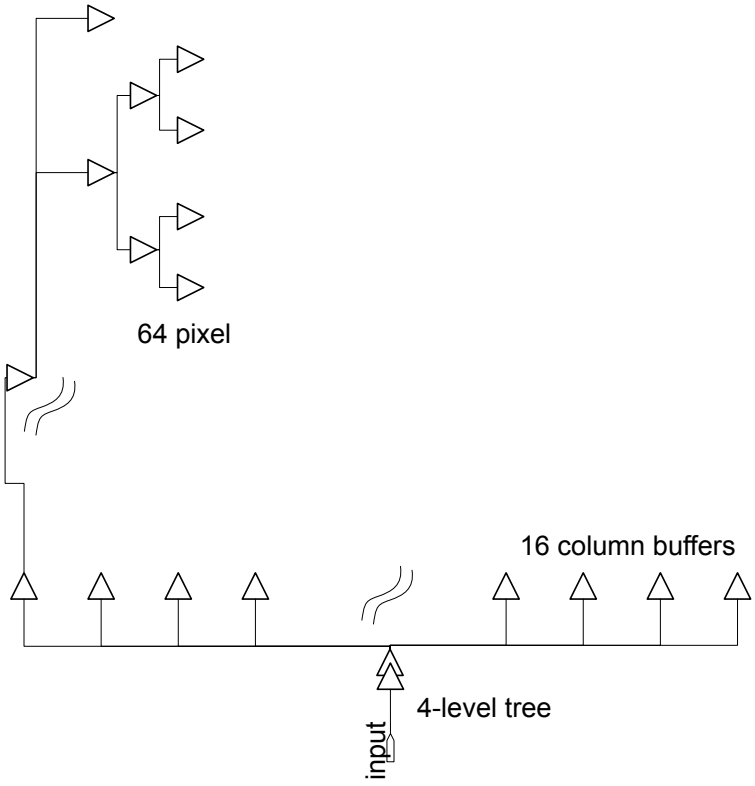


Figure 4.2: Simplified diagram of the H-tree clock implemented for every global signal present in Concolor.

4.2.2. Modes of operation

Two modes of operation are possible: frame-based and event-driven. In frame-based mode, the global shutter signal is opened for a fixed time and the detector captures the events whose information is available after the shutter is closed. This mode is preferred when events are synchronous with the system clock, such as time-of-flight cameras. In event-driven mode, the shutter remains open until an event occurs. This event is defined by the ratio of photons per unit of time, which

is an externally configurable parameter. If this condition is not achieved, the self-reset module resets the pixels and TDCs that have fired, thus rejecting DCR and background noise. The level of the latter is defined for the given application; when such levels are reached, an event is detected, the digital core closes the shutter after the predefined integration time [4] the information becomes immediately available. Event-driven operation reduces dead times, throughput, power, and is particularly effective when the events are uncorrelated with the system clock and the shutter (e.g., PET).

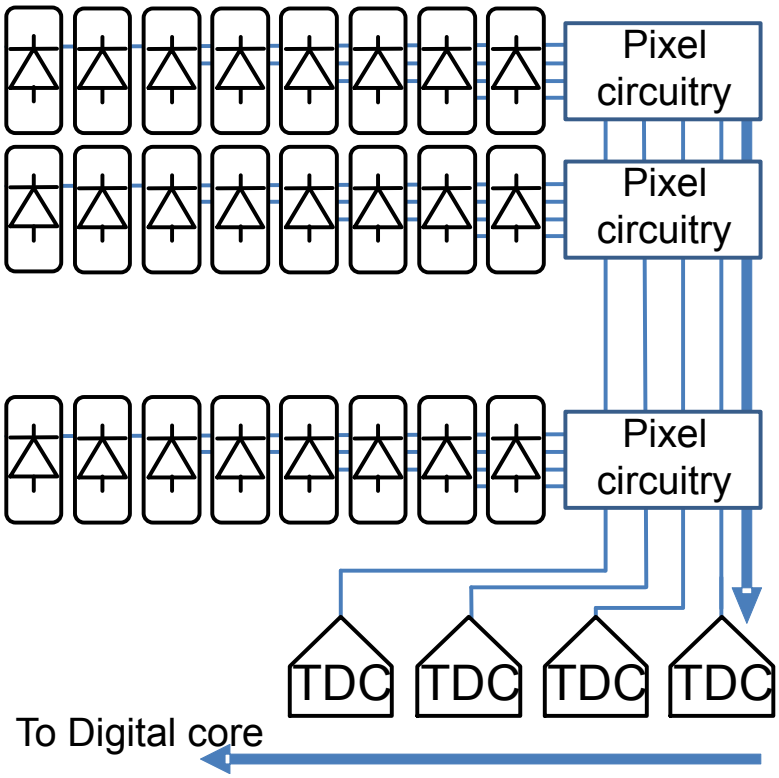


Figure 4.3: Simplified diagram of the connections between the SPADs and the TDCs that represents a panel.

4.2.3. General description of the digital core

The digital core performs the control and the readout of the MD-SiPM, which includes: masking, SPAD and TDC reset, configuration for window frame, frame mode, readout of pixels and TDCs, and synchronization operation. The core accepts the commands by serial communication and the readout is a 16-bit 2.5-V CMOS clocked bus that can be connected through a CMOS USB FIFO directly to the PC. A dedicated synchronization input is provided to work with several modules simultaneously. The maximum event rate depends on the mode of operation, the

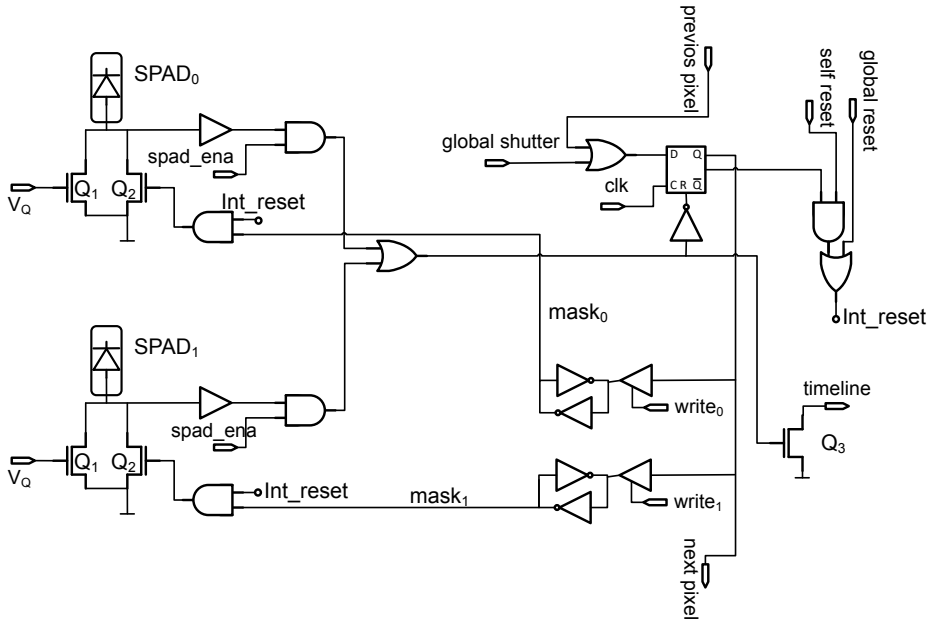


Figure 4.4: Pixel circuit: a Flip flop is chained with the previous and next pixels in scan-mode fashion to configure the masking memory and read out the data.

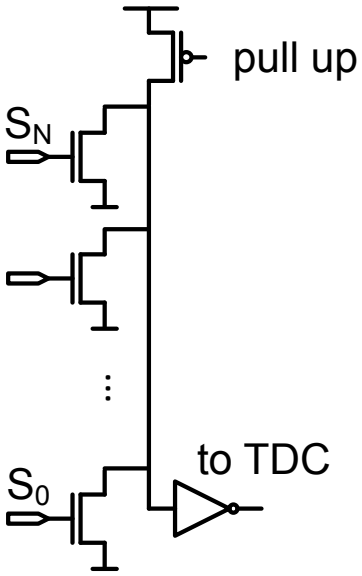


Figure 4.5: Simplified connection of the SPAD column and TDCs pairs.

length of the frame, and the activity for the given application. However, it finds its upper limit in the transmission bandwidth of the system. For frame-based mode, the package to be transmitted comprised of TDCs (2560 bits) and pixels (4096 bits). At a frequency of 80 MHz, it takes about 5.2 µs for the bus to transmit these data packages, thus reaching 24.6M time-stamps/s and 192K frames/s. For the event-driven mode, TDCs (2560 bits), pixel addition (32 bits), and global time (32 bits) are sent. It takes about 2µs to transmit the data packages, thus reaching 64M time-stamps/s.

4.3. SPAD array

The FSI SPADs of Concolor were fabricated in STMicrolelectronics 40nm technology [2]. In this section, the layout and the measurements on the SPAD array are shown in detail.

4.3.1. Layout

The floorplan of the SPAD array can be seen in Fig. 4.6. The whole array of Concolor comprises 32768 SPADs. This array is split into panels of 8x64 SPADs that can be fed by different high-voltage supplies. Each MD-SiPM has 8 panels, for a total of 8192 SPADs. Every panel has 4 pixel columns and 1 electronics column (e-columns) whose sizes are equal in order to enable 3D integration. The SPADs work in duets to form one pixel that although they share 1 electronics module, they have their own masking, quenching and reset circuitry.

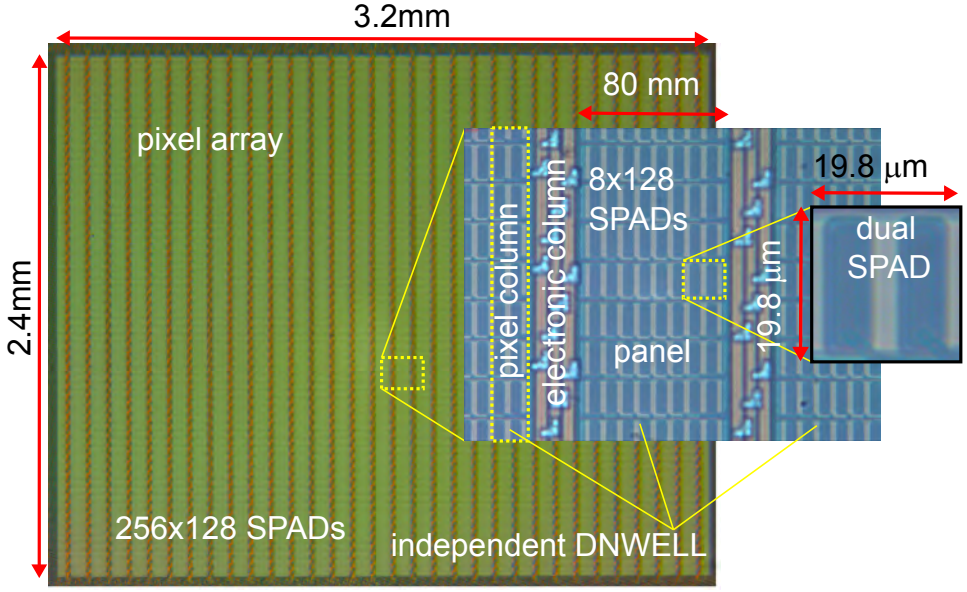
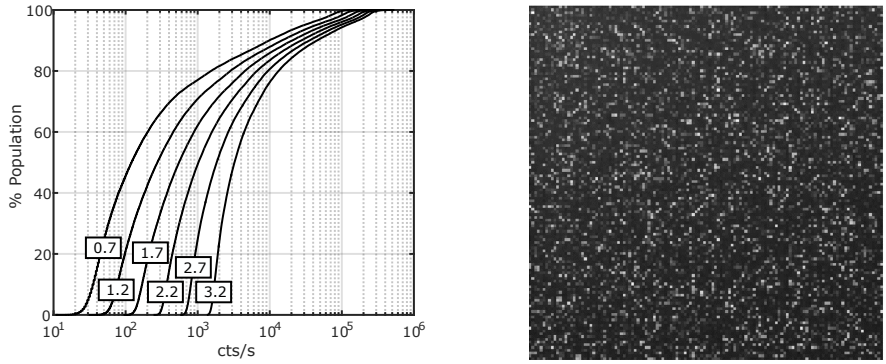


Figure 4.6: SPAD array where panel, columns, e-columns and SPADs are displayed.



(a) DCR of the whole population for various excess bias (b) DCR 2D map of the DCR @$V_{EB} = 1V$ voltages.

Figure 4.7: DCR characterization for Concolor.

0.20%	-0.03%	-0.01%	0.02%	0.00%	0.15%	-0.11%	-0.19%
0.08%	0.60%	0.14%	1.06%	0.56%	0.23%	0.30%	0.04%
0.06%	5.48%	—	0.54%	13.67%	2.26%	0.44%	0.09%
0.14%	1.04%	0.17%	2.25%	1.92%	0.78%	0.36%	0.21%
0.24%	0.29%	0.23%	0.22%	0.16%	-0.02%	0.06%	-0.12%

Table 4.1: Cross-talk between SPADs in the same panel.

4.3.2. Characteristics

Dark count rate: the SPADs were measured in darkness for different excess bias voltages. Results are shown in 4.7a. The knee of DCR is approximately located at 70%. The number of screamers for PET application was found to be between 10% and 15%.

Geometrical uniformity: non-uniformities in the layout, differences in the SPADs and physical variations in the process lead to a general non-uniformity along the array. The achieved uniformity when the sensor is illuminated by flood white light, measured as FWHM/mean, equals 3.8%. Histogram is shown in Fig. 4.8.

Cross-talk: the cross-talk among SPADs in a section of panel was calculated using the method aggressor-victim used in [5]. The section is shown in Fig. 4.9. The selected aggressor for the measurement was the SPAD $(x_0; y_0) = (3, 4)$ and the results are shown in Table. 4.1. The SPAD $(2, 4)$ shares its output with the aggressor so it cannot be considered for this measurement. It can be observed that the horizontal cross-talk is higher than the vertical cross-talk, explained by the shape of the SPADs.

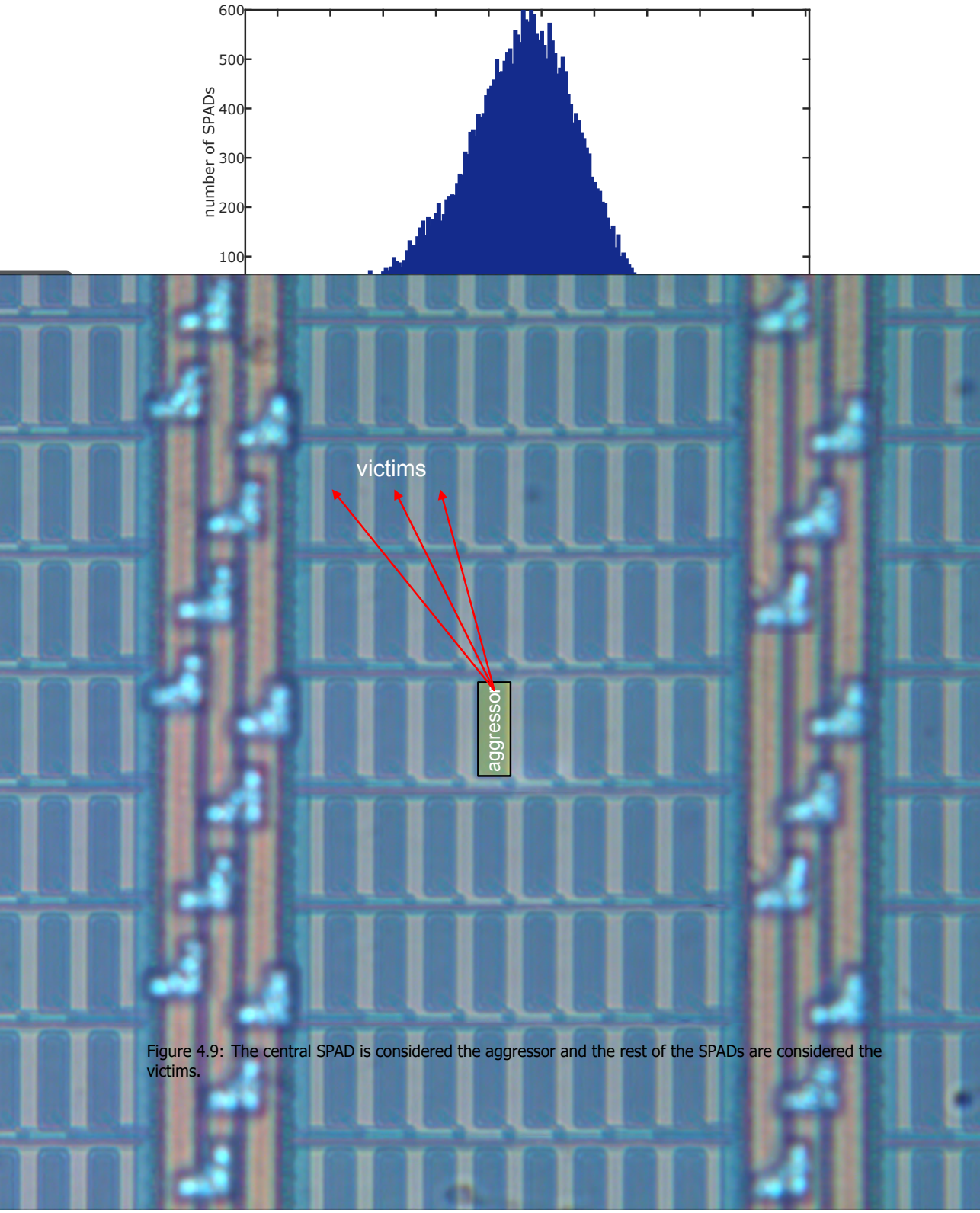


Figure 4.9: The central SPAD is considered the aggressor and the rest of the SPADs are considered the victims.

4.4. Digital core

The inclusion of a programmable digital module to completely configure the sensor, operate it and read out the information made full integration of Concolor possible. The block diagram of the digital core is shown in Fig. 4.10.

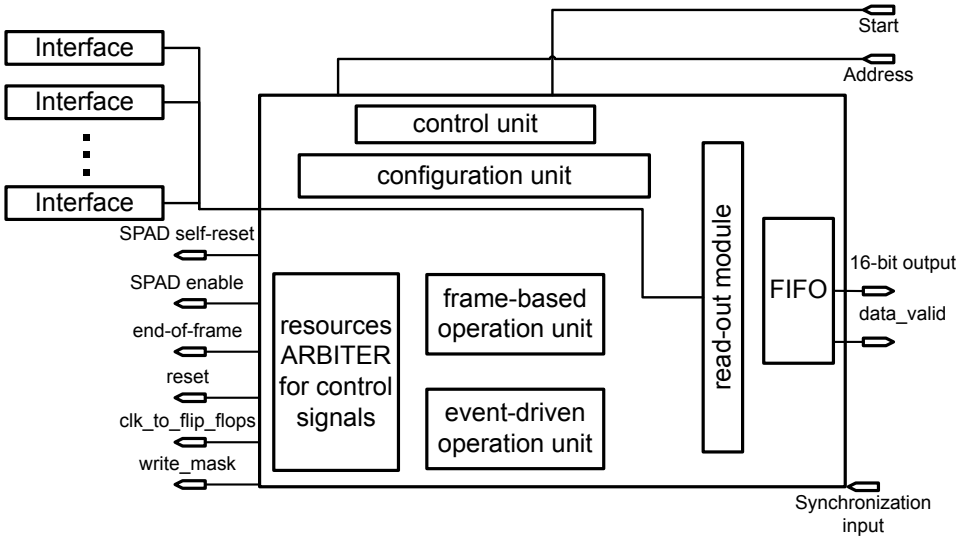


Figure 4.10: Diagram of digital core, written in VHDL. Critical timing components were custom-made.

The digital core has all the modules that are required to operate Concolor. The control unit is the interface with external world that can be either an FPGA or an SPI controller from a computer. It receives commands, processes them and passes the information to the rest of the modules. The masking unit takes the masking information from the control unit and operates all the signals to store it to the internal in-pixel memory of the array.

To operate the chip for measurements, there are two modules: frame-based operation unit and event-driven operation unit. Those units, as explained before, are meant for the two main types of applications: synchronous and asynchronous. Both the modules, along with the masking unit need access to all the internal control signals of Concolor that go to the pixel array and TDCs, such as self-reset, SPAD enable, end-of-frame, reset, etc.. These hard resources, signals and modules cannot be shared for operation. This is resolved by the arbiter of the chip that takes the commands from the control unit and grants access to a particular module that needs to operate the sensor. The digital core was programmed in VHDL, tested with ModelSim, synthesized with Vision Design and placed and routed with Encounter. Post-layout simulations were carried out in both ModelSim using .sdf files and in Analog Design Environment (ADEL). The simulation makes use of Spectre for analog components and veriloga for functional and non-synthesizable modules. For the full digital simulation, all the main components of Concolor as TDCs, pixel array and electrical columns were described by their behavioral coun-

teparts. This method speeds up the verification process and facilitates debugging stage in a faster environment.

4.4.1. Configuration

Configuration of Concolor begins with reset of the digital core. The reset sets the state of all the internal components to idle. As next step, a ECHO command should be sent to check that communication and digital core are up and running. For this, Concolor receives the ECHO command with a value that sends back to the FPGA, along with a hard-coded register can be checked on the computer. If this is done successfully, the digital is ready for operation.

Figure 4.11: An arbitrary masking pattern was tested on chip.

At this point, the in-pixel memory that stores the masking pattern has random power-on values. The mask command should be sent to the control unit to complete the second step. Then the masking unit takes control of the signals through the arbiter and starts filling the memory with the masking pattern row by row, pixel by pixel. The masking can absolutely arbitrary and an example of how this unit works is shown in Fig. 4.11.

The digital core counts on several commands that are used in the masking process. The pseudo code is as follows:

function StoreMask(bool $DATAMASK[][]$)
  for $side$ in [0,1] **do**
    for $row_i = 0 \rightarrow 63$ **do**
      for $col_i = 0 \rightarrow 63$ **do**
        push $DATAMASK[col_i][row_i] \rightarrow REG_{mask}$
      end for
      push $REG_{mask} \rightarrow in - pixel - memory$
    end for

```
        set write(side) HIGH
        wait 1μs
        set write(side) LOW
        wait 1μs
    end for
end function
```

At the end of this procedure, the register is filled up with a logic 1 as that is the reset value of the flip-flops. Therefore, when the array memory is read out, the in-pixel flip-flops get automatically reset as in a FIFO memory due to destructive read operations. Pseudo code is as follows:

```
function PrepareFIFO(void)
    for col_i = 0 → 63 do
        push 1 → REG_mask
    end for
end function
```

In a real measurement, the masking pattern should be calculated based on a measurement of the sensor in complete darkness. The DCR map is measured for every MD-SiPM and the SPADs are sorted from the least noisy to the most noisy. The application states how to define screamers and what DCR level is acceptable. The SPADs that do not comply with this limit are excluded by setting their masking value to 0.

As third step, it is necessary to chose one of the two available modes of operation and to set the parameters that add flexibility to each mode. For frame-based mode, it is required to define the length of the frame and the strength of the reset. For event-driven mode, along with the strength of the reset, it is necessary to set the double-threshold of the derivative of light/DCR to define the presence of an event. Once communication has been tested, the masking information has been stored and the mode was set, Concolor awaits measurements commands.

4.4.2. Operation

Frame-based mode for synchronous applications (e.g. LiDAR)

For this mode of operation, the time of exposure is known. The sensor becomes sensitive to light for a known length of time upon the synchronization signal has been asserted. After this time has elapsed, the frame-based operation unit shuts off the pixel array and asserts end-of-frame signal that latches the reference TDC. Pixel and TDC information is ready to be read out. Fig. 4.12 shows the waveforms for this case. The read-out module reads the internal memory pixel by pixel, followed by the TDC information.

Event-driven mode for asynchronous applications (e.g. PET)

In this mode of operation, after the assertion of the synchronization signal, the SPAD array becomes sensitive to light for indefinite length of time until an event has occurred. The two parameters for the double-threshold aforementioned help to define what is considered an event in the following fashion. The SPADs in the array

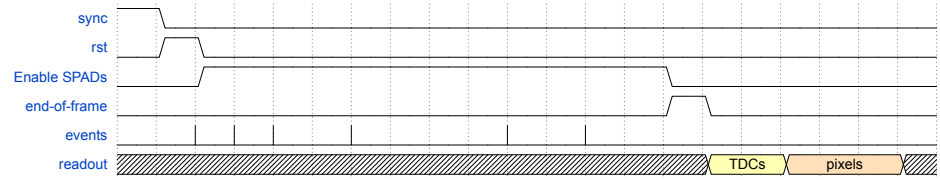


Figure 4.12: Waveforms when frame-based mode is used.

are generally triggered by photons or DCR effects. The event-driven operation unit calculates the DCR and compares it with the first parameter set during the configuration phase. If the DCR is higher than this parameter, the operation unit immediately switches the threshold to the second parameter and compares the new values of DCR with it. If this second threshold is surpassed, the unit concludes that there is an event and finishes the acquisition after a predefined integration time [4]. It shuts down the pixel array and asserts the signal end-of-frame to capture the reference TDC and the measurement is complete. In case the number of time-stamps does not exceed the defined threshold, the operation unit concludes that it was only noise and asserts the self-reset signal that will be spread along the pixel array and TDC banks. Thanks to the self-reset circuit it will only reset the pixel and TDCs that got fired; thus, minimizing the adversary effects of insensitivity-on-reset. The read-out module sends TDC information first; at the same time, the operation unit sums up all the pixels that have been fired so that the result is available by the time the TDC information has been sent. The read-out transmits the the addition result and sends the global internal time of Concolor. This internal timer is a 32-bit counter fed by the main clock of the system that corresponds to the exact moment the measurement window was ended. In a system composed by several Concolor sensors, this time is absolutely necessary when the system is operating in event-driven mode in order to have a global reference to frame the time-stamps captured by the TDCs. The waveforms of operation are shown in Fig. 4.13.

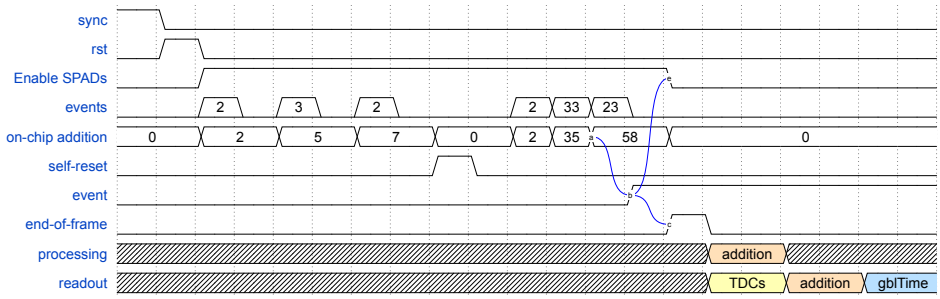


Figure 4.13: Waveforms for event-driven mode.

4.4.3. Commands of the digital core to test Concolor

The digital core has several commands that are meant for test that are optional but very useful when operating the system.

ECHO command: the digital core receives an ECHO command and retransmits the value to the FPGA. This command is used to check that the chip is alive. Specially the digital core, is alive.

READFIXREG: there is a register with a hard-coded value that can be read at any moment to check the communication is working.

MASKRDBK: this command is a combination of many much smaller commands and the purpose is to read back the masking pattern from the in-pixel memory. This procedure checks that the pixel array memory, digital core and global clock are fully operational.

4.5. Sliding Scale Time-to-Digital converters: first generation

4.5.1. Design principles

The TDCs were designed to achieve high time resolution while maintaining low power consumption. Several aspects of the MD-SiPM were considered to choose the architecture; the DCR and the time resolution of the SPADs being the most important. For low-level background light applications, like PET and indoors LiDAR, DCR represents the main source of hits for the TDCs. The probability of a TDC to be triggered by DCR is governed by the Poisson distribution as follows in Eq. (4.1).

$$P(hits > 0) = 1 - e^{-\lambda(1-M)NT} = 1 - 0.89 = 0.11, \tag{4.1}$$

where λ is the mean DCR, M is the masking factor (estimated 5%), N is the number of SPADs (64), and T is the full range of the TDCs ($1\mu s$). N and M are parameters that can be tuned at design time and operation time, respectively, to ensure the TDC availability at any given time during their maximum range. The bin size was chosen considering the time resolutions of state-of-the-art SPADs. The total jitter is given by Eq. (4.2)

$$J_T \propto \sqrt{J_{SPAD}^2 + 2.2Q^2/12 + J_{TDC}^2}, \tag{4.2}$$

where J_{SPAD} is the jitter of the state-of-the-art SPADs (estimated 100 ps), Q is the bin size, and J_{TDC} is the jitter of the TDC all in FWHM. The bin size was chosen to be 40 ps to make the second and third terms negligible with respect to the first term, which is the limiting factor. The term $2.2Q^2/12$ represents the quantization noise.

4.5.2. Architecture

Each MD-SiPM is equipped with a bank of 32 VCOs based on ring oscillators (ROs) that are constituted by nine pseudo differential stages. The RO schematic is shown in Fig. 4.14. Schematics and layout of the pseudo differential stages are shown in Fig. 4.16. The layout was equalized in terms of traces distance and capacitance to minimize DNL. The input impedance of the flip-flops varies at the rising edge of the clock; this is the reason every stage counts with two output buffers to prevent any interference of the latching process into the oscillation. The layout was designed such as the differential inputs (inp and inn) are located to geometrically match the output of the previous stage (outn and outp), so as to enable the ring oscillator to be perfectly abutted. One tap per stage has been placed to improve layout symmetry. The delay of the stages (t) can be configured within 40 and 150 ps by means of an off-chip PLL (implemented in FPGA in this case). The range of the operation frequency [calculated as $1/(18t)$] extends from 0.46 to 1.38 GHz. The VCOs are phased coupled along the sensor with the main objective of reducing the phase noise as demonstrated in [6]. The phase that is coupled is properly sized to compensate for extra capacitive loads.

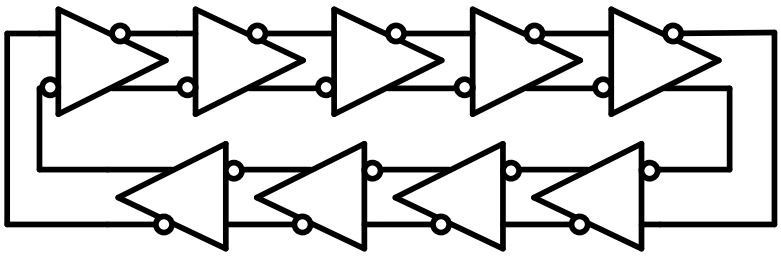


Figure 4.14: Schematic of the pseudo-differential ring oscillator with 9 phases used in the TDCs.

The coupling, shown in Fig. 4.17, is made through two nMOS–pMOS transistor pairs that connect the internal nodes inp and inn of the first phase of each $VCO_{(i)}$ to $VCO_{(i-2)}$ and to $VCO_{(i+2)}$. At the ends, the last VCO and the first one (the reference) make the bridge between the odd and even VCOs. The transistors can be externally controlled to provide the different degrees of coupling to achieve the different performance as explained in [6]. Four logic modules are attached to each VCO, thus totalizing 128 TDCs plus one extra reference TDC. The TDC architecture is shown in Fig. 4.18. The logic modules include five components: 1) a buffer; 2) a 10-bit LFSR counter; 3) a phase sampler; 4) a tri-state bus to read out the information; and 5) a buffer and 1-bit counter. The layout is shown in Fig. 4.19.

At last, the full layout of TDC bank of Concolor is shown in Fig. 4.20. The counters and sampling circuitry are supplied by a different source. The lines used for power supply are fully symmetric and distributed along the full length of the bank so as to avoid IR drops or imbalances in the heat generation. The oscillators are coupled forming a circular ring, thus the order of the TDCs in the bank is as follows: the first oscillator is connected to the third which is connected to the fifth, then the seventh and so on until the thirty-second oscillator. The ring is completed by

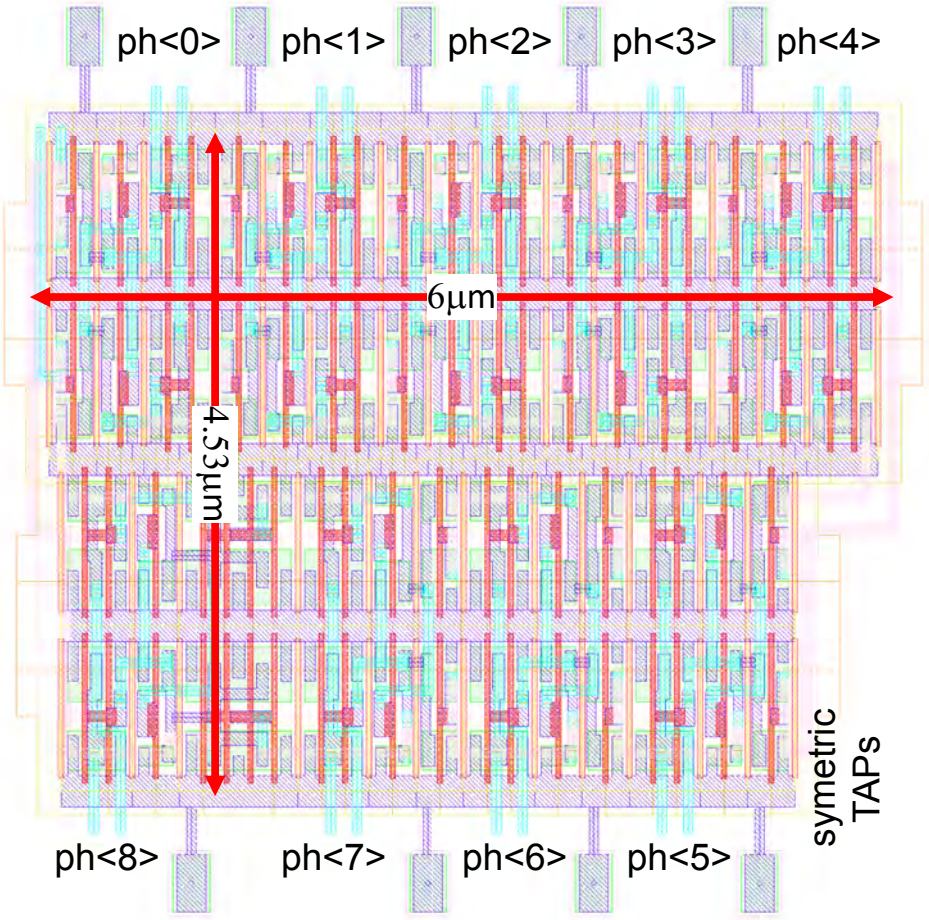
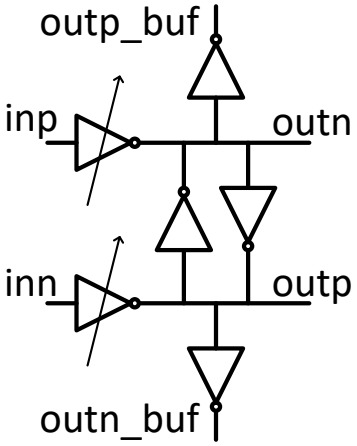
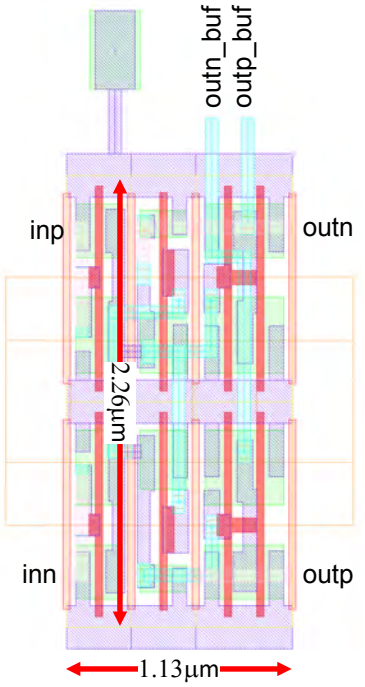


Figure 4.15: Layout of pseudo-differential ring oscillator with 9 phases used in the TDCs.



(a) Schematic of a single stage of the implemented ring oscillators. The buffers isolate the oscillation from the sampling circuit, thus preventing disturbances in the oscillation.



(b) Layout of single stage.

Figure 4.16: Pseudo differential stage of the ring oscillator.

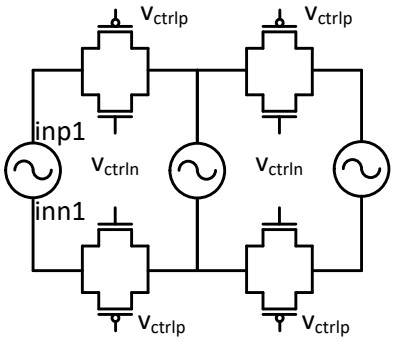


Figure 4.17: Schematic that shows the chosen coupling method. Only one phase of the VCOs is coupled.

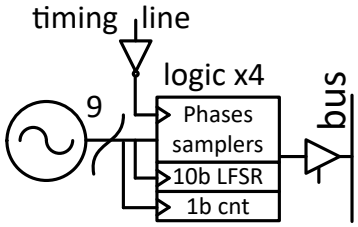


Figure 4.18: Diagram of the TDC sampling logic. There are 4 samplers per VCO. They are connected through tri-state buffers to a common bus to be read out.

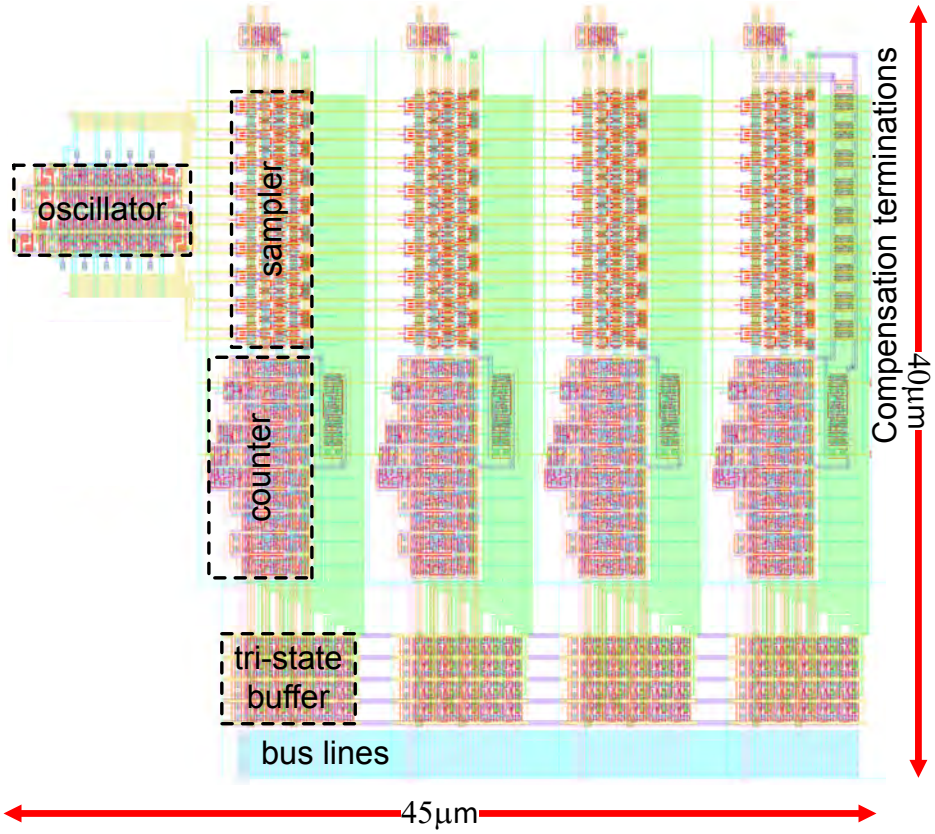


Figure 4.19: Layout of a ring oscillator and 4 sampling circuits.

connecting the thirty-second to the thirtieth, then to twenty-eighth and so on until the second oscillator is reached. The first and second oscillators have a connection loop against the reference.

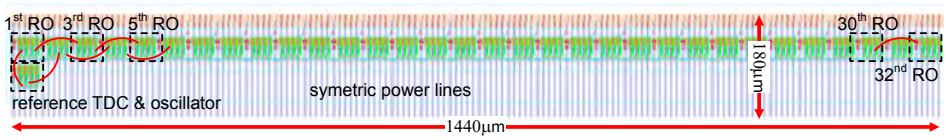


Figure 4.20: Layout of TDC banks.

4.5.3. Operation

The TDCs measure the time from the moment an event occurs to the end of the frame by means of a coarse LFSR counter and a fine scale formed by the phases of the VCO. Since the system can work in two modes, event-driven and frame-based, the end-of-frame signal could come at any time synchronized with the clock of the system. Fig. 4.21 shows the waveforms and the way the TDCs operate. This operation is essentially the same for both modes.

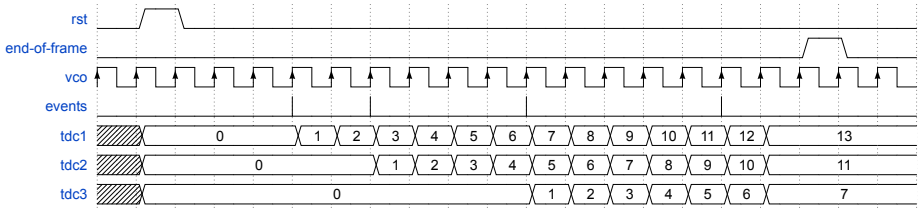


Figure 4.21: Basic operation of a TDC. After the reset signal is asserted, the counter remains at 0 until an event occurs. Then it counts until the end-of-frame signal is asserted by either of the modes (event-driven or frame-based).

Eq. 4.3 shows the way every time-stamp is calculated. The time-stamp T, in number of bins, is expressed as

$$T = (cnt + CC) * N + P_{vcof} - P_{vcos}, \tag{4.3}$$

where cnt is the LFSR counter, CC is a correction applied to the counter, N is the number of phases (18), P_{vcof} is the final phase, and P_{vcos} is the start phase of the VCO. Although this is the most natural way to calculate the time-stamp, it would double the area and the power of the samplers and the transmission time since every TDC requires two phases to obtain the time-stamp. Taking advantage of the fact that the phases are coupled, only one phase (reference) is sampled at the end of the frame. Eq. 4.3 can be rewritten as follows in Eq. 4.4:

$$T = (cnt + CC)N + (P_{reff} - PS(vco)) - P_{vcos} \tag{4.4}$$

where P_{reff} is the final phase of the reference VCO and PS is the phase shift between a given VCO against the reference. Phases and counts are expressed in number of bits.

The sliding scale technique has two undesirable effects that introduce errors in the coding of the TDCs. Fig. 4.22 shows the first effect that takes place when the trajectory from the start phase of the VCO to the final phase of the VCO encloses the phase of the counter. If both the cases are compared (left and right), the actual total time is less than 1 count, but the case where the phases enclose the phase of the counter will have 1-count error in excess. This needs to be corrected. The second effect, shown in Fig. 4.23, happens when the start phase of the Vco occurs when the clock of the counter is at high-level; thus, making the counter instantly increase 1 count and needs to be corrected. In case the start phase occurs when the clock of the counter is at low-level, the counter remains unchanged and nothing should be done. Fig. 4.24 shows raw data coming from the TDCs after decoding by Eq. 4.4 (a). In this case the counter might have 1-count error or even 2-counts error if both the effects are present. The plot (b) shows the TDC values after correcting the type-1 effect. The plot (c) shows the TDC values after appling type-2 effect correction. At last, in plot (d), both the effects are corrected and the TDC values are error-free.

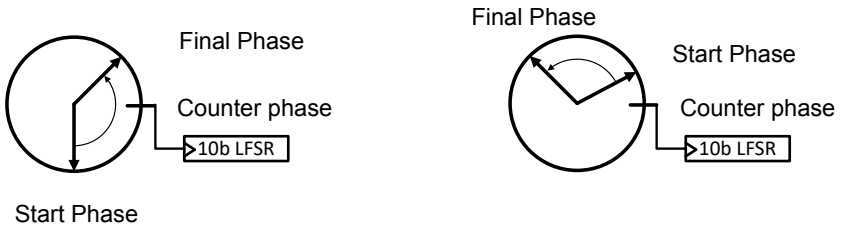


Figure 4.22: Error type 1. The final phase and start phase encloses the phase of the counter. On the left, the counter has 1 count in excess; needs to be corrected. On the right, the counter is correct.

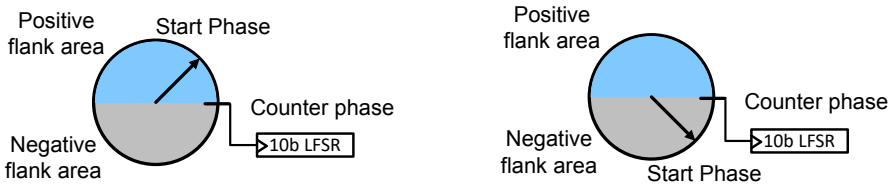


Figure 4.23: Error type 2. The start phase pulse comes when the clock of the counter is high; thus, causing an instant increment of the counter. On the right, the start phase comes when the clock of the counter is at low-level, so no increment occurs.

In principle, the VCOs of Concolor are running all the time and they are available at any given moment a measurement begins, thus achieving the maximum

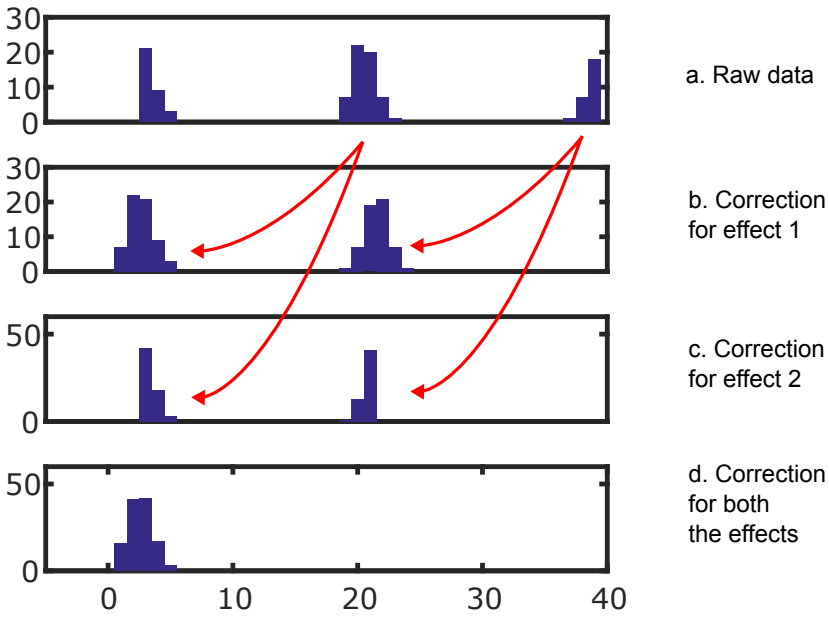


Figure 4.24: Coding of the TDCs. Compensation for both the effects caused by sliding-scale technique.

throughput of the chip. However, in order to save power, when Concolor is not operating at its maximum speed, they can be shut down and turned on only a few microseconds before it's needed. For intermittent operation, this enables a type of "stand-by" mode because the VCOs consume most of the power of Concolor, the digital core consumes only leakage power when is sleeping, and the array does not take any power if it is not getting periodically reset. In the same fashion, for applications where the resolution is not extremely important, or the maximum resolution of Concolor is not needed, the VCOs can be slowed down. In the tested implementation, this was achieved by means of a DAC that varies the voltage control of the VCOs.

The detailed operation is as follows: after a photon arrival, the timing line is pulled down, and the buffer activates the samplers that latch the phases of the VCO (Pvcos). The counter, fed by one of the phases of the VCO (Pc), starts running from that moment until the end of the frame when it is stopped by the shutter. The shutter signal is controlled by the digital core and is asynchronous with the VCO. As consequence, if those events happen simultaneously, it is not possible to distinguish whether the counter included the last VCO cycle or not (off-by-one error). In order to mitigate this problem, a 1-bit counter was added. Fed by a different phase (Pb), this 1-bit counter can be checked to know if the main counter should be odd or even; the variable CC in 4.4 will get the values 0 or 1 accordingly. The phases Pc and Pb are 180◦ apart to ensure that at least 1 counter is always correct. If P_{vcof} is equal to Pc, the counter might have incurred in an off-by-one error and should be modified according to the value of the 1-bit counter. If P_{vcof} is different from PC, the counter is correct and it does not require any correction.

4.5.4. Calibration of the TDCs

Though the phases of the VCOs are tightly coupled, there is a small phase shift between every $VCO_{(i)}$ and $VCO_{(i+2)}$ that accumulates over the sensor. This effect worsens as the speed of the VCOs increases. There is a second effect in the sampling logic due to the distance between the samplers and the VCO they are connected to. Both effects can be accounted in a look-up table with every phase of VCO plus phase of the sampler. The table can be built up by firing all of the SPADs at the same time with a synchronous laser. Fig. 4.25 shows the phase shift for every VCO+sampler of the sensor when operating at maximum speed.

4.5.5. Sliding Scale technique study

The sliding scale technique is a proven method that has worked very well for ADCs; in this thesis, it was used to reduce the DNL of the TDCs [7]. The VCOs are asynchronous with the clock of the system and the window frame, therefore every time-stamp taken is measured by a different phase of the VCO. This method can compensate any mismatch in the layout of the ROs, and furthermore any local and chip-level transistor mismatch. In order to calculate the impact of the sliding scale, the VCOs were measured with a random pulsed laser in 300-ns frames. The minimum and maximum DNL were calculated for the start phases of the VCOs and for the time-stamps. The results, shown in Fig. 4.26, exhibit an improvement of 6.25

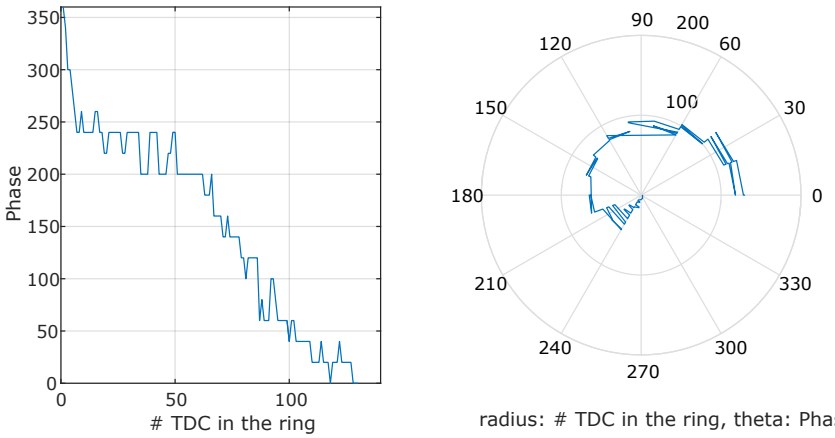


Figure 4.25: Left: phase shift of every VCO+samplers with respects to the reference. Right: polar plot of the phase shift of every VCO+samplers. This information is used to calibrate the TDC bank.

times. Fig. 4.27 is a 2D histogram of all the phases for all the TDCs that form the bank. For this measurement, the chip was exposed to very low-level light to avoid saturation of the TDCs shich will mislead to a higher DNL levels.

The reader might notice that 2 TDCs have a missing code. This only happens when the maximum frequency is used due to a glitch of one specific bin in the TDC spectrum. However, it does not represent a problem since the probability to encounter this glitch is as low as 0.014% and it only happens for few TDCs. The maximum DNL of those TDCs is still at a level of 0.12 LSB when that problematic bin is discarded. The cause of this effect is an asymmetry in the layout that leads to uneven IR drops in the TDCs supply.

4.5.6. Timing performance of the system

A laser, synchronous with the system, was employed to characterize the time response of every pair semicolumn-TDC. The hits originated by any SPAD in the semicolumn were used to build a histogram and the jitter was calculated at full-width at half-maximum (FWHM). The results are shown in Fig. 4.28. The total jitter calculated includes the jitter contributions of the laser, SPAD, timing line, TDC, and FPGA. The SPAD is the main contributor. In some applications, particularly in PET, the single photon time resolution (SPTR) is an important parameter to characterize the system as explained in [7]. It essentially describes the uncertainty in time of the whole system when a single photon impinges the sensor. Fig. 4.29a was obtained by measuring 1 million hits and shows the resolution for 850nm source for FWHM, full-width-at-tenth-maximum (FWTM), and full-width-at-1%-maximum (FW1pM). Resolution at FWHM is 161ps. The same procedure was used to obtain the results for 766-nm wavelength. The SPTR is 194ps at FHWM, 529ps (FWTM), and 1.12ns (FW1pM). Results are shown in Fig. 4.29b.

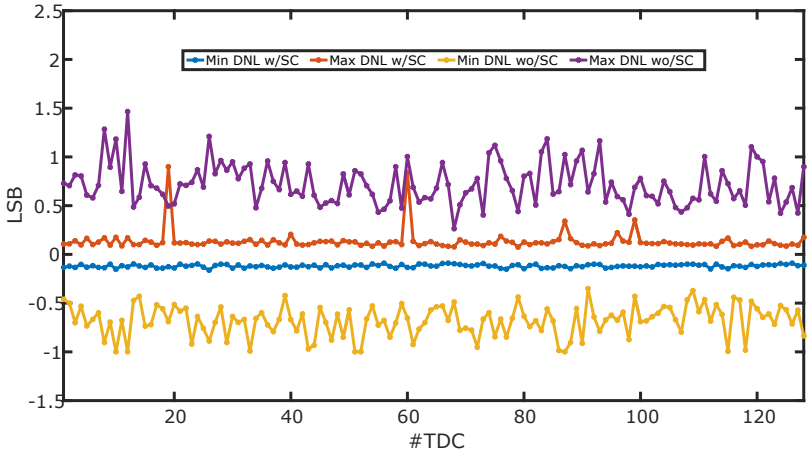


Figure 4.26: Comparison between the DNL of the VCO itself and the DNL of the TDCs when using the VCO asynchronously with the clock of the system (sliding scale technique). Maximum and minimum DNL are shown for the VCO in purlple and yellow lines respectively. Maximum and minimum DNL of the sliding scale technique are shown in blue and red lines respecteively. The DNL 6.25 times.

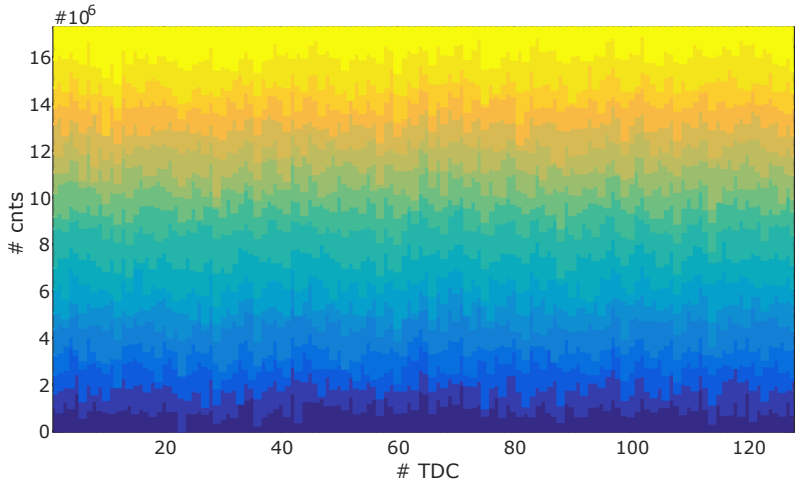


Figure 4.27: Stack plot for each TDC. Each stack is divided into 18 phases represented by the colored segments whose lengths are proportional to the histogram of the DNL measurement.

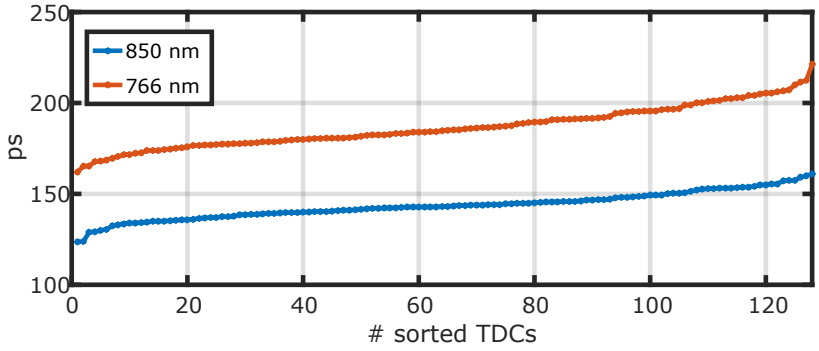
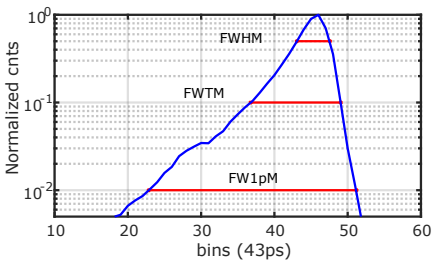
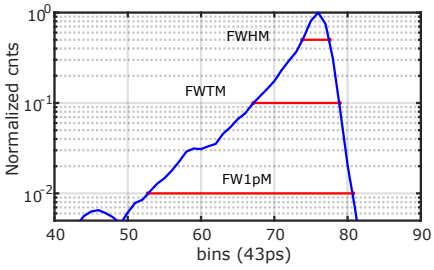


Figure 4.28: Time resolution per TDC for both 766 and 850 wavelengths. The TDCs are sorted from best to worst jitter.



(a) SPTR is 194ps (FWHM) for a 766nm source.



(b) SPTR is 161ps (FWHM) for a 850nm source.

Figure 4.29: SPTR for different wavelengths.

Another important parameter is the multiphoton time resolution (MPTR) that describes the resolution of the system when several photons impinge any combination of SPAD-TDC. Fig. 4.30 shows the resolution of the time of arrival when multiple photons impact the sensor. Four different methods were used to calculate the time of arrival. Averaging shows a poor result due to the non-Gaussian time response of the SPAD. The first-photon method considers only the first photon to calculate the time of arrival. Last, the maximum-likelihood-estimation (MLE) [8] method exhibits the best result as it accounts for the TDC and SPAD response. It finds the estimator with the highest probability given a set of values. Interpolated Maximum Likelihood is a purposed method in this work that linearly interpolates the response of the TDCs to mitigate the quantization error. For this technique, a table with 40 bins by 128 positions for every TDC is built; every position of the table stores the responses of the TDCs when events occur in the line the TDC is serving. As an example, the response of the first TDC is shown in Fig. 4.31. As expected, the jitter of the branch highly corresponds to that of a single SPAD as the SPAD is the physical limiting factor. Table 4.2 shows the performance summary for the TDCs in Concolor.

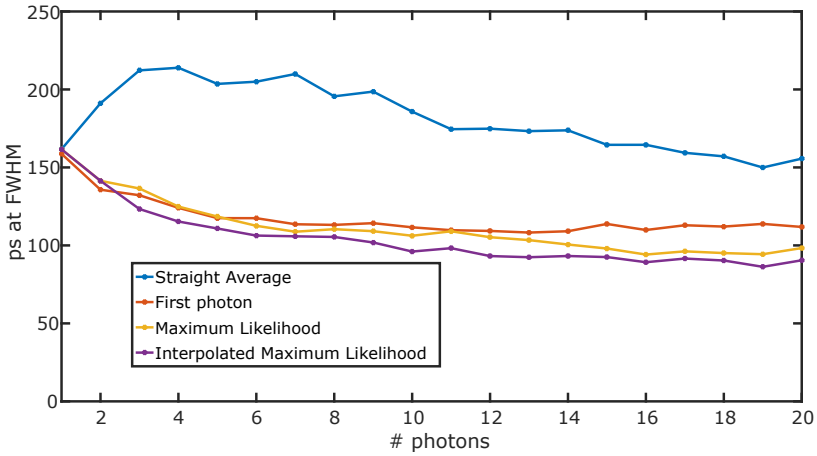


Figure 4.30: Multi Photon Time Resolution for 4 different methods. Interpolated maximum likelihood (purposed in this work) improves by reducing quantization error.

4.6. Distributed PLL for frequency adjustment

In previous sections, TDCs were thoroughly presented, along with the time calculation formula and a calibration method to enhance the accuracy and resolution. However, nothing was said about the specific frequency that the VCOs are running at. Concolor has 4 quadrants whose TDC banks are independent from each other. This means, that when the 4 quadrants are utilized for the same application, the frequencies their VCOs are working at, since they are open-loop, will vary from

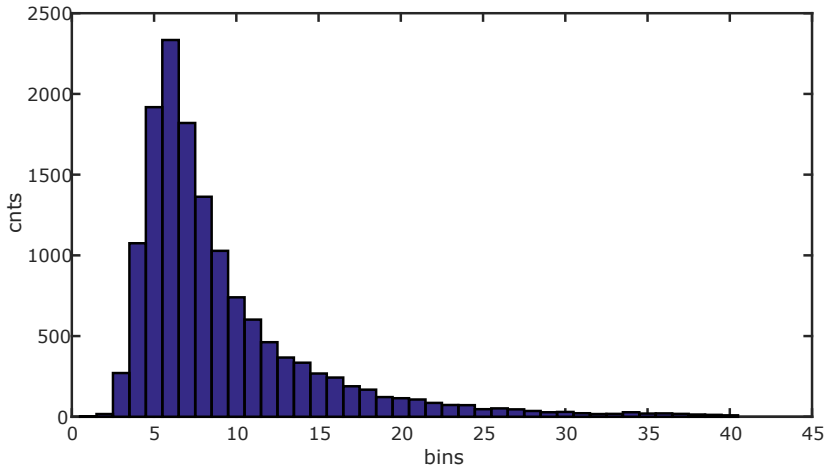


Figure 4.31: Response for the first branch of the system. The response comprehends SPAD, timing-line and TDCs all combined.

Type	Value
Bin size	40 to 150 ps (V_{ctrl} = 1.21V and 0.8V
Peak-to-peak jitter	6,3 ps
Added jitter at 100 ns	55ps
Power per TDC	56uW when working 100%
Power full chip 512 TDCs	29 mW (at 75ps)
Sliding scale	<0.4
DNL Range	0.9 to 1.3 us
Total DNL	-0.1 + 0.12
INL	<1
Area	625 (um)^2

Table 4.2: Table of performance of TDCs

quadrant to quadrant; moreover if the whole system is composed by several Concolor modules. Temperature will also affect the frequency and the calibration. PLLs are a well-known method to control the frequency of oscillators to make it match to that of a reference. In an image sensor, area is a scarce resource and designs aim at maximizing active area to increase sensitivity. In this work it is purposed a distributed PLL that takes the biggest parts of PLL systems off the chip.

4.6.1. Architecture

The architecture of the distributed PLL (DPLL) is shown in Fig. 4.32.

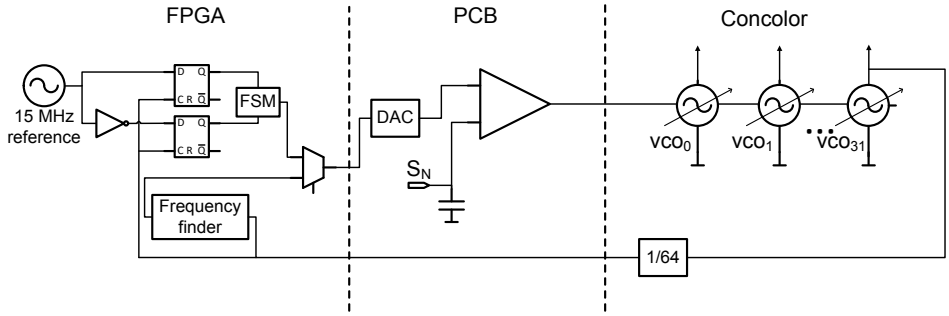


Figure 4.32: Circuit of Distributed PLL. VCOs and counter are on-chip. Phase comparator on the FPGA and DAC plus amplifier on the PCB.

Only the necessary parts of the PLL are on-chip. VCOs need to be sampled by the TDCs so they must be as close as possible to the samplers. The counter, that in principle can be programmable, is also located on-chip so as to reduce the frequency that travels off-chip through the bonding wires. The rest of the components are located off-chip. Making use of the crystal that is on board in FPGA systems, a frequency finder and phase comparator was designed. The FPGA sends the digital number of the voltage that needs to set to an external DAC that sets the low-noise amplifier to control the VCOs.

4.6.2. Operation and results

The DPLL operation has two stages. After power-on, the frequency of the VCOs is unknown and can be far away from the reference. For this reason the firmware activates the frequency finder module, whose diagram is shown in Fig. 4.33. It uses the internal clock to count the number of pulses coming from the VCOs for a programmable number of times that is used to average out the result. It compares the result with the period of the clock system. If it is lower, the firmware increases the voltage and it decreases the voltage otherwise. The number of times that the result is averaged is increased to make the frequency of the VCOs converge to the reference value. After this first stage, the VCOs and reference are oscillating at same frequency with an error of $1/(N+1)f$, where N is the number of measurements averaged and T the period of the reference. In case the reference is 15MHz and $N = 64K$ (16 bits), the error of the frequency is 228Hz. This is already enough

to operate Concolor as the phase comparison is done when the time-stamps are calculated. Additionally, a bang-bang phase comparator [9] is used to finely tune the phases of the VCOs, thus leading to fully in-phase synchronized TDC banks.

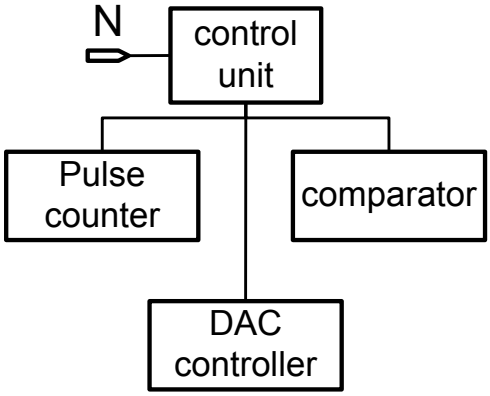


Figure 4.33: Block diagram of the frequency finder module.

The results for the frequency finder are shown in Fig. 4.34. The open-loop frequency / N across different quadrants and chips vary from 15.5MHz to 16.5MHz, this means the VCOs frequency band is between 992MHz and 1.056 GHz. Once the frequency finder is activated, all the VCOs run at 1.024GHz that is the reference multiplied by 64.

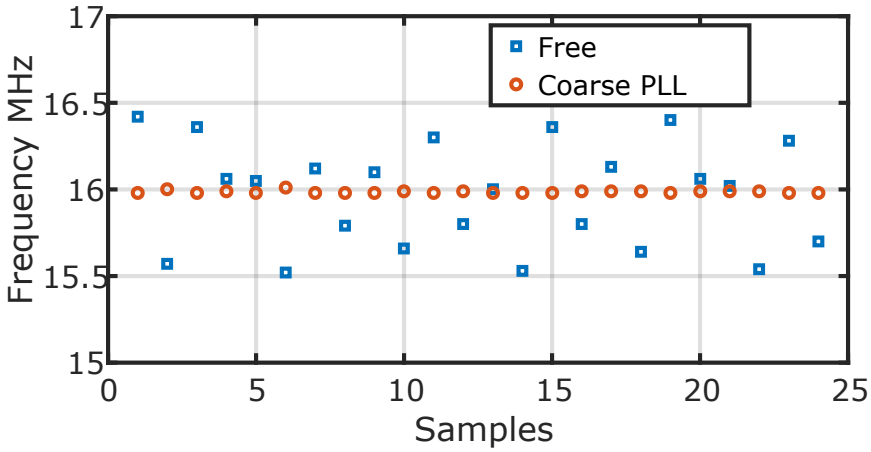


Figure 4.34: PLL measurements for 24 samples for open-loop and closed-loop.

4.7. Time-of-Flight Applications

PET is a noninvasive medical imaging technique to generate a 3-D image of the tissue of interest as explained in [10]. An image sensor can indirectly detect gamma photons (511 keV) by means of a scintillator that absorbs gamma radiation and generates a shower of visible photons that can be time-stamped by the sensor. Fig. 4.35 shows the sensor coupled with an LYSO scintillator

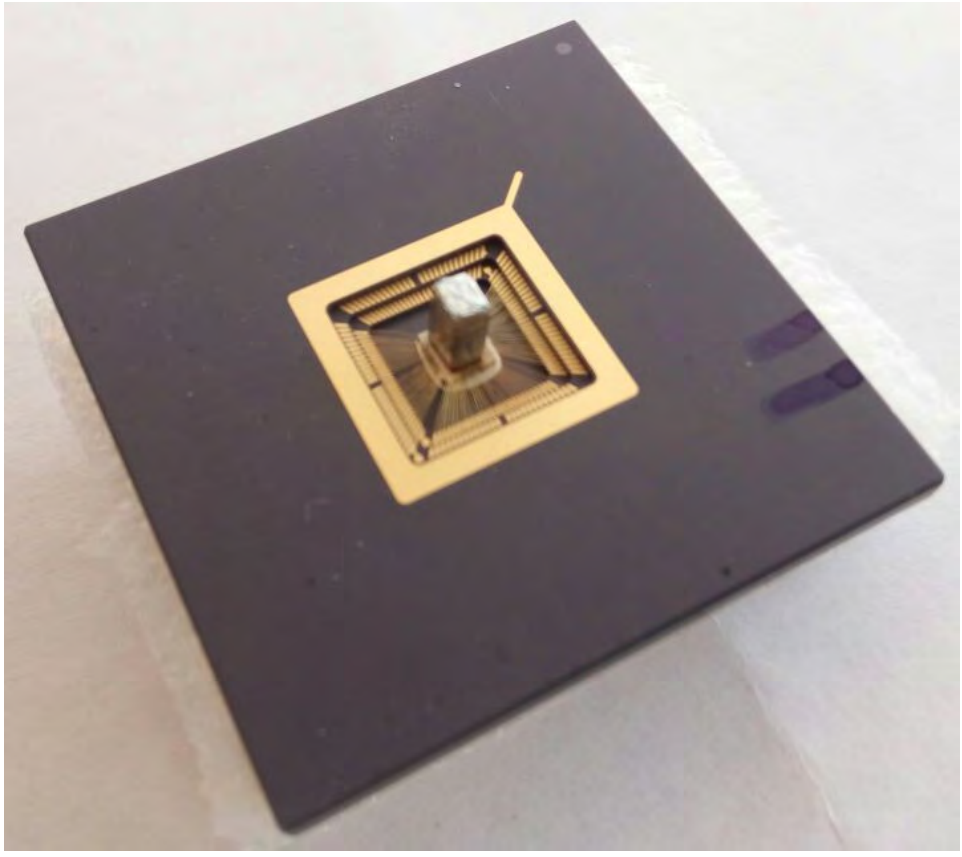


Figure 4.35: Scintillator coupled to Concolor. The scintillator is covered by high-reflective white interface and aluminium foil.

4.7.1. Positron Emission Tomography

Energy Resolution

The number of photons detected is proportional to the energy deposited by the gamma photon into the crystal. A lower energy than the initial energy signifies that the gamma photon lost energy by scattering meaning it cannot be used for the tomographic reconstruction and should be dismissed. Hence, the importance of the measurement of the energy deposited into the crystal. The scattering process,

fully described in [11], is ruled by the Compton's law and it can be deduced that the minimal energy that a gamma photon might lose equals 1/3 of its initial energy, therefore the resolution must be within that limit. Fig. 4.36 shows the spectrum of a ^{22}Na source, exhibiting an energy resolution of 20%. Both peaks of ^{22}Na are shown.

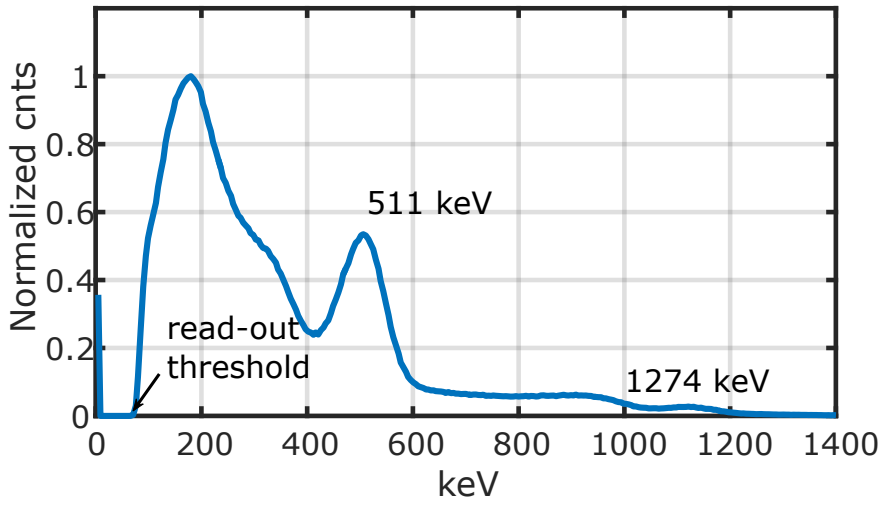


Figure 4.36: Resolution for Sodium source is 20%. Both the peaks at 511keV and 1274 keV are shown.

Linearity

Some other applications, like spectroscopy, depend on the linearity of the system. This assessment was performed by measuring five different radiation sources using the whole sensor that includes four MD-SiPMs. A histogram like Fig. 4.36 was built for five radiation sources for each MD-SiPM and its peak coordinates were extracted to build the linearity plot shown in Fig. 4.37. The peak of ^{22}Na can be seen at (x = 511 keV and y = 450 keV). The nonlinearity obtained is 2% and it can be further improved by applying the saturation-correction curve of the MD-SiPM explained in the introduction. Fig. 4.38 shows the same information expressed as a percentage deviation with respect to the ideal response.

Coincidence time resolution:

The single-photon coincidence time resolution (SPCTR) is a very important parameter when two modules are working synchronously in PET systems. For this measurement, two systems using Concolor were employed with a synchronization module. A 405-nm laser in combination with a laser splitter and two diffusers were employed to make the dual optical path. Fig. 4.39 shows the scheme of the measurement. The synchronization module is actually part of the firmware in the FPGA. It takes the clock alternatively from the FPGA board or from an external PLL to distribute the

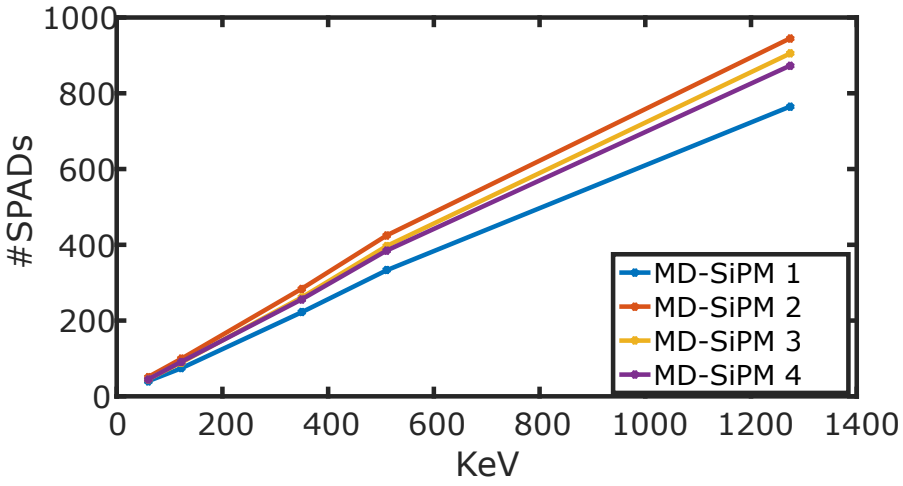


Figure 4.37: Radiation linearity of Concolor for the 4 different quadrants.

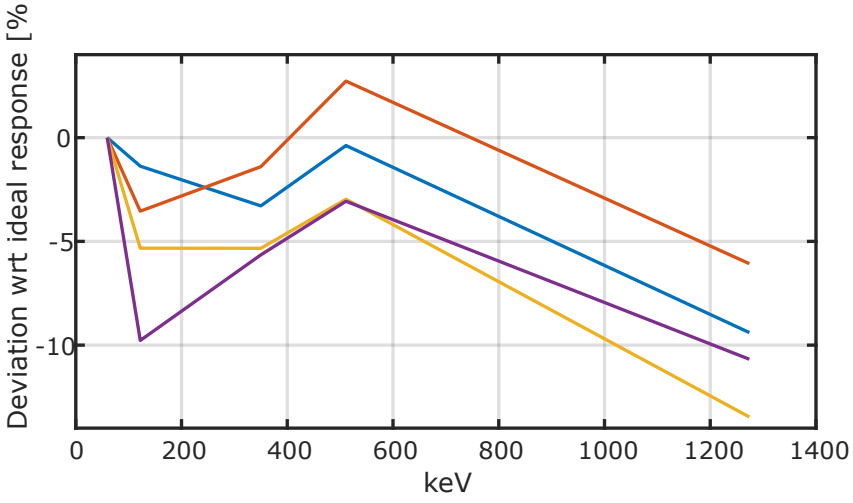


Figure 4.38: Deviation of the radiation linearity with respect to the ideal response, expressed in percentage.

clock along the modules. The results are shown in Fig. 4.40, where the measured CTR is 244 ps.

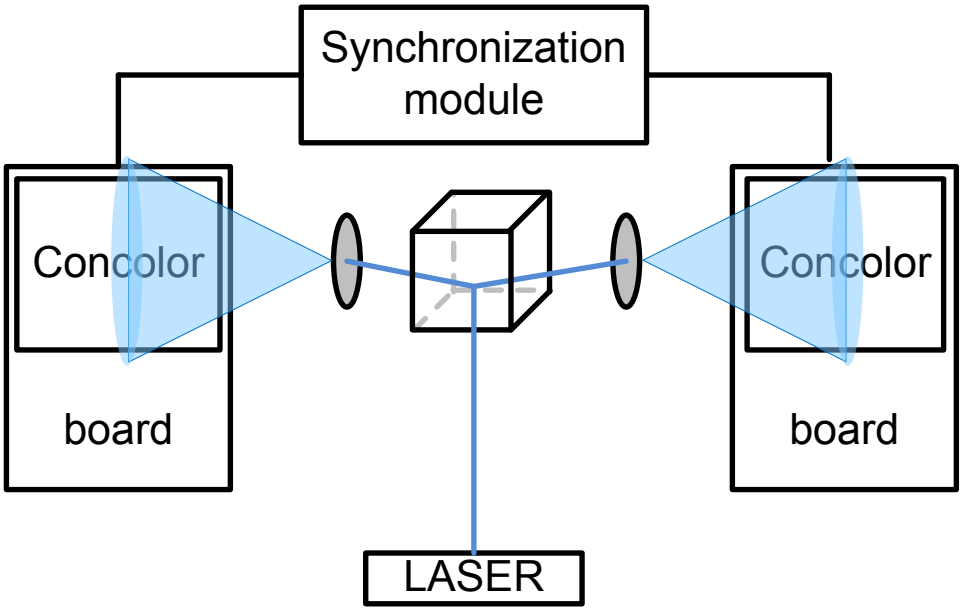


Figure 4.39: Measurement set-up diagram for PET operation.

The total jitter of CTR can be expressed as $J_{CTR} = \sqrt{J_{system1}^2 + J_{system2}^2 + J_{sync}^2}$, where J_{sync} is the jitter introduced by all the electronics for synchronization: both the FPGAs, the PLL and the connections.

Since the systems are both Concolor-based,

$$J_{system1} = J_{system2} = \sqrt{J_{SPAD}^2 + J_{ampl}^2 + J_{line}^2 + J_{TDC}^2}.$$

Thus,

$$J_{CTR} = \sqrt{2(J_{ampl}^2 + J_{line}^2 + J_{TDC}^2) + 2J_{SPAD}^2 + J_{sync}^2} = \sqrt{J_{electronics}^2 + 2J_{SPAD}^2} = 244\ ps.$$

It means that the jitter introduced by the electronics all combined is less than 50ps. It is negligible with respects to the jitter of the SPAD.

4.7.2. 3D Imaging/LiDAR

3D imaging is a topographic method to create a 3-D graphical representation of a physical target. The working principle is based on the illumination of a scene with a pulsed laser and the detection of the photons that reflect off the target. By calculating the time of arrival of these photons, it is possible to create a representation model with X, Y, and depth information. There are two main approaches: 1)

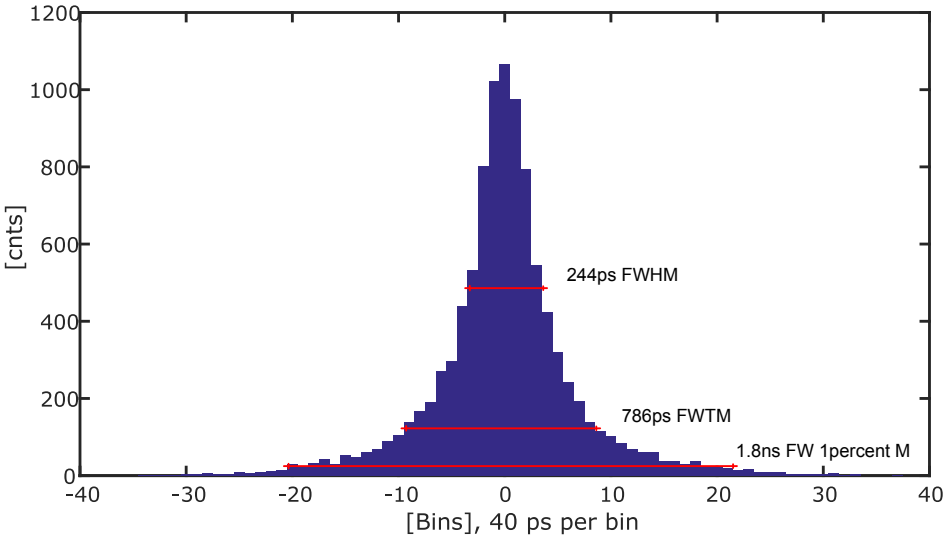


Figure 4.40: Coincidence Resolution Time (CRT) of the system.

flash and 2) scanning methods; both are explained in detail in [12] and [13]. Every photon absorbed by the sensor has a time-stamp used to calculate the depth, and its position in the array is used to calculate X and Y in the scene. Fig. 4.41 shows a 3-D image generated by a flash technique; it displays a resolution of 6.5mm. For every single photon, the space resolution is about 2.4cm (160ps). For flash LiDAR, multiple measurements were performed to improve the resolution by averaging. Assuming that every measurement can be represented by a Gaussian distribution with mean μ and standard deviation σ, the average of M measurements has a Gaussian distribution with media μ and standard deviation $\sqrt{(M)}/M\sigma$. As M increases, the resolution becomes finer and finer. In this example, 1 LSB (43ps) was used to define the spatial resolution. If the previous equations are combined, $M = 13.8$; thus, at least 14 measurements per point are required. The system, working in the frame-based mode, can provide up to 128 depth data per 5.2µs (1 frame). In Fig. 4.42, a actual 2D picture taken with a camera is shown.

4.7.3. Distance Measurements, ranged method

A pulsed non-diffused laser is pointed to the object whose distance to the sensor is wanted. In this experiment, the shutter is open for the whole range. Fig. 4.43 shows the error of the measurement for M = 1000. For short distances, the parallax problem between the laser and sensor dominates.

4.8. Summary of the sensor

The performance of the sensor is summarized and compared to other state-of-the-art works in Tab. 4.3 that were relevant for PET applications at the time this sensor

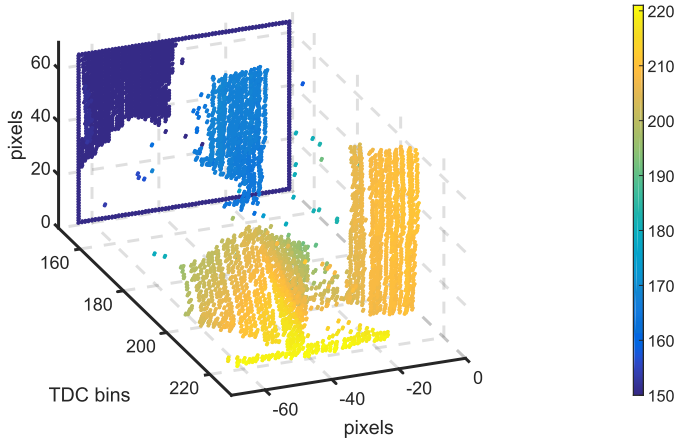


Figure 4.41: 3D picture reconstructed from X-Y-Z position information of the sensor. The resolution in Z is 6.5 mm. The methods utilized to generate the image employ a 3x3 matrix that cannot be calculated for the border of the image, thus creating a frame with no value at the border.

Figure 4.42: Actual 2D picture of the scene taken by a regular camera.

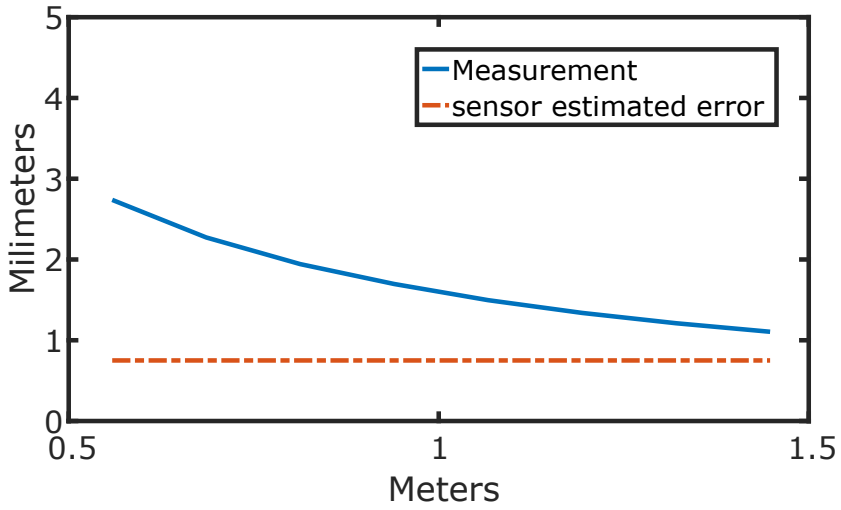


Figure 4.43: Near distance resolution after parallax correction. The error cannot be fully corrected due to the uncertainty on the exact position of the chip respects to the frame in the prototype.

was introduced to the scientific community.

	Braga et al. JSSC 2014	Carimatto et al. ISSC 2015	Frach et al. JSSC 2014	Concolor
SPAD/SiPM	180 ev/10 ns	416	6396	8192
TDC LSB (ps)	64.5	48.5	23	40
TDC DNL (LSB)	-0.24+0.28	-.75+.75		-0.1 +0.12
TDC INL (LSB)	-3.9+2.3	-2+4		<1
Pitch (mm x mm)	16.2x16.2	30x50	59x64	19x5
SPTR (ps)	266	327	153	162
Power/TDC (mW)	948	1500		171
Energy res (%)	10.2	15	11	20

Table 4.3: Comparison to the state-of-the-art designs.

4.9. Conclusion

The first MD-SiPM in 40-nm technology has been designed and presented. Its performance was demonstrated for PET, 3-D vision, and light ranging applications. The new TDC architecture was proven to obtain the expected resolution of 161 ps (FWHM) with a low power consumption of 12 mW per TDC bank which makes it suitable for image sensors. The sliding scale technique reduced the DNL of the VCOs (from 0.75 to 0.12 LSB). The level of integration achieved by the digital core largely facilitates the usage of the system and the scaling for synchronous applications.

The calibration by means of a synchronous laser that it is needed to find the phase relationship between the oscillators and the reference is too time-costly and can be replaced by an electrical calibration in next versions of the chip for simplicity. Alternatively, an in-situ self-calibration or phase comparison is possible. This improvement was included in MindHive, which will be describe din 6.

References

[1] A. Carimatto, A. Ulku, S. Lindner, E. Gros-Daillon, B. Rae, S. Pellegrini, and E. Charbon, Multipurpose, Fully Integrated 128 × 128 Event-Driven MD-SiPM With 512 16-Bit TDCs With 45-ps LSB and 20-ns Gating in 40-nm CMOS Technology, IEEE Solid-State Circuits Letters **1**, 241 (2018).

[2] S. Pellegrini, B. Rae, A. Pingault, D. Golanski, S. Jouan, C. Lapeyre, and B. Mamdy, Industrialised SPAD in 40 nm technology, in *2017 IEEE International Electron Devices Meeting (IEDM)* (2017) pp. 16.5.1–16.5.4.

[3] A. Carimatto, S. Mandai, E. Venialgo, T. Gong, G. Borghi, D. R. Schaart, and E. Charbon, 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.

[4] M. Chin, M. F. Bieniosek, B. J. Lee, and C. S. Levin, Integration time window for pulse width modulation readout of silicon photomultipliers for 0.5 mm resolution 3-D position sensitive PET scintillation detectors, in *2014 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)* (2014) pp. 1–2.

[5] N. Calandri, M. Sanzaro, L. Motta, C. Savoia, and A. Tosi, Optical Crosstalk in InGaAs/InP SPAD Array: Analysis and Reduction With FIB-Etched Trenches, IEEE Photonics Technology Letters **28**, 1767 (2016).

[6] A. Ronchini Ximenes, P. Padmanabhan, and E. Charbon, Mutually Coupled Time-to-Digital Converters (TDCs) for Direct Time-of-Flight (dTOF) Image Sensors, Sensors **18** (2018), 10.3390/s18103413.

[7] B. Markovic, S. Tisa, F. A. Villa, A. Tosi, and F. Zappa, A High-Linearity, 17 ps Precision Time-to-Digital Converter Based on a Single-Stage Vernier Delay Loop Fine Interpolation, IEEE Transactions on Circuits and Systems I: Regular Papers **60**, 557 (2013).

[8] T. Brox, Maximum Likelihood Estimation, in *Computer Vision: A Reference Guide*, edited by K. Ikeuchi (Springer US, Boston, MA, 2014) pp. 481–482.

[9] S. Tertinek, J. P. Gleeson, and O. Feely, Binary Phase Detector Gain in Bang-Bang Phase-Locked Loops With DCO Jitter, IEEE Transactions on Circuits and Systems II: Express Briefs **57**, 941 (2010).

[10] J. W. Cates and C. S. Levin, Evaluation of a TOF-PET detector design that achieves ≤ 100 ps coincidence time resolution, in *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)* (2017) pp. 1–3.

[11] M. K. Nguyen, T. T. Truong, M. Morvidone, and H. Zaidi, Scattered Radiation Emission Imaging: Principles and Applications, International Journal of Biomedical Imaging **2011**, 913893 (2011).

[12] M. Beer, O. M. Schrey, C. Nitta, W. Brockherde, B. J. Hosticka, and R. Kokozinski, 1×80 pixel SPAD-based flash LIDAR sensor with background rejection based on photon coincidence, in *2017 IEEE SENSORS* (2017) pp. 1–3.

[13] K. Ito, C. Niclass, I. Aoyagi, H. Matsubara, M. Soga, S. Kato, M. Maeda, and M. Kagami, System Design and Performance Characterization of a MEMS-Based Laser Scanning Time-of-Flight Sensor Based on a 256 × 64-pixel Single-Photon Imager, IEEE Photonics Journal **5**, 6800114 (2013).

5

Panther: 2D and 3D system fabricated in 40nm ST technology for LiDAR and Positron Emission Tomography assessment.

When you change the way you look at things, the things you look at change.

Max Plank

The sensor presented in this chapter is an implementation for LiDAR systems. By exploiting 3D integration, large amounts of electronics can be utilized while keeping enough space for the active area. Two nodes with different capabilities were used for two very distinct goals. Implementation of the electronics benefited from 40nm ST technology to achieve high level of integration.

5.1. Introduction

Panther was conceived as a platform to test new methods and ideas on a brand-new technology to be used for the main ToF applications nowadays which are LiDAR and PET. The nature of the technology enabled 2D and 3D-integrated silicon, making the SPADs Front Side Illuminated (FSI) and Back Side Illuminated (BSI) respectively. Due to the structure of these types of SPADs, FSI SPADs are preferable for those applications where wavelength are closer to blue spectrum and BSI SPADs are a better choice for applications where wavelengths are closer to the red spectrum. This is due to the depth of penetration of the red and blue wavelengths on silicon. This would make, in principle, BSI SPADs not useful for PET applications. However, thinning techniques were used to try to overcome this problem. By thinning down the back part of the silicon, it is possible to make BSI SPADs suitable for the depth of penetration of blue wavelengths. 3D integration enables two interconnected tiers, one for SPADs and another for electronics. As a consequence, increasing the available area for electronics to enable new features and capabilities to the sensor. The results for these SPADs were published in [1]. From this idea, two sensor families were born: Concolor (2D) and Panther (3D). In this chapter, all the components of Panther will be explained and its results will be shown and discussed. The sensor was jointly designed with Dr. Augusto Ximenes who introduced the concept of decision makers and implemented the main TDCs (the method is fully explained in [2]) and the timing lines that will be succinctly explained in order to understand the system. Dr. Scott Lindner also participated in the design with a flexible programmable quenching module that can work in the passive and active modes and the cascode quenching system to allow for higher excess bias [3]. This work presents the digital core of the system, along with the memory for the TDCs, auxiliary and extended TDCs, electronics and static memories.

5.2. Architecture

Panther was designed in 3D ST 65-40nm CMOS technology. It essentially has two tiers interconnected face-to-face with pillars. The top tier holds 4096 18.2x18.2 μm^2 SPADs and the bottom tear has all the electronics that serves those SPADs. Fig. 5.1 shows the complete block diagram of the bottom tear of the sensor. The active area of the sensor is organized in four quadrants and the electronics for each quadrant is folded in such a way that takes half of the space of the SPADs that are placed on top in order to leave area to allocate the rest of the components. Each quadrant has 16 dual columns of 32 pixel-circuitry modules, for a total of 1024 pixels. In the center of the sensor, 64 TDCs are placed to time-stamp the hits coming from the pixels through timing lines that are equidistantly distributed to minimize the skew along the sensor. The timing line goes side-by-side with the address line to catch the addresses of the pixels that were hit. A memory right below takes the TDC values to be transmitted to the digital core which collects not only timing and address information from the hits but also the addition or energy count to send to the FPGA. The digital core also handles the configuration information such as masking and several parameters that are stored in registers. Additionally to the

main TDCs, the sensor has a group of 8 auxiliary TDCs and a group of 64 eXtended TDCs (XTDCs) that can digitally multiply the range of the main TDCs by a factor of 16.

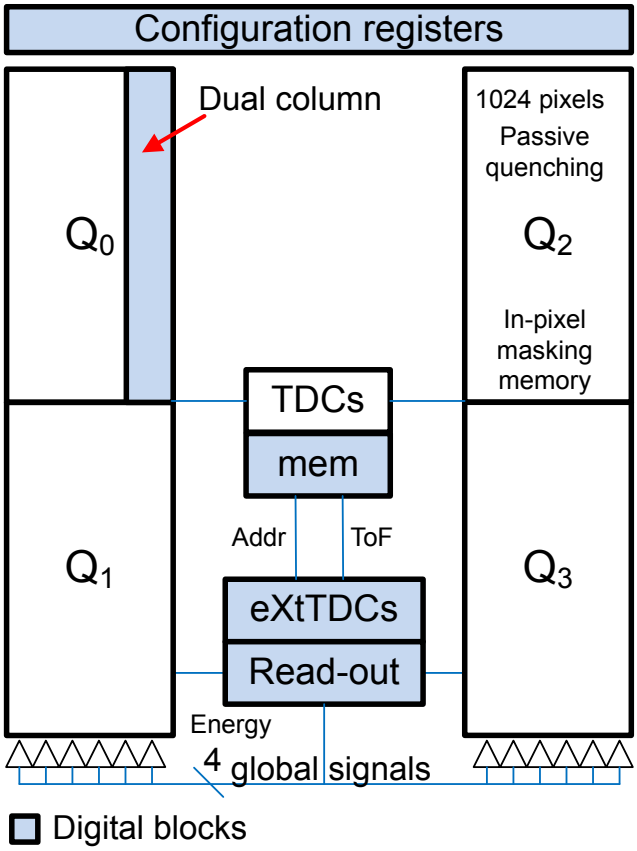


Figure 5.1: Block diagram of Panther. Four quadrants with 16 dual columns to serve 1024 pixels each.

5.3. Dual columns

The dual columns serve a total of 64 SPADs. They have quenching circuit, masking memory and a register to perform configuration and read-out. The register functions also as an in-pixel memory to store the hit of the SPAD. The timing information goes to a decision tree that propagates time and addresses along the sensor to the read-out module. Decision trees propagate the time information and at the same time the address of the pixel that originated the pulse. The dual columns have skew-free H-tree for global signals as the reset, clock and global shutter.

5.4. Digital Core

5.4.1. Architecture

Fig. 5.4 shows the block diagram of the digital core of Panther. It has a control unit that takes commands from the FPGA and redirects them to the destination module. Unlike Concolor, where the digital core performs read-out, configuration and operation, in Panther we opted to use the digital core for read-out and configuration, leaving the operation task to the FPGA to achieve more flexibility for ample variety of applications. The scheme for the read-out system is similar to that of Concolor. It comprises several interfaces which make use of address scanners and wrappers when needed. TDCs, XTDCs, Pixel Addresses, Pixel map and registers can be read by sending a command to the control unit along with the number of the information that is required. A serializer and a FIFO complete the read-out scheme, which are the interface to the FPGA.

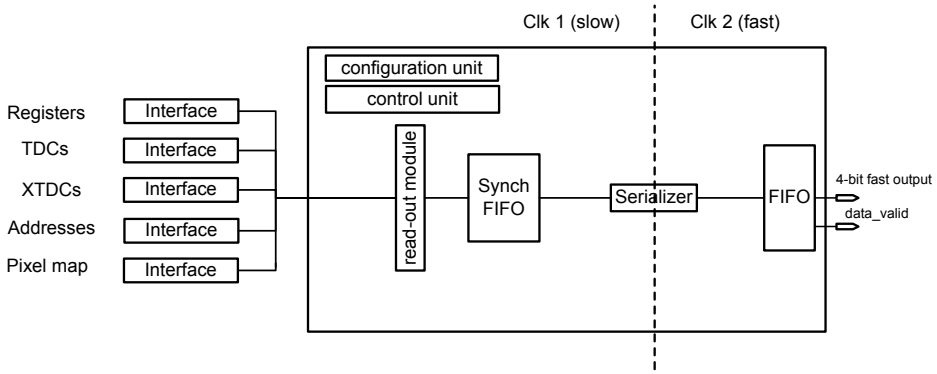


Figure 5.2: Block diagram of Panther. Four quadrants with 16 dual columns to serve 1024 pixels each.

5.4.2. Operation

Set-up stage:

Configuration of Panther: there are parameters to modify the behavior of Panther that should be set before getting Panther started.

Configuration of the pixel array: masking information is sent to the digital core which pushes the value through a circular register to the columns in a very similar fashion as Concolor.

Test: Panther has many commands to internally check registers, values, frames etc. It has fixed registers that can be read at any moment, ECHO commands to test communication, etc..

Read-out: the digital core accepts commands to read out any module of the system. This can be performed in two different ways. The read-out can be on demand, meaning the data will be sent if and only if a request is received; or, alternatively the read-out can be read-upon-frame that will set Panther to read out the data, previously specified, right after a frame has ended. This can save time and synchronization issues with the FPGA. The read-out module that takes control of the interface and gets the information, does not have direct connection with the FPGA but to a serializer that is fed by 2 different clocks that have to belong to the same domain as the serializer is not an asynchronous module. It takes the 16-bit bus and converts it into a 4-bit bus in at a higher-frequency clock and sends it to a FIFO that readable from the FPGA. The system is designed in a way that the output clock has to be at least as twice as faster than the internal clock, but since the conversion is from 16 bits to 4 bits, the system will not benefit from an output clock faster than 4 times. The single-ended PADs also limit the speed of the communication which was proven to work up to 160 MHz for 15cm traces from the chip to the FPGA.

Calculation on read-out for PET: The digital core can perform in-situ calculations as the data is gathered from the pixels and transmitted. Zeroth and first momenta of the pixel information are calculated by serial adders and multipliers; thus obtaining total energy and Anger position in X and Y axis. Fig. 5.3 shows the the circuits implemented.

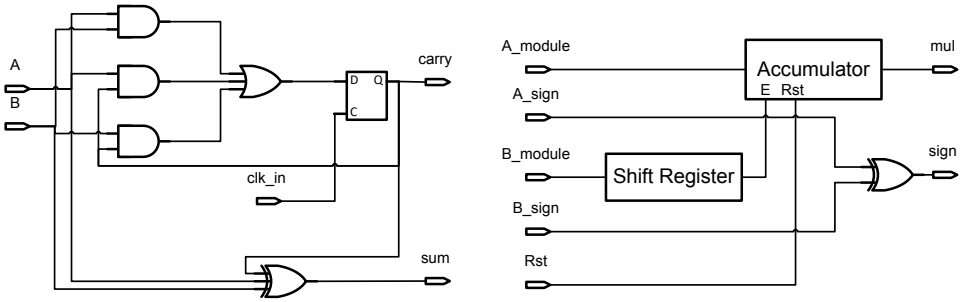


Figure 5.3: Serial adder (left). Multiplier based on shift register and accumulator (right).

These modules are used in Panther to implement the formulas to calculate the momenta m_{th} shown in Eq. 5.1.

$$m_i = \sum_{x=1}^{N} p(i)(i - m_0)^r \tag{5.1}$$

Where i is the column or the row, p is the pixel addition and r is the order of the momentum. For PET applications the zeroth ($r = 0$) momentum corresponds to energy:

$$E = \sum_{y=1}^{N}\sum_{x=1}^{N} s(x,y) \tag{5.2}$$

and the first momentum corresponds to Anger position ($r = 1$):

$$A_x = \sum_{x=1}^{N} p(x)(x - mx_0), \tag{5.3}$$

$$A_y = \sum_{x=1}^{N} p(y)(y - my_0), \tag{5.4}$$

where $p(x) = \sum_{y=1}^{N} s(x,y)$ and $p(y) = \sum_{x=1}^{N} s(x,y)$ where $s(x,y)$ is the spad/pixel information.

5.5. Time-to-Digital converters

5.5.1. eXtended Time-to-Digital Converters (XTDCs)

The main TDCs of Panther are based on ring-oscillator that get sampled when a hit arrives. The ring oscillator was designed to run at frequency with a minimum bin of 20 ps (6.3 GHz for 8 phases). The samplers have a 10-bit counter and 4 flip-flop to latch the phases from the RO. Therefore, the range of the TDCs is between 160 ns for a bin size of 20 ps and 819 ns for a bin size of 100 ps. The best time resolution of the sensor is achieved when it is operating with a bin size of 20 ps. The distance that can be measured in these conditions is $d = 20 \cdot 8 \cdot 2^{10}/2 \cdot 1ps\dot{c} = 24.5\ m$, that is not enough for LiDAR systems intended for self-driven vehicles. To increase the distance, this work introduces the eXtended TDCs (XTDCs) that can improve the range of the main TDC by using the PLL on the FPGA; thus, XTDCs can work with the accuracy of the crystal.

Architecture

The block diagram of the XTDCs architecture can be found in Fig. 5.4. The XTDCs were designed in VHDL and are part of the main digital core. The XTDCs design is similar to that of regular TDCs; it has a 4-bit Gray counter connected to 4 flip-flops to latch the counter and a FIFO to save the values for read-out. In this way, this module was included into the digital core scheme was through an address scanner that was explained in chapter 2. The hit information is buffered after the timing line not to interfere with the main TDCs.

Operation

In a frame-based system, which is commonly the case for LiDAR application, both TDCs and XTDCs get reset so that their values start in all-0 and the mode is set to '0'. When the frame starts, the system is ready to operate. Fig. 5.5 shows the waveforms when the system is active. The phases of the main TDCs run freely

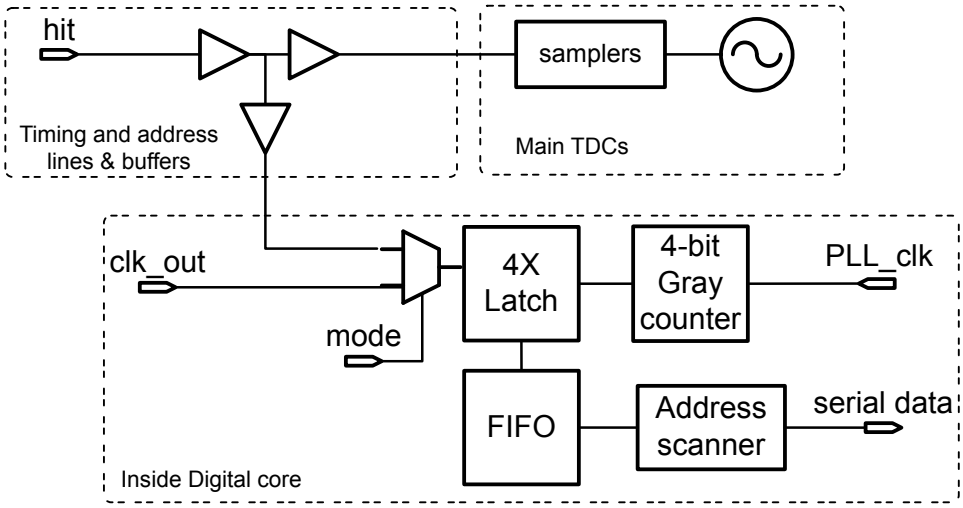


Figure 5.4: Block diagram of XTDCs

according to the local oscillator that can run as fast as 3.125 GHz. The main counter increases by 1 with every cycle of the same VCO. In this condition if an event comes, phases and counter are sampled by the logic and the maximum distance is 24.5m as stated before. The XTDCs can extend that range by using and external low clock generated by the FPGA to feeds an internal Gray counter that is sampled by the hit along the main TDCs. The clock of the TDCs and XTDCs come from different sources, thus they are uncorrelated and asynchronous. This is the reason whereby Gray counters were employed in the design. The clock of XTDCs is external and can be tuned according to the requirements of the system and the measurements to be performed. Since it has 4 bits, it can extend the range by 16 times. Longer increments in the range can be achieved by simply adding a counter in the FPGA because at this point, times are slow enough to catch the hits from the FPGA without any special care in the signals.

5.5.2. Auxiliary array supported by Time-to-Digital Converters

Panther was designed to be a very versatile sensor that can be used in LiDAR and PET as main ToF applications. However, for very simple tests and ranging measurements, the system might be too complex. This motivated to design Panther to be able to work in simplified mode with a small array of SPADs connected to a group of 8 TDCs that can be read using a simple SPI protocol that makes it suitable for simple applications. Fig. 5.6 shows the scheme of the auxiliary array with the layout in Fig. 5.7.

Architecture

Every SPAD has a masking memory and a quenching circuit that though can be programmed, they have default value after power-on reset to be ready to work.

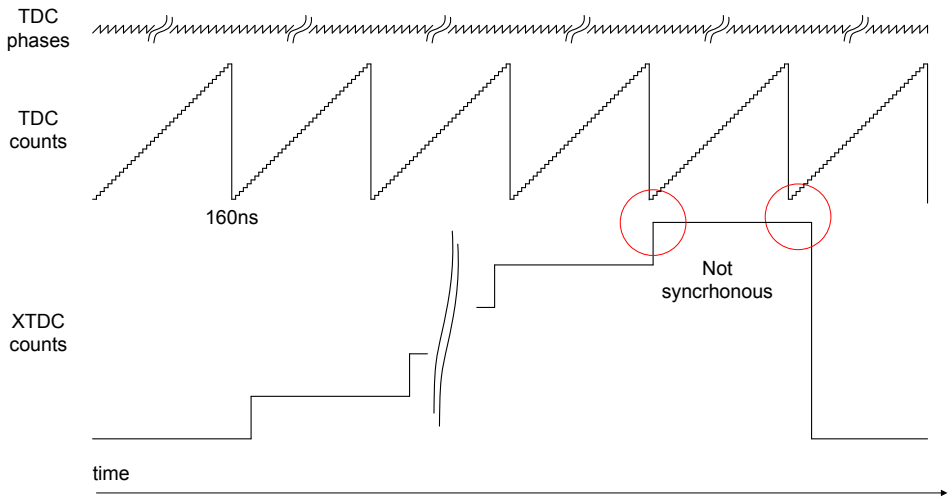


Figure 5.5: Combined operation of the main TDCs with on-chip XTDCs.

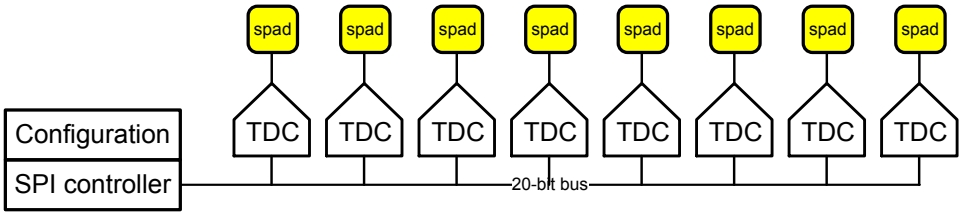


Figure 5.6: Diagram of auxiliary array of SPADs connected to TDCs.

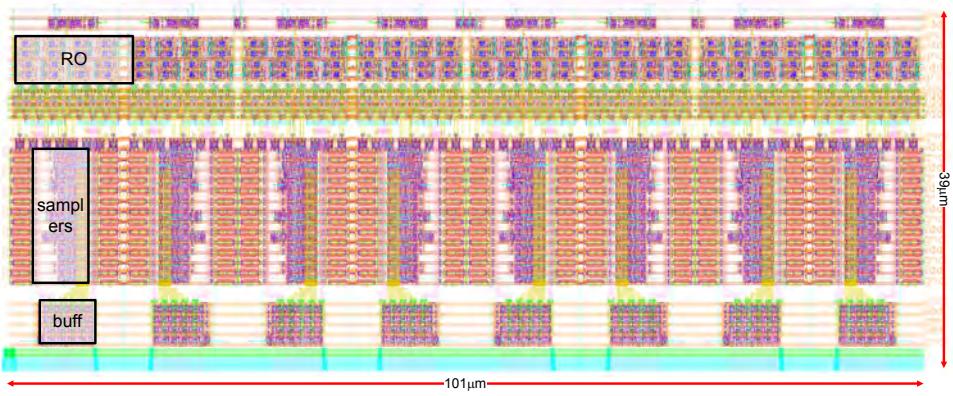


Figure 5.7: Layout of the TDCs used in the auxiliary SPAD array.

Eight TDCs are connected to the SPADs and register the time-of-arrival of the hits. The information can be read by an SPI module that has access to a 20-bit bus to read out the data by sequentially selecting the tri-state buffers from every TDC. The registered times are compared to an external signal, also used for the main array. The TDCs are very similar to the TDCs of Concolor. It has some improvements/differences that will be highlighted in the following:

- The ring oscillators have 8 phases instead of 9. Non-power-of-2 numbers affect the efficiency of the read-out. Layout is shown in Fig. 5.8
- The coupling was made by direct connections instead of using analog gates.
- Connections between stages were perfectly equalized to guarantee connections with no skew.
- Symmetry was improved so that the resulting parasitic capacitance is better compensated.

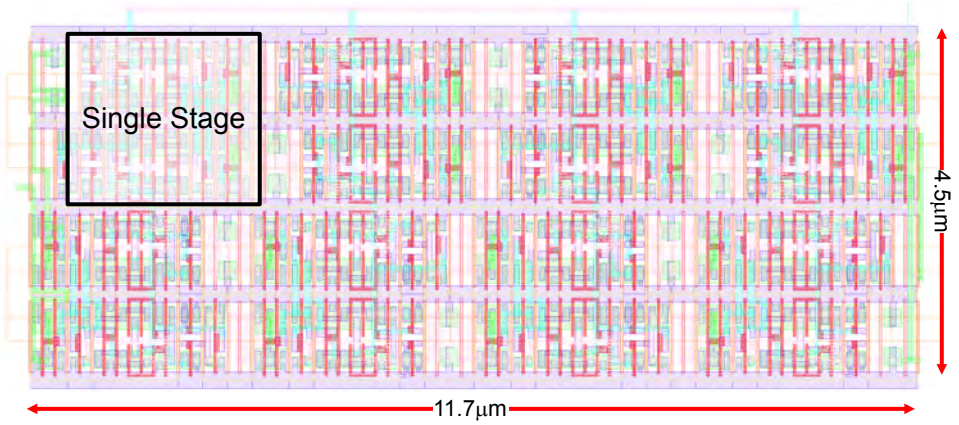


Figure 5.8: Layout of the improved Concolor-based Ring oscillators used in the auxiliary SPAD array.

Results

The auxiliary array was tested with a 405 nm laser and the results are exhibited in this subsection. The photon resolution of the system was measured, obtaining 174 ps at FWHM and 375 ps at FWTM. Fig. 5.9 shows the histogram of the acquisition. DNL was also measured in the entire range. The results are shown in Fig. 5.10. DNL for TDC 5 is -0.51 +0.34 LSB. The first 10 bins were excluded from this or any measurement because it is the time that the multi-coupled VCOs take to stabilize the oscillation; thus resulting in a time of 670 ps that although the system works, it does so out of specifications. The DNL for all the TDCs is shown in Fig. 5.11. For simplification, these TDCs do not use sliding scale technique like the TDCs of Concolor. That is the reason the DNL cannot achieve the results shown in chapter 4.

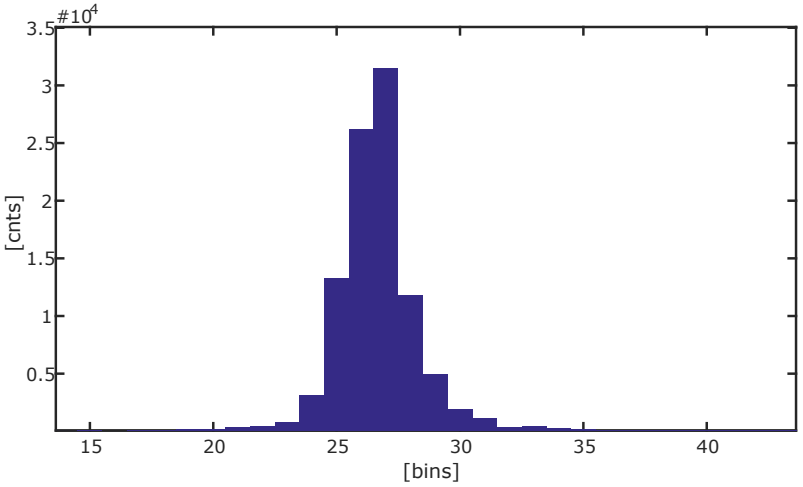


Figure 5.9: Temporal resolution of the system measured by means of laser is 174 ps at FWHM and 375 ps at FWTM. Bin size = 65 ps.

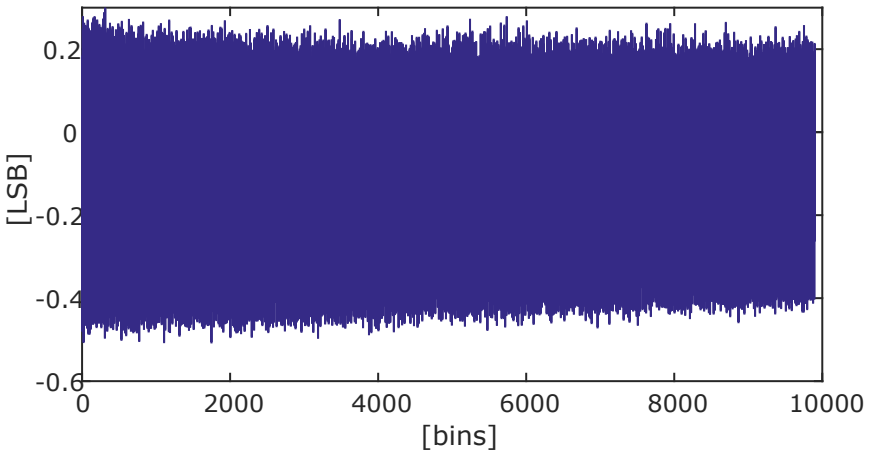


Figure 5.10: DNL through the whole range for TDC 5. Bin size (LSB) is 65ps.

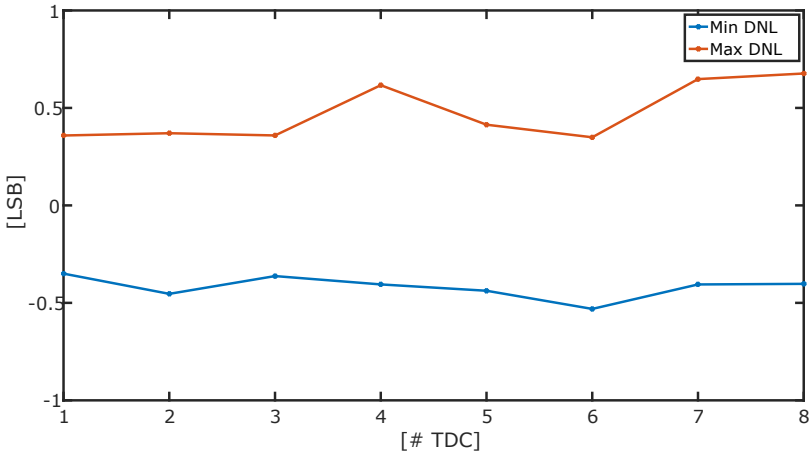


Figure 5.11: DNL for all the TDCs in the small array. Bin size (LSB) is 65ps.

5.6. High-speed Register-alike FIFO

In chapter 2, the concept of TFIFO was introduced. Its architecture and operation were fully explained and the results of the test of the memory were shown, discussed and compared with the most performing state-of-the-art memories in terms of speed and power. TFIFO was part of an idea to conceive static memories that are FIFO by design, used in image sensors. This concept led to two subtypes of True FIFO memories: pointer FIFOs where the data is static and the pointer information moves in a circular fashion, and register-alike FIFOs (RFIFO) where there is no pointing mechanism but rather the data moves as in a shift register. The first category was the one presented in chapter 2 and the second category is covered in this chapter. The idea behind the second category is that, although it takes much more power than the first category since words are moving instead of the pointer, the electronics is simpler, the words cells can be connected directly without any further modification of the structure, there is no overhead electronics (sense amplifiers, buffers, flag circuitry, etc.), and they are still advantageous for very small sizes compared to registers. The diagram of one word is shown in Fig. 5.12. Additionally, it does not require clock trees for synchronization. The classification of RFIFO is: static, FIFO, volatile, asynchronous, single domain.

It comprises a word driver and a memory element. Similarly to the TFIFO already presented, the word driver acknowledges a read-write operation that in this case there is no distinction. RFIFO reads and writes at the same time; as a result, it acts as shift registers. The word driver operates as follows: after a positive edge of clk-in, the flip-flop asserts Q and a very small ring oscillator creates a pulse that goes directly into the memory element that captures the data that is in the input D at that moment. The same pulse propagates to the 'Reset' input of the flip-flop stopping the cycle. The pulse is passed to the next word driver that repeats the operation in order to store the next word at its input. The memory element is a

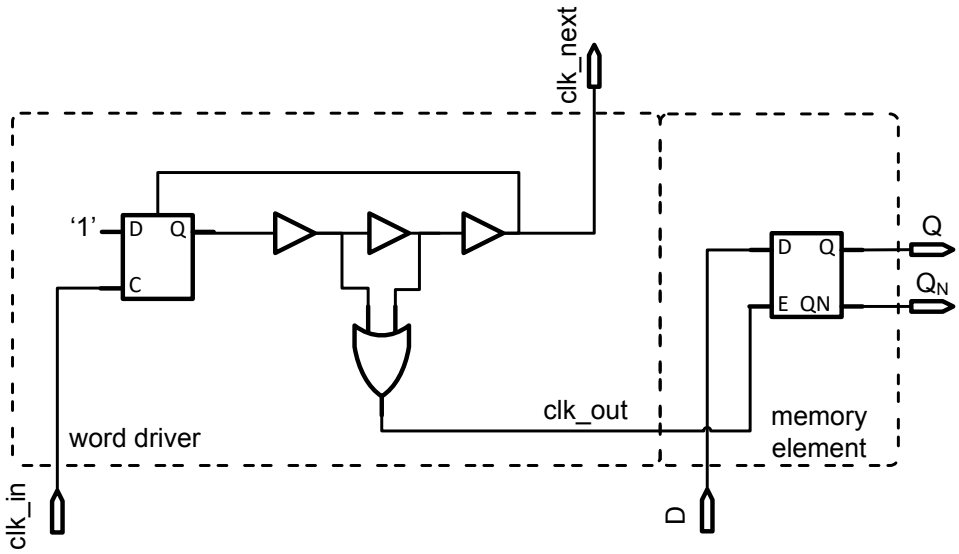


Figure 5.12: Schematics of the TDCs used in the auxiliary SPAD array.

D latch or static 6T SRAM memory cell. The layout of RFIFO is presented in 5.13. Because the clock moves backwards with respects to the data to always ensure the timing constraints, write operations have a latency of $N\Delta T$ where N is the depth of the memory and ΔT is the delay of each stage, measured around 240ps. On the other hand, reads are responsive with no lag. RFIFO can be used as programmable memory that is written by the configuration module and read at any given time to quickly extract its values, for instance, a serial weighted adder as the one used in Panther or a the multiplication modules in the neural network used in MindHive.

When RFIFO is operating 10% overvoltage, it can achieve a frequency of 4.16GHz. It was tested utilizing a data provider and a data consumer on-chip. A SPI interface controls the data transmission from and to the FPGA.

5.7. Characterization

In this section, the characterization of Panther is discussed. The results are organized in three subsections. Firstly, the sensitive matrix was characterized in terms of DCR and also the crosstalk among the SPADs was analyzed. The second part includes hardness test for chip to assess the compatibility with PET applications. At last, there are measurements for 3D imaging.

5.7.1. Dark Count Rate (DCR) and crosstalk

The sensor has two versions where different modules and types of SPAD were implemented. The first version of the sensor (Panther I) has two different Pixel set-ups. One half of the active area is composed by SPADs sized 18.6 x 18.6 μm^2 and the other half has SPADs sized 9.3 x 9.3 μm^2 that are connected by groups

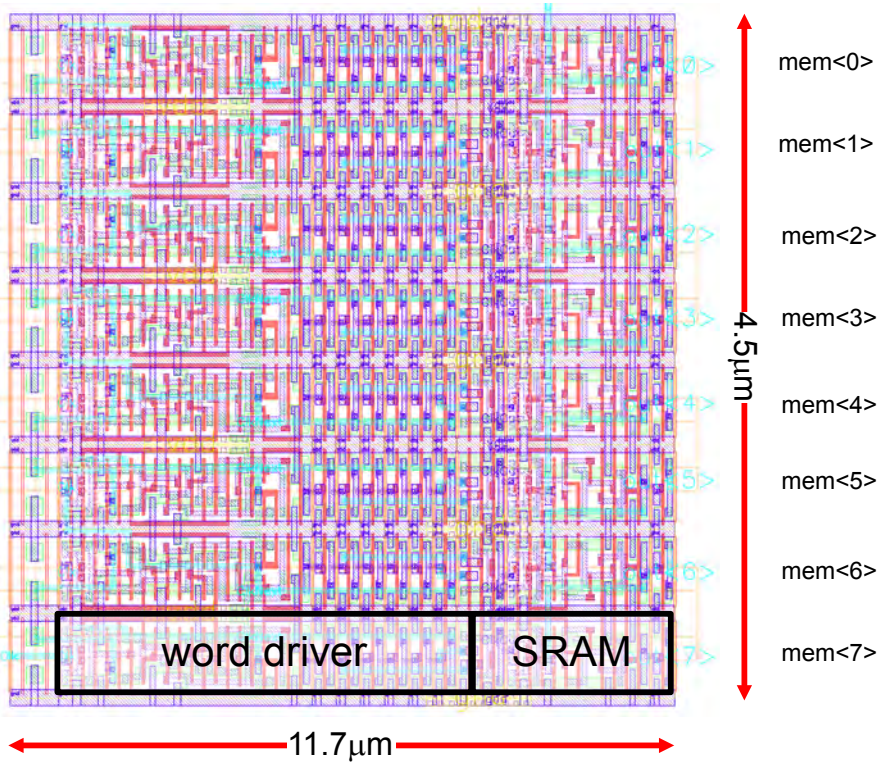


Figure 5.13: Layout of the TDCs used in the auxiliary SPAD array.

of 4 to form 1 pixel. These two types of arrangements only affect the DCR and crosstalk on the chip but it does not have any impact on the way the two versions function which remains the same. DCR population plot for both the types of SPADs is shown in Fig. 5.15. Panther II has only small SPADs covering the whole area. Fig. 5.14 has the DCR 2D map of the sensor in two versions. The 2D map on the left corresponds to all the SPADs, while in the plot on the right the 5% of the most noisy SPADs were saturated to the value of the SPAD at 95%. This is to improve the visibility of the rest of the SPADs in the plot.

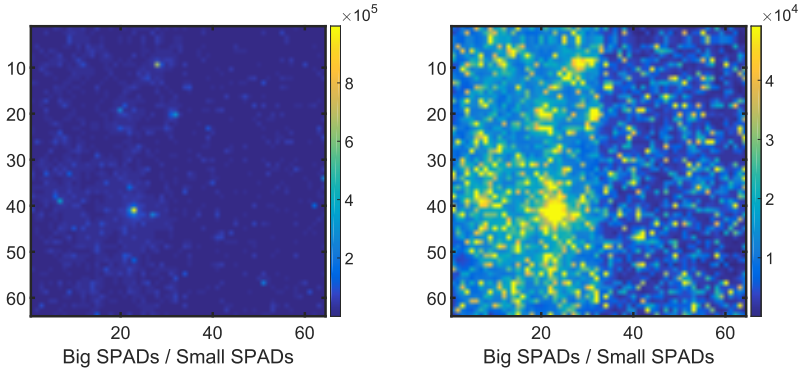


Figure 5.14: DCR of Panther, the whole population is shown for large and small SPADs for the array on the left. The same plot is shown on the right where the screamers have been saturated to the maximum value for clarity.

The reason to change the size of SPADs in the second version was the high crosstalk of the former SPADs. Fig. 5.17 shows the DCR for both the sizes of the SPADs. On top of the DCR, the average DCR of the 8 neighbor SPADs in their vicinity is marked with a red dot. In Fig. 5.16, again DCR for both the sizes of SPADs are plotted but the red dots have been randomized. The crosstalk can be seen when comparing these two plots. There is a high correlation between the vicinity DCR and the DCR of the central pixel. Unfortunately, we cannot make a quantitative analysis because there was no optical masking implemented in this chip.

5.7.2. Radiation hardness

Radiation hardness test was carried out on Panther I in order to check the behavior of the chip under radiation conditions. Whether the sensor be used for PET or for LiDAR or space applications, it is necessary to know how the chip can withstand to radiation exposure. A test was designed in collaboration with LETI in order to estimate the life span of the chip under radiation and to analyze how the performance of the chip degrades over exposure time.

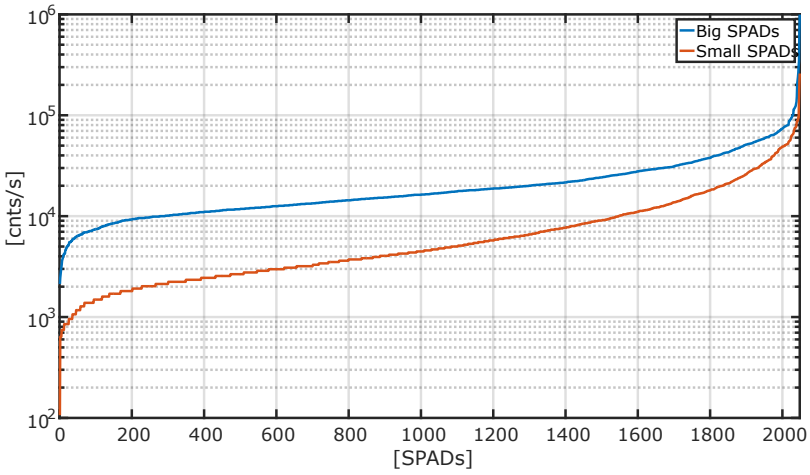


Figure 5.15: DCR of Panther, the whole population is shown for large and small SPADs sorted from the least noisy to the most noisy.

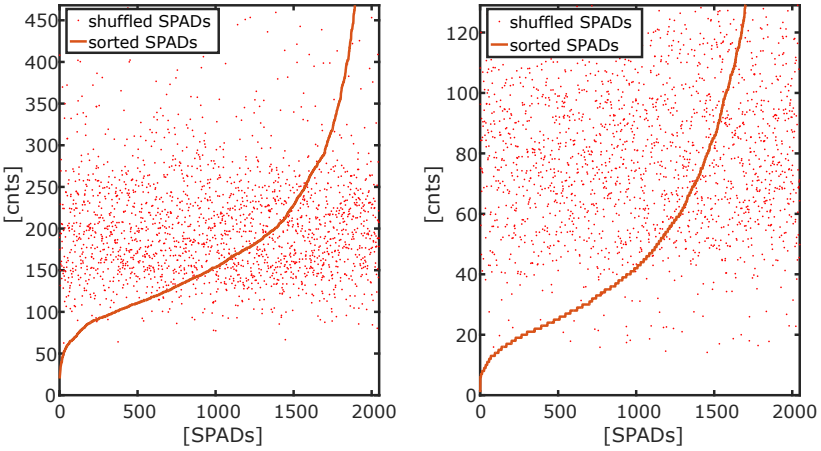


Figure 5.16: DCR of Panther for large and small SPADs. The whole population is shown, sorted by their level of DCR. The red points represent the mean DCR of the neighbour SPADs; however, these dots were randomly shuffled to show how a non-crosstalk system should look like.

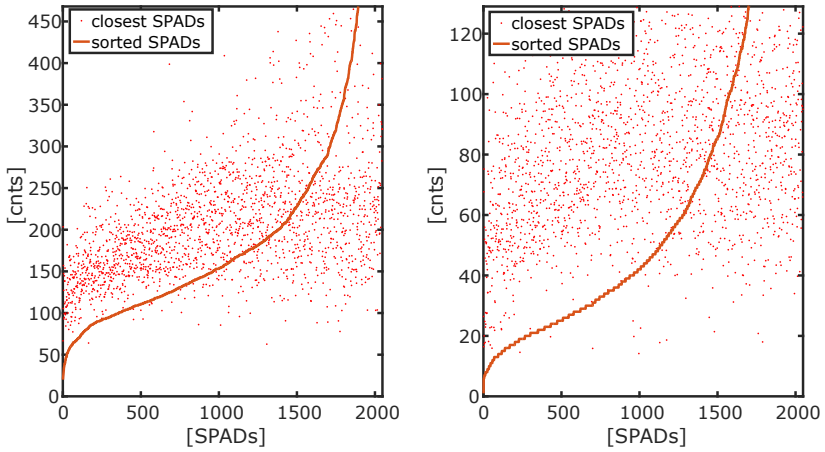


Figure 5.17: DCR of Panther for large and small SPADs. The whole population is shown, sorted by their level of DCR. The red points represent the mean DCR of the neighbour SPADs.

5

Test design The laboratory where the test was carried out counts with a powerful radiation Cobalt source of 1.17 and 1.33 MeV with a dose of 1-2 kGy/hour that can be freely moved between the chamber where the objects can be exposed to radiation and a water pool that keeps it isolated from the rest of things or people. This isolation guarantees to have a activity comparable to background from space at sea level. The radiation is expected to have a negative impact on the chip, increasing the DCR of the SPADs. Fig. 5.18 shows a picture of the place.

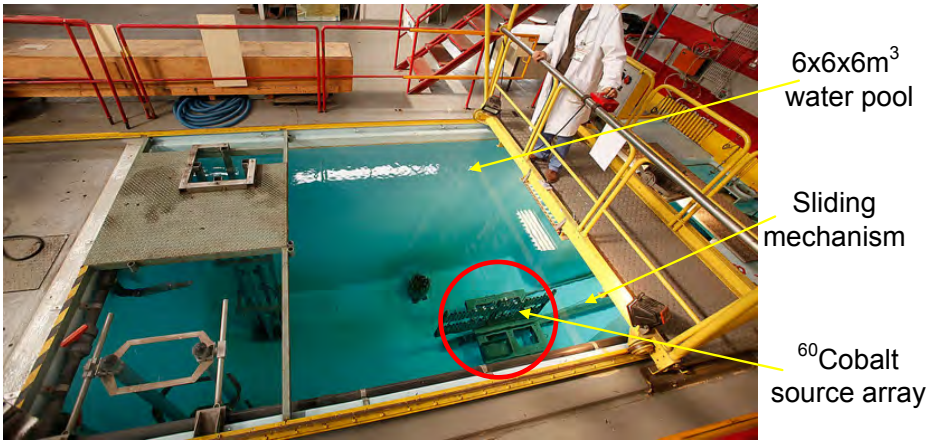


Figure 5.18: LETI facilities where the experiment was carried out.

Preparation: all the electronics, chip, power supplies are placed inside the chamber. As Fig. 5.19 shows, the chip is placed orthogonally to the location where the radiation source is going to moved to. The power supplies, FPGAs and electronics in general are conveniently covered by lead bricks to reduce the dose taken by them as the sensor should be the only electronics under test. The FPGA was connected via a long 10-m USB cable equipped with an amplifier to a laptop outside the chamber. This layout also avoids letting the laptop be exposed to the radiation. Once all the components were placed, a dry-run test was realized:

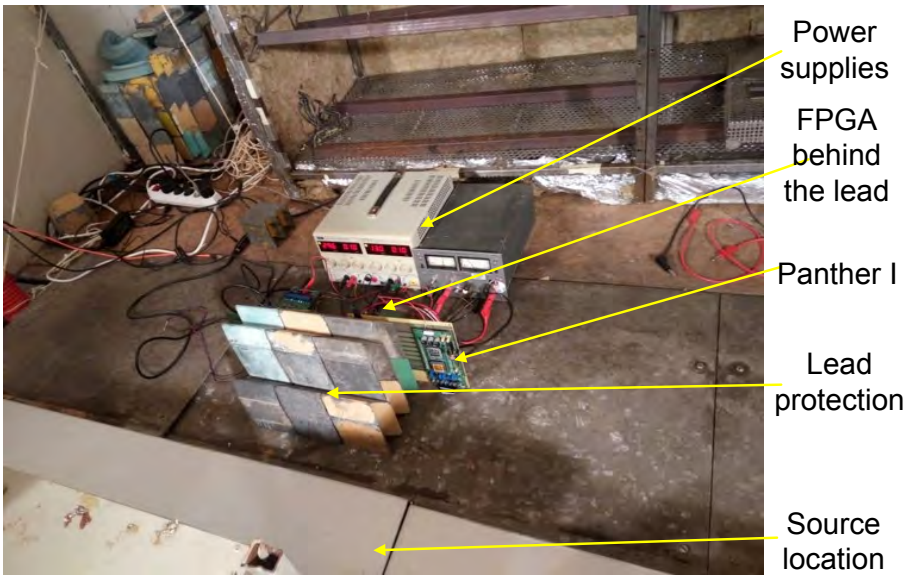


Figure 5.19: LETI facilities where the experiment was carried out.

- Lights off.
- Door closed and locked. The security officer performed the safety check.
- DCR measurements were carried out.
- Door open and lights on.

After some adjustments, the DCR measurements were fine and the chip was ready for the real test.

Radiation test: the security officer and the expert on radiation measurements made sure everything was functional, explained the procedure and security measures and kept control on the background radiation present in the lab. Next step, the test was carried out as follows:

- Lights off.
- Door closed and locked. The security officer checked the chamber was empty of people.

- DCR measurements started.
- The radiation source was moved in the chamber.
- DCR measurements stopped after 2 hours due to major failure on the electronics.
- The radiation source was moved out of the chamber.
- The security officer checked the resting position of the source.
- Door open and lights on.

Post measurements: in order to analyze and properly calculate the dose taken by the chip, it is necessary to know the activity of the radiation source at its location. The activity can be calculated using the composition and shape of the source and the geometry of the set-up. However, it is not easy to do this calculation, so the preferred method is to do a small test placing radiation reference sensors in the same location as the chips under test were placed. This allows us to get precise information about the activity in every point of interest.

Results and discussion: the total time of exposure was 137 minutes before a major failure on the electronics occurred due to the action of the irradiation on the peripheral components. Fig. 5.20 shows the DCR of the sensor over time. The DCR increases as the the chip is exposed for longer. In order to be able to analyze the data more in detail, the same DCR information is shown in Fig. 5.21 but in a different way. The DCR median value is plotted along the DCR at the first and at the last decile (each of ten equal groups a population can be divided into).

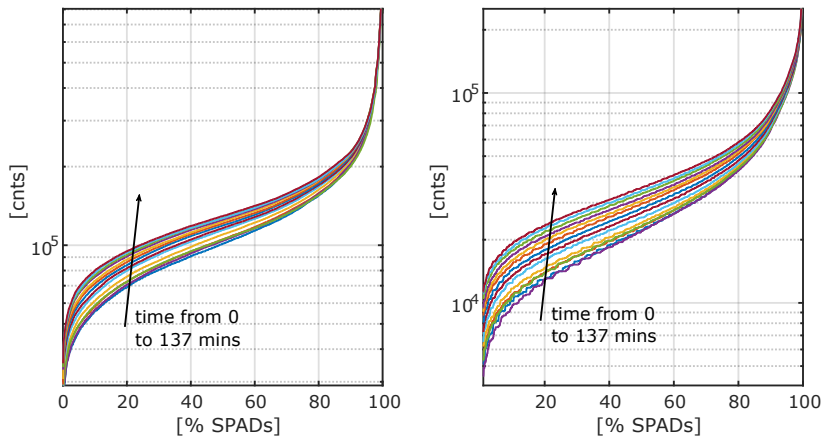


Figure 5.20: DCR of Panther for large and small SPADs during the irradiation process.

The total irradiation dose taken by the sensor was 0.2611Gy/s in a lapse of 137 minutes (which makes it suitable for PET by large amount). The DCR degradation of the large SPADs was 11.8cnts/s/Gy/A where A is the area of the SPAD; while the degradation of the small SPADs was 11.8cnts/s/Gy/A. An observation on the plots

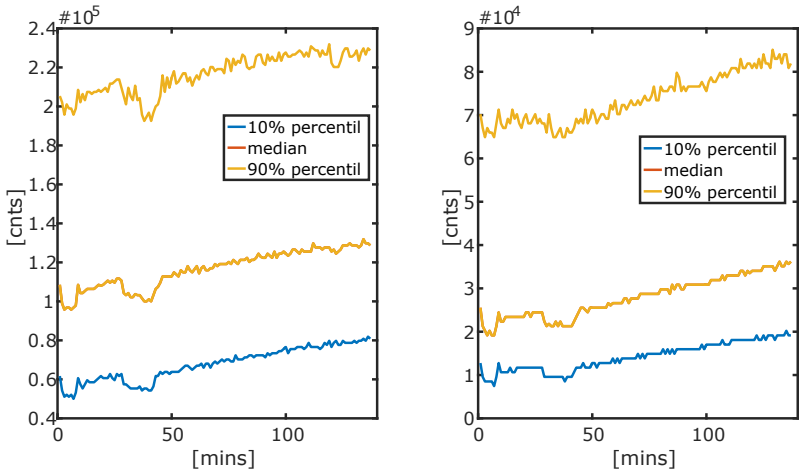


Figure 5.21: DCR of Panther for large and small SPADs on the process of irradiation.

5

is to be made for the two points where the DCR abruptly drops. Fig. 5.22 shows the mentioned effect. In this situation, due to other tests that were concurrently running wit o some handwork c ck, the DCR contin

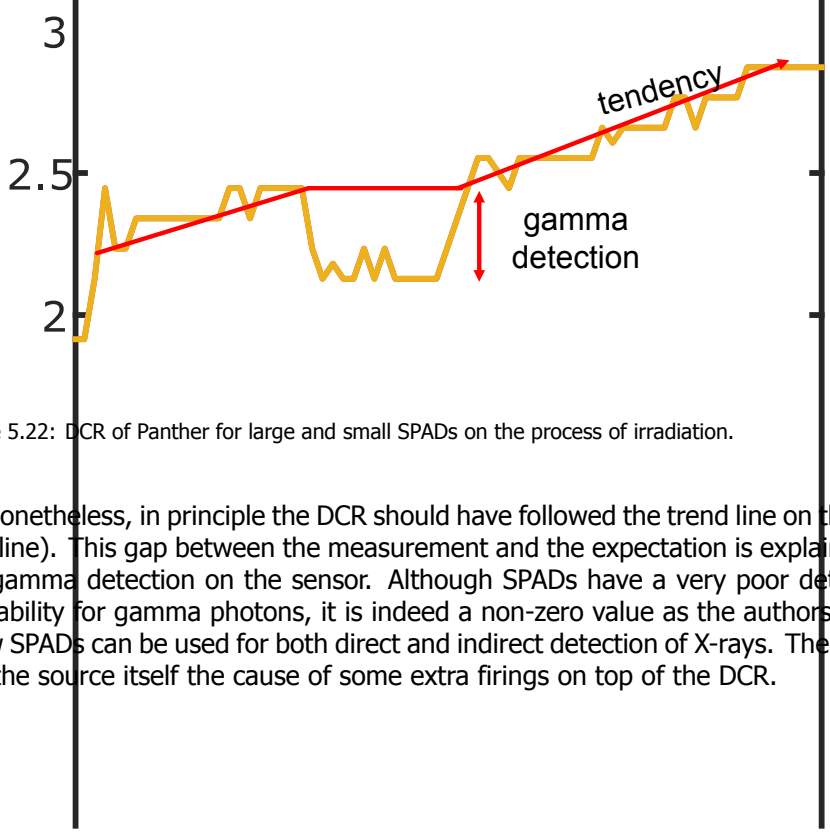
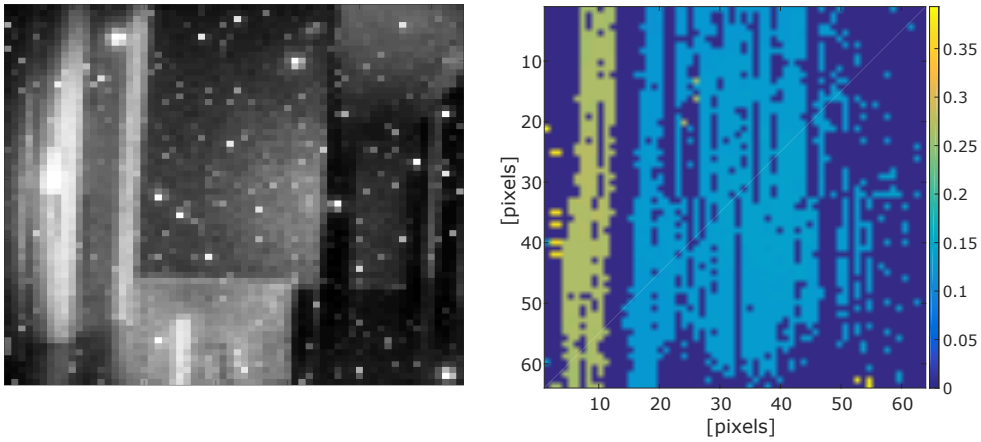


Figure 5.22: DCR of Panther for large and small SPADs on the process of irradiation.

Nonetheless, in principle the DCR should have followed the trend line on the plot (red line). This gap between the measurement and the expectation is explained by the gamma detection on the sensor. Although SPADs have a very poor detection probability for gamma photons, it is indeed a non-zero value as the authors of [4] show SPADs can be used for both direct and indirect detection of X-rays. Therefore, it is the source itself the cause of some extra firings on top of the DCR.



(a) 2D image captured by Panther II.

(b) 3D information from the TDCs.

Figure 5.23: 2D and 3D images to give precise information about the scene.

5.7.3. Results for 3D imaging

Panther was used indoors to perform LiDAR measurements. In this experiment, a 865 nm LASER was used synchronously with the system in flash mode [5]. An appropriate filter for the same wavelength was used to reduce the background noise. The filter is a band-pass filter centered at the desired wavelength with a bandwidth of 10 nm. Every frame of operation, a unique pulse was triggered from the laser. The sensor captures all th events in that frame that is sent to the PC and analyzed. TDC data and pixel map are used and combined to compute 2D image, 3D image and timing histogram that are shown in Fig. 5.23a and Fig. 5.24 respectively.

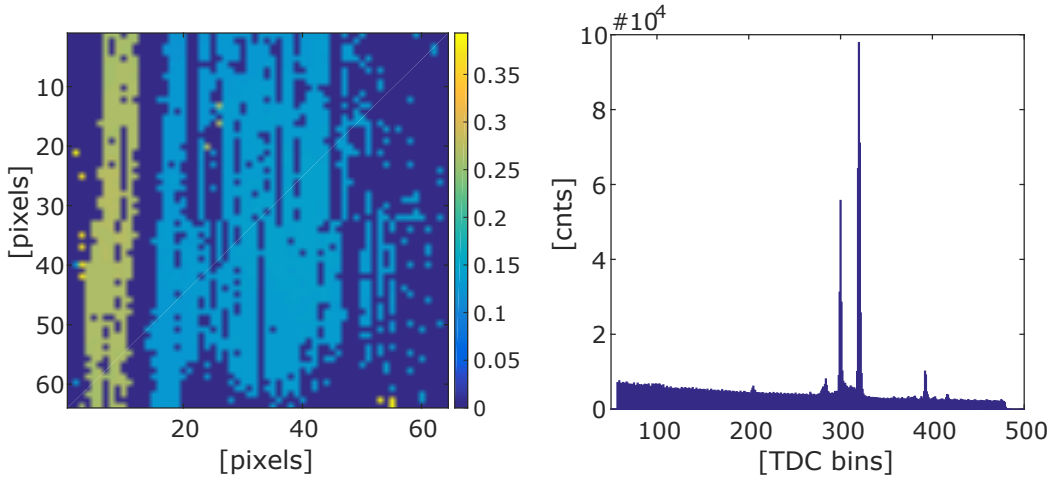


Figure 5.24: 2D histogram of the scene (left), histogram of all the TDCs (right).

5

5.8. Conclusion

A 3D integrated image sensor was presented, characterized and tested. The results for LiDAR were shown and irradiation hardness test has been performed.

The amount of dose absorbed by the sensor is shown in Eq. 5.5.

$$D = D_s \cdot T = 0.2611 Gy/s \cdot 137 \cdot 60\ s = 2146\ Gy \tag{5.5}$$

but

where D_s is the dose per unit of time and T is the time of exposure. The amount of dose during a PET test can be estimated from the equivalent dose usually absorbed by a patient that is $D_p = 5.4mSv$, [6]. This equivalent dose, in case of gamma-rays, can be translated into a dose of 5.4mGy. The dose that can be absorbed by one sensor depends on the geometry of the system; which includes the size and position of the patient, their distance to the sensor, size of the sensor and size of the scintillator that is coupled to it. This can vary for different types of systems and tests and therefore it is quite complex to estimate. However, even in the most pessimistic calculations, if the dose absorbed by only one sensor were the same as the total dose absorbed by the patient, the sensor could withstand a total number of exposures expressed by:

$$N = \frac{D_{sensor}}{D_p} = 397452 \tag{5.6}$$

This means the system has 45 years life time considering that PET tests take 1 hour. Consequently, we can assert the irradiation does not represent a problem for the sensor for PET applications. There is another aspect to discuss, which is the wavelength for PET applications. In systems where LYSO sensors are used as scintillator, the typical wavelength generated after a gamma absorption is 420 nm.

In despite of the efforts to thinning down the back side SPADs to make it more sensitive to blue spectrum, the photon detection efficiency at that wavelength is around 1% [1]. This represents a significant disadvantage as a very low PDE can be detrimental for the final coincidence time resolution of the system [7].

References

[1] T. Abbas, N. Dutton, O. Almer, S. Pellegrini, Y. Henrion, and R. Henderson, Backside illuminated SPAD image sensor with 7.83µm pitch in 3D-stacked CMOS technology (2016) pp. 8.1.1–8.1.4.

[2] A. Ronchini Ximenes, P. Padmanabhan, and E. Charbon, Mutually Coupled Time-to-Digital Converters (TDCs) for Direct Time-of-Flight (dTOF) Image Sensors, Sensors **18** (2018), 10.3390/s18103413.

[3] S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf, and E. Charbon, A High-PDE, Backside-Illuminated SPAD in 65/40-nm 3D IC CMOS Pixel With Cascoded Passive Quenching and Active Recharge, IEEE Electron Device Letters **38**, 1547 (2017).

[4] B. Dierickx, B. Dupont, A. Defernez, and N. Ahmed, Indirect X-ray photon-counting image sensor with 27T pixel and 15e−rms accurate threshold, Digest of Technical Papers - IEEE International Solid-State Circuits Conference , 114 (2011).

[5] M. Beer, O. M. Schrey, C. Nitta, W. Brockherde, B. J. Hosticka, and R. Kokozinski, 1×80 pixel SPAD-based flash LIDAR sensor with background rejection based on photon coincidence, in *2017 IEEE SENSORS* (2017) pp. 1–3.

[6] A. Kaushik, A. Jaimini, M. Tripathi, M. D'Souza, R. Sharma, A. Mondal, A. Mishra, and B. Dwarakanath, Estimation of radiation dose to patients from 18 FDG whole body PET/CT investigations using dynamic PET scan protocol, Indian J Med Res **142**, 721–31.

[7] M. Fishburn and E. Charbon, System Tradeoffs in Gamma-Ray Detection Utilizing SPAD Arrays and Scintillators, Nuclear Science, IEEE Transactions on **57**, 2549 (2010).

6

MindHive: new-generation SPAD image sensor for computer vision in TSMC 40nm technology.

Your assumptions are your windows on the world. Scrub them off every once in a while, or the light won't come in.

Isaac Asimov

This chapter presents the next natural step in image sensors. Researchers have been improving the image resolution, time resolution and properties specially designed for particular applications like PET, LiDAR, FLIM, etc.. So far, the information given by image sensors is quantitative and comprises distances, times and amount of light. The following step for image sensors is the inclusion of intelligence on-chip that can process the raw data and deliver also qualitative information in order to make the systems more versatile, robust and faster.

6.1. Introduction

This work introduced Concolor in chapter 4; the sensor is mainly focused on PET but it was also designed considering ranging applications. In chapter 5, we showed a design for LiDAR that also held features for PET as in-situ energy and position calculation and gamma event detection. Different technologies, SPADs, methods and circuits were tried and tested for these two applications covered in this work. Other sensors are specifically designed for one sole application as designs that are multi-purpose usually have to face issues that are originated from the trade-off performance needed for each of those applications. This makes the sensors hardware-dedicated units that perform well for the purpose they were designed for. The next logical step is to design an image sensor that is application independent. This reminds to the origins of microprocessors back in the 70's. At that moment for each task, there was a specific hardware solution. Later on, the idea of one unique hardware that executes several programs for different tasks came to the scene and that gave birth to microprocessors. In the same fashion, we present MindHive, an image sensor that, equipped with a matrix of 64x64 SPADs, 64 TDCs and a distributed neural network, can be "reprogrammed" (by loading the weights of the NN) at run-time to perform any task without redesigning the complete hardware. With the help of well-known high-level processing capabilities of neural networks, MindHive was designed to deliver qualitative information in micro seconds, recognizing patterns. In the future this qualitative information could include highly sophisticated useful information, such as "car in front" or "the vehicle is off the road" or "sign ahead". This chapter will show the whole design and show limited results because the sensor is still under test.

6.2. Design concept

In this section the design concept and the ideal architecture will be discussed, yet not discussing the realizable architecture and the implemented structure in the physical chip that was designed, taped out and partially tested. The main goal of this concept is to provide qualitative information generated by image streams and captured by the optical modules. This can be carried out by a neural network that receives the information from the SPAD matrix. Ideally we would like the neural network to have total access to every SPAD in a temporal sequential way such that time, hits and history be available at any time for the NNs to process. This would enable processing on not only the images captured by the sensor but also the sequence of them thus processing video streams in real time in order to deliver high-end qualitative information to the system. The kind of information we are thinking of here is "presence of danger", "accident in front", "a person is walking down the road", etc.. All this could be processed in microseconds thanks to speed boost the silicon can offer to processors and neural networks in general. The design should be part of a bigger system that gathers the raw information, trains the neural networks, downloads the weights to the sensor and starts the cycle again in order to train the NNs for specific application.

Multiple applications: as mentioned in the introduction of this chapter, we explained that the idea was system design to be used in any kind of applications. For instance, in case of Positron Emission Tomography, a neural network could be trained to calculate the first three momenta of the detected light (m_0, m_1 and m_2) to estimate energy, Anger position in X and Y axis and Depth of Interaction (DoI) estimation respectively [1]. The NNs can be trained for sign detection, weather conditions assessment, optical character recognition or anything that could be of interest.

Down to (silicon) earth: the realizable architecture for MindHive certainly cannot implement the ideal concept into CMOS technology. Limitations on technology, space and cost forced us to simplify the concept to make it possible to become silicon. Nonetheless, the primeval idea stands: MindHive processes light events coming from the SPADs in order to deliver qualitative high-end information in very short time that cannot be matched by off-chip processing solutions.

6.3. Architecture

Since at this point the reader is familiar with several architecture of image sensors, the architecture of MindHive, unlike the case of Concolor and Panther, will be explained in a bottom-top fashion. This means we start from the SPAD to the whole sensor as this will give more insights about the thinking process that materialized MindHive.

6.3.1. Cells, rows and macro-cells

The sensor has a SPAD matrix of 64x64, every octagonal SPAD with a radio of 6mm, considered the atomic cell of the hive, is connected to electronic circuitry for all the functions needed (masking, timing, hit counting, read-out, etc.). As the reader is very familiar with these circuits by now, in the next sections only the improvements from the previous chips will be explained. The pixel circuitry is shown in Fig. 6.1.

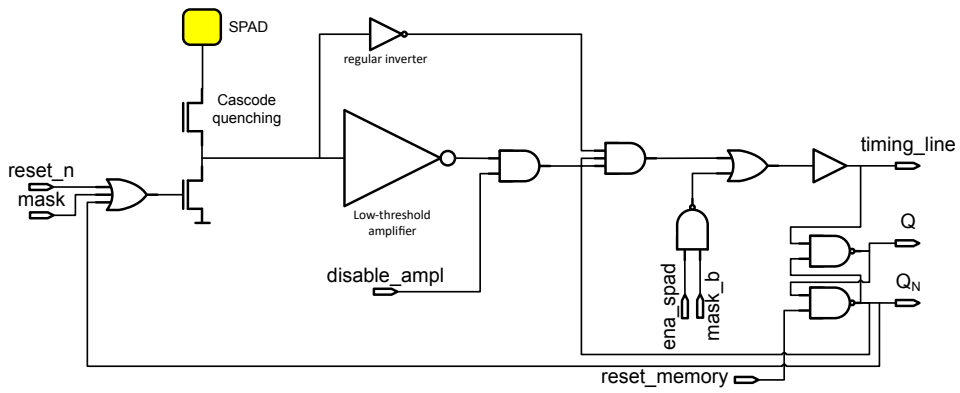


Figure 6.1: Pixel circuit of MindHive. Passive quenching, active self-reset system, hit-counting output, timing output processed by two different amplifiers.

Masking circuit, self-reset module and quenching work in the same fashion as in Concolor. The addition to the quenching circuit is another transistor cascoded with the main quenching transistor in order to increase the maximum excess bias voltage applicable to the SPAD as explained in [2]. The pulse, once is generated by the SPAD, it is detected by both the low-threshold amplifier (refer to section 3.3.1 of chapter 3) and a regular inverter. If ena_ spad is equal to '1', meaning the global shutter is on and the signal mask-b is also '1', the pulse propagates to the balanced-OR (refer to section 3.3.3 of chapter 3) timing line. At the same time, an internal latch-based memory is asserted. The output of the memory is connected to a 20-bit counter through XOR tree for hit-counting (refer to section 3.2.2 of chapter 3). The pixels are organized in rows of 8 pixels each and these rows are organized by 8 to form a macro-cell.

The rows bring the timing and hit-counting signals to the macro-cell level. For hit counting the row has an XOR-tree and the macro-cell another XOR-tree to connect the 20-bit counter. In case of the timing line, this is not as simple. If we were to connect every timing line through an OR tree to the main timing line that goes to the TDC, it would not be released until the SPAD that originally started the pulse gets reset. In many implementations there are in-pixel monostable modules that release the lines after a predefined time. However, they take considerable area, so in our design we opted to use an OR tree at row-level with 1 monostable per row and an OR tree at macro-cell level. The implication is that any SPAD that fires blinds only the rest of the SPADs of the same row that SPAD belongs to. Meanwhile, the rest of the rows are still capable of generating time-stamps. The reset can be asserted at any time since the self-reset module will ultimately determine if that reset should be attended for every particular SPAD. The monostable does not affect the hit-counting capabilities. The SPADs then do not blind any other SPAD when firing. Fig. 6.2 displays the hit-counting line from SPAD to the counter while Fig. 6.3 shows the connectivity from the SPAD to the TDC.

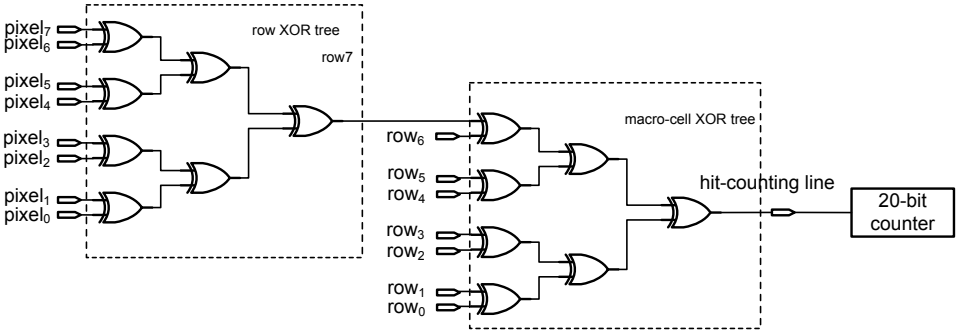


Figure 6.2: Full connection from every SPAD out of the 64 SPADs of the macrocell through the XOR tree to the 20-bit counter for hit-counting capabilities.

As described, the cells are organized in macrocells of 8x8 that are considered the minimal intelligent unit. Thus, the number of macrocells present in MindHive is 64 and they are distributed in a square shape of 8x8 as can be seen in Fig.

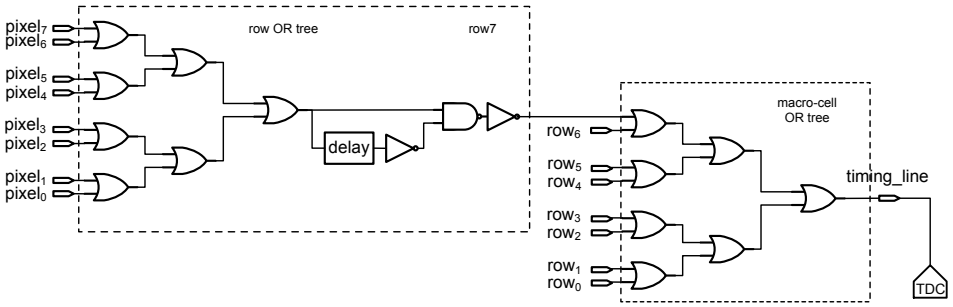


Figure 6.3: Timing connections from SPADs to the TDC that serves the macro-cell. The monostable could be at any level, the trade-off is area vs time-stamps capabilities. In this implementation the monostable is at row level.

6.4. Macrocells are capable of hit counting and timing capabilities through the OR and XOR trees (explained in section 3.2.2 of chapter 3). In this way, there are 64 TDCs in total, serving 64 macrocells. The inter-cluster skew can be removed by calibration and the intra-cluster skew is negligible, also explained in section 3.3.3 of chapter 3. The 20-bit counter has intensity information that is used in the first layer of the neural network.

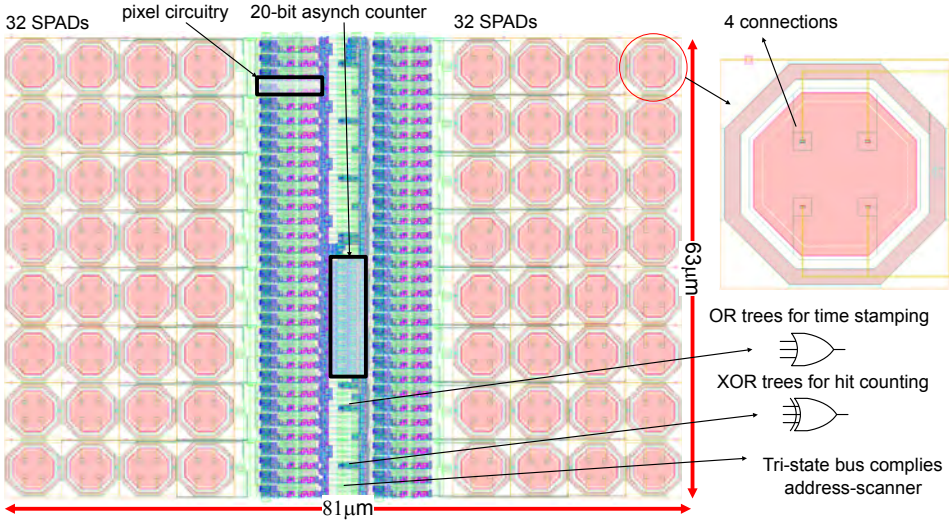


Figure 6.4: Macro-cell of the hive. The macro-cell comprises 64 SPADs, OR and XOR tress to propagate hit and time information and a 20-bit counter for intensity processing.

6.3.2. Columns

MindHive is organized in 8 columns with 8 macro-cells each. They have a mini TDC bank of 8 TDCs, a FIFO to capture events time-stamped by the TDCs and a read-out

system. Fig. 6.5 shows the layout of the columns. The counters are all connected to a bus using tri-state buffers to ease the read-out using address scanners. The NNs can also get the information in the same step. These connections do not need to be fast or de-skewed since they do not hold any timing information but intensity information. For this reason, they were implemented in parallel straight lines in thin metal layer. For timing lines, it is the opposite. Although the skew is corrected by calibration, the propagation of the signal should be as clean as possible. This was implemented in thick metal layer and surrounded by dummies. Fig. 6.6 shows these two types of connections. Masking, voltages and other slow signals were laid out using thin low metal layers.

Read-out: the time-stamps captured by the TDCs are stored in an internal synthesized FIFO of depth 4. Three out those four positions can be used by the TDCs while the last position is reserved for the end-of-frame time. In This fashion, unlike in Concolor, every TDC is responsible of getting its own reference. This avoids that the phase shift along the chip requires a complex calibration. This improvement is part of the conclusion of chapter 2. The read-out system sends the FIFO information serially to an external FPGA. Then, by means of a address scanner (see section of chapter 2 for reference), the module reads all the hit counters to harvest the information of every macro-cell in the hive. One of the differences to the other chips in this design is that the read-out system does not have individual access to every SPAD as they are not of interest. However, in order to perform the masking process, the information of every SPAD is needed. The solution is to carry out the masking in 64 steps by masking 63 SPADs of every macro-cell leaving the SPAD of interest unmasked. The hit counter will therefore count only events coming from that SPAD. The second difference is that, except for the neural network connections, every column is fully independent to capture hit events, getting the time-stamps and transmitting everything to the FPGA. Furthermore, these columns can be abutted and replicated as many times as needed thus facilitating the construction bigger arrays for other sensors.

6.3.3. Neural Network

Neural networks are a mathematical representation of knowledge that can solve very complex problems. They need to be trained as explained in the Introduction chapter of [3]. Neural networks can solve problems that were not showed to them as a part of the training set thanks to the ability of "Generalization" explained in chapter 4 of [3]. A neural network can also be defined as an universal interpolation system whose parameters are adjusted with a known set of cases and can predict the result for new cases. The implementation of the neural network of MindHive was done in four layers. The first layer has 8 inputs that process the information of every column of macro-cells, the second and third layers have 8 hidden neurons and the last layer is composed by a single output neuron. The diagram of the neural network is depicted in Fig. 6.7.

The structure of the NNs is still open for debate; it was proven that there does not exist a problem that three-layers networks could not solve that four-layers net-

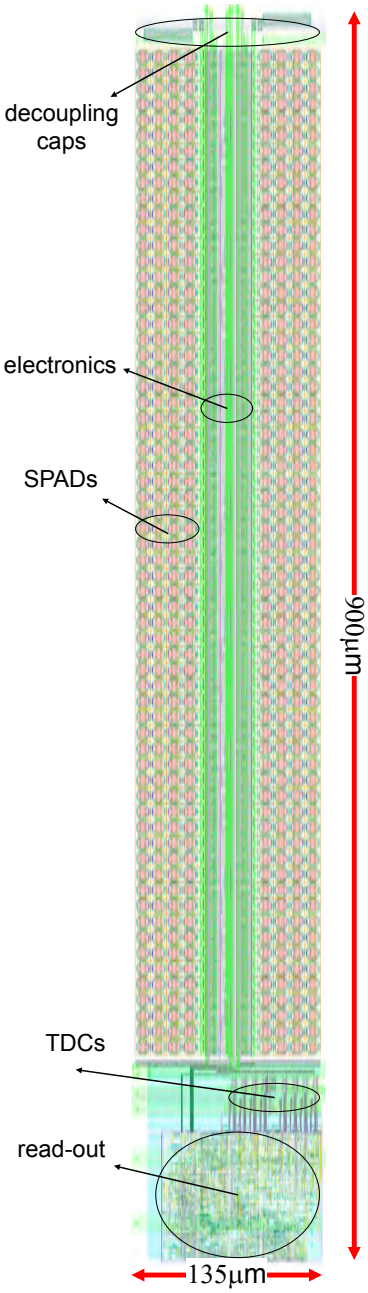


Figure 6.5: Layout of the columns. Every column hold 8 macro-cells, a mini bank of 8 TDCs and read-out electronics, and it is completely independent from the rest of the columns.

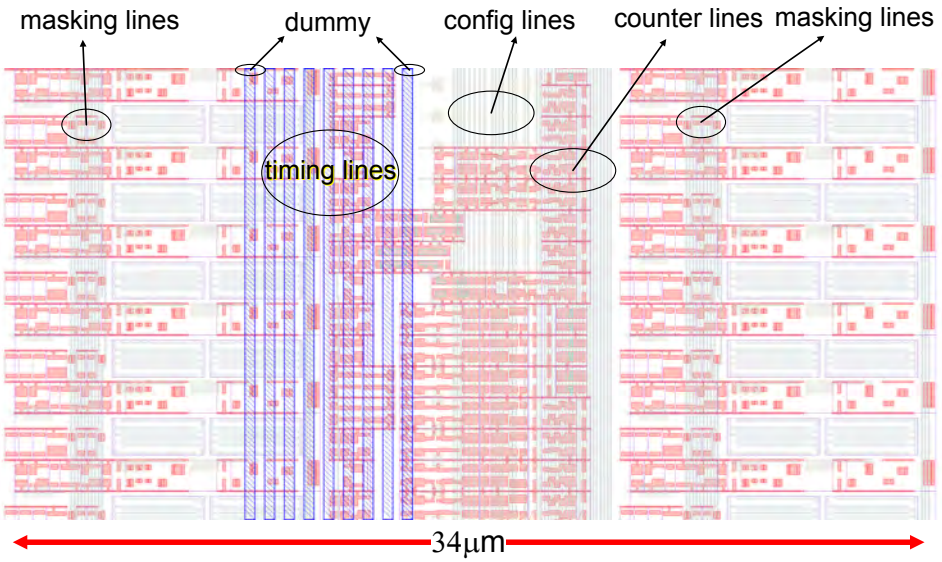


Figure 6.6: Section of the column. Different metals and widths depend on the functionality of the lines.

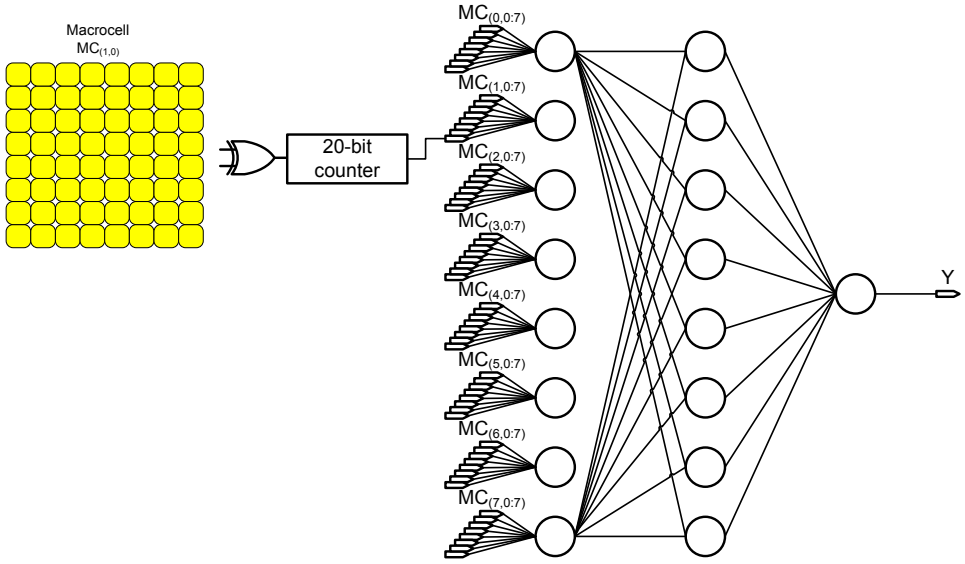


Figure 6.7: Neural network connection from the SPAD to the digital output. Intentionally, not all the connections between first hidden layer and the second hidden layer are shown for clarity.

works can. Three layers are enough to solve any problem that performs a continuous mapping from inputs to outputs [4]. Then why has a four-layers network been designed for MindHive?. The theorem uses fully-connected networks to demonstrate the conclusion. In this case, restricted by the shape factor of the chip and limitations of the technology, the network implemented cannot fulfil that rule. An extra hidden layer was added to compensate for this non-full connectivity. Hidden layers do not put up problems for the prediction phase but for the training phase. These two phases will be explained in the following sections.

Neurons:
the neurons of the network are the smallest unit of processing taking N inputs and giving one output. The equation that governs the behavior of neurons is shown in 6.2 and its graphical representation is showed in 6.8. Equations are fully explained in the introduction chapter of [3].

$$Y = f\left(\sum_{i=0}^{N} W_i X_i + b\right) \tag{6.1}$$

where X_i is the i^{th} input, W_i is i^{th} weight, f is a non-linear function known as "activation function" defined in $f : \mathbb{R} \rightarrow (-1; 1)$ and b is the bias or constant added to the neuron. The non-linearity of this function is crucial because otherwise the whole neural network reduces to a sole neuron. This can be easily demonstrated; a combination of linear equations is still linear and can be represented by a linear neuron. Thanks to the activation function, the output Y is a non-linear transformation of the outputs multiplied by variable and programmable weights. The equation 6.2 can be rewritten in matrix form as:

$$\mathbf{Y} = f\left(\mathbf{W} \times \mathbf{X} + b\right) \tag{6.2}$$

where **W** is the weight vector, **X** is the input vector and the operator (x) is the vector multiplication.

Implementation of the equation: in any physical implementation, moreover in silicon, there are limitations for implementing mathematical equations. The first compromise to make is to choose between floating-point and fixed-point representations. Floating-point operations are area and time-costly; for many implementations therefore fixed-point is preferred. For this design in particular, floating-point mathematics was discarded. The number of bits chosen for the operations will determine the accuracy of the network as anaylized in [5]; in this work it is N = 10.

Neuron circuit: the operations to be performed essentially are multiplications, additions and at the end the activation function has to be applied. There are 8 multiplications and 8 additions to be performed that can be done by approaching the problem from two different aspects. The first aspect is about the multiplicity of hardware: concurrent operation vs sequential. For example in the first case, 8

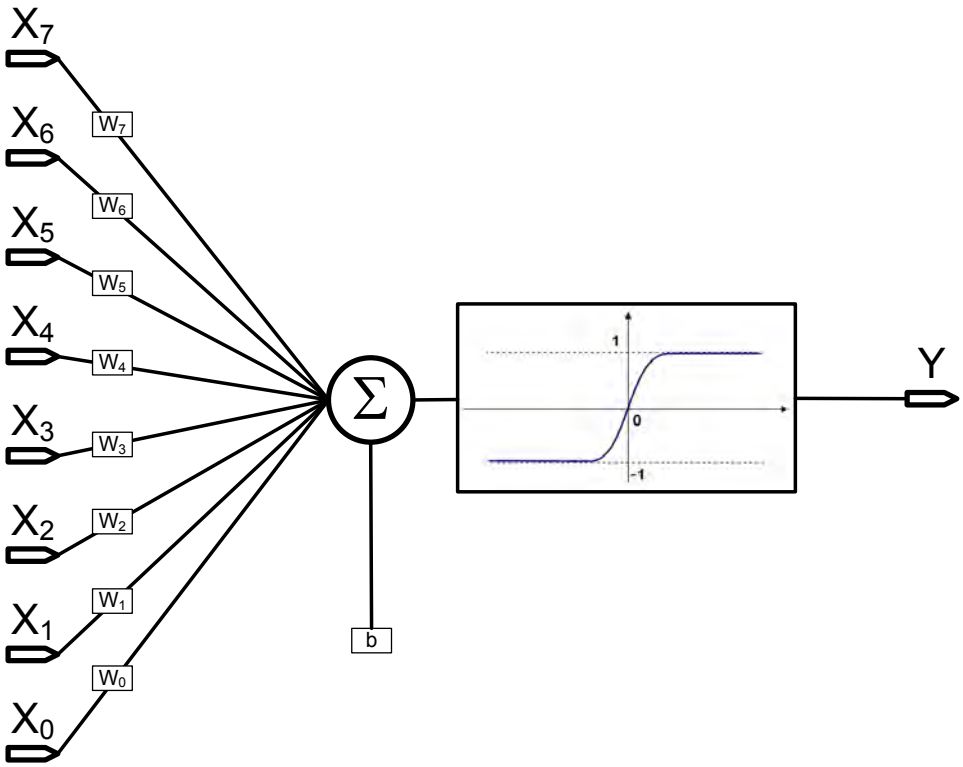


Figure 6.8: Diagram of the neuron. The neurons in the implemented net has 8 inputs, 8 multipliers with a programmable weight, an adder and a LUT for the transfer or activation function.

multiplications are done concurrently in one step while in the second case 8 steps of 1 multiplication each is performed. The obvious trade-off is area vs speed. The second aspect to evaluate is parallel vs. serial. Either multiplications and additions can be done in parallel, thus every operation will take one step, or alternatively, they can be done serially, thus every operation will take N steps, N being the number of bits of the operands. Fig. 6.9 shows two examples. The one of the right is the actual implementation of the neuron in MindHive. The area taken and number of clocks needed by the different alternatives to perform the full equation were thoroughly analyzed. The preferred architecture uses serial concurrent multiplications and sequential parallel additions. This is the best trade-off found to cope with the constraints of area and speed of operations. For M operands of N bits, the system takes $T = T_{mul} + M.T_{addition}$ steps, where T_{mul} is the time of 1 multiplication and $T_{addition}$ is the time of 1 addition which is multiplied by M since it is sequential unlike multiplications that are concurrent. Since $T_{mul} = 2N + 2$ and $T_{addition} = 1$, $T = 2N + M$. For this implementation $M = 8, N = 10$, then $T = 28$. There are 4 more clock cycles needed for synchronization and acknowledge of the internal modules , thus $T = 32$.

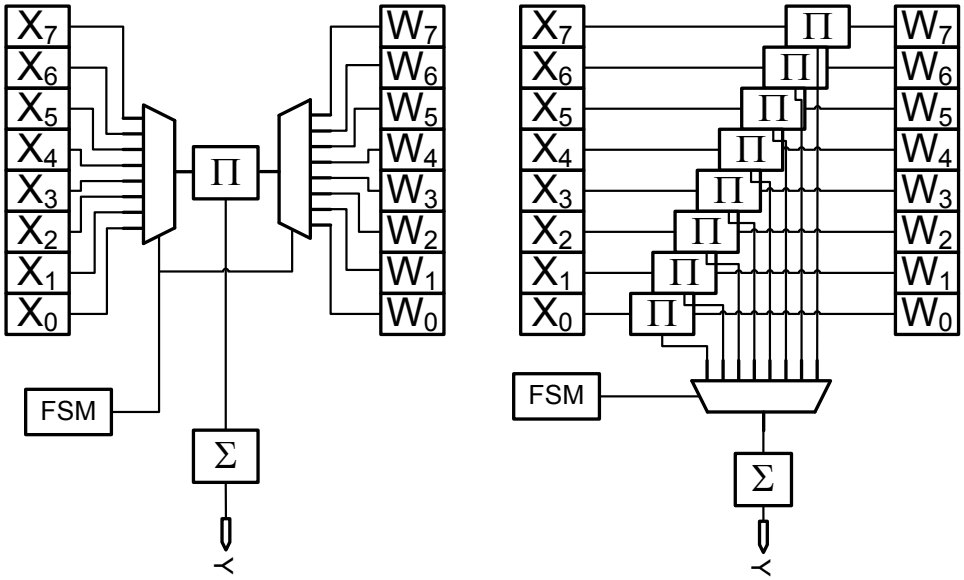


Figure 6.9: Sequential and parallel multiplication and sequential and parallel addition (left). Concurrent and serial multiplication and sequential and parallel addition.

Activation function:

there are many functions that can be used for neurons, and, in principle, any option should work provided they fulfil certain conditions: the function must be monotonic, non-linear and fully differentiable (chapter 1 of [3]). However, to easy the training algorithm and to reduce the computational cost, it is desirable that the derivative

of the function can be expressed as function of the function itself. For instance: $f'_{(x)} = Af_{(x)} + B$.

The function used in this work is the widely-known Sigmoidal function showed in equation 6.3 (chapter 1 of [3]):

$$S(x) = \frac{1}{1 + e^{-x}} \tag{6.3}$$

This function fulfills with all the requirements for an activation function. It is monotonic, differentiable and non-linear. As strange it may sound, the derivative of the this function: $\frac{d}{dx}S(x) = \frac{1}{1+e^{-x}}'$ can be written as function of $S(x)$ like in equation 6.4.

$$\frac{d}{dx}S(x) = S(x)(1 - S(x)) \tag{6.4}$$

Implementation of the activation function: in the actual implementation, it was a symmetric version of $S(x)$ to consider both positive and negative numbers. The way this has been done is $S_{sym}(x) = 2S(x) - 1$ and its derivative gets doubled. The activation function is fixed in a ROM memory to take less area than if it were possible to program it on run-time. This fixed implementation saves at least the area of the flip-flops that would be necessary to implement a small RAM. The outputs of the neurons are 9-bit length (8-bit for negative and positive numbers). A ROM with this depth for 17 neurons which is the number of neurons existent on every neural network of the implementation would make the total number of required MUXes be $N_{MUX} = U(N(2^N) - 1) = 78319$, where N is the number of bits (9) and U the number of neurons (17). In the actual implementation this number was reduced by two different methods:

a. Symmetry: the estimation was based on Two's complement number implementation. Taking advantage of the symmetry of the activation functions, if we represent the numbers in "sign and magnitude" we could re-utilize half of the table for negative numbers; thus, reducing the table to a half of its size. The sign-and-magnitude representation also makes multiplications easier. The three systems are compared in table 6.1.

Both one's complement and Sign-and-magnitude have double representation for 0. It can be a positive 0 or a negative 0. This does not represent an issue for the calculations or the fetching process in the LUT. Another advantage of sign-and-magnitude is that the conversion to and from one's complement is as simple as N inverters. As conclusion, multiplications, additions and activation function transformation can be performed with relative ease using this coding.

b. Interpolation: for small intervals, the activation function can be linearly interpolated so as to reduce the number of bits of the table. If only 5 bits are used instead of 8, the table will be shrunk from 256 positions to only 32. Therefore, the

decimal	binary	2's comp.	1's comp.	sign/mag.
+0	000	000	000	0/00
-0	000	000	111	1/00
1	001	111	001	0/01
2	010	110	010	0/10
3	011	101	011	0/11
−4	100	100	-	-
−3	101	011	100	1/11
−2	110	010	101	1/10
−1	111	001	110	1/01

Table 6.1: Table for the most popular number coding systems.

5 MSB are taken to address the ROM to evaluate the function and the 3 LSB are directly used for interpolation.

With these two simplifications, the table for the activation function has been reduced to $N_{MUX} = U(N(2^N) - 1) = 2703$ where $U = 17$ and $N = 5$. The area taken by this implementation is only 3% of the full table. Larger MUXes with 3, 4 or 6 inputs can be used but the ratio would stand. Fig. 6.10 shows the three cases described here. The version with 5-bit chopping plus linear interpolation using the input improves the accuracy where it matters (linear region) and worsens it where it does not matter much as the function has already saturated. The numbers in both the axes X and Y go from -255 to 255 but the numerical interpretation is -1 to 1 since all the operations are fixed point "0.N". As an example the following multiplication: $234_{0.N} * 105_{0.N} = 95_{0.N}$ and the addition $234_{0.N} + 105_{0.N} = 169_{0.N}$.

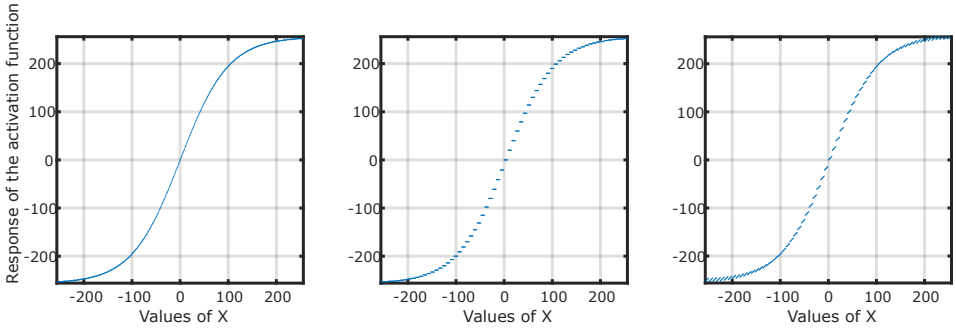


Figure 6.10: Full implementation of activation function (left), 5-bit chopping implementation (center). 5-bit chopping plus linear interpolation using the input (right).

A possible improvement is to modify the interpolation according to the range of the input that can be easily implemented by the assessment of the MSB part of the input.

Layer

Layers are groups of neurons at the same level that do not have interaction among themselves. They take inputs from the outputs of the previous layers and give outputs to the next one in the chain. A layer is called **input layer** when its inputs are the inputs of the neural network; it is called **output layer** when its outputs are the outputs of the network; and it is called **hidden layer** when its outputs are not the outputs of the network. It is said a neuronal network is fully connected when all the neurons of a layer are connected to all the neurons of the next layer. In the case of MindHive all the layers are fully connected except the first hidden layer because this would have been excessively demanding in terms of silicon real estate for space needed for the multiplications on the sensor. In this way, the macro-cells of the columns are fully connected but the inter-column connections is made at the second level of the neural network through the hidden layer. This makes the neural network performance be different in both the directions, thus making the chip non-omnidirectional and susceptible to rotation.

Phases of neural networks

In this subsection, the two different phases of neural networks will be explained. Particular details of the implemented neural network will be shown and discussed.

Training phase: the first phase is called ”training” and it is the process of adjusting all the weights and biases of a net in order to ”teach” the net how to respond to certain stimuli. There are many algorithms and methods to carry out this task. Almost all are based on using a training set of inputs and targets that the net has to accomplish. The algorithm calculates the error, that is the difference between the desired output and the actual output, and backpropagates that error through the net to adjust weights and biases. As this process continues, the error lowers and it stops when the error reached a predefined level. At this stage it is said that the network is trained.

Prediction phase: this is the stage when actually the net is used to solve a problem. New inputs, never seen by the neural network, are presented at its inputs. The net executes all its additions, multiplications and transformations and as a results gives an output. This output represents the solution to the problem presented to the net.

Evaluation of performance: the error of the network at training phase, based on MLE, is not directly related to final performance of the net defined by the cases that the neural network was able to classify correctly vs the total cases. Those two values should have a co-joint tendency (same derivative sign) when the network is fitting well. However, at some point in the training phase, the error could decrease even further but the performance will actually decrease as well, contrary to intuition. This very well-known problem is called ”over-fitting” and happens when the neural network is getting over-trained and ”learns” the particular cases instead of the

general features of them. This means that in the training phase, the performance should be checked while the neural networks is being trained.

Tests on the neural network

The neural network was entirely written in VHDL code, this includes adders, multipliers, LUTs for the activation function, FSMs, etc.. The verification process involved behavioral simulations carried out in ModelSim. Then the VHDL was synthesized and placed and routed using Innovus. The layout was checked in Cadence tools. The layout extraction information was included in ModelSim in order to do account for parasitics in a full post-layout simulation. In this type of simulations it is not possible to simulate and check the modules individually as the resulting file after all the synthesis process is a compound of low-level gates and RC delays. However, the overall results can be checked. The neural network was trained to perform simple additions and XOR operations to validate functionality. The results were compared for behavioral and post-layout simulations.

The list of entities written in VHDL code is shown in Fig.6.11. Layers do not have an entity by themselves. The neural network top level creates the layers by instantiating and connecting the neurons.

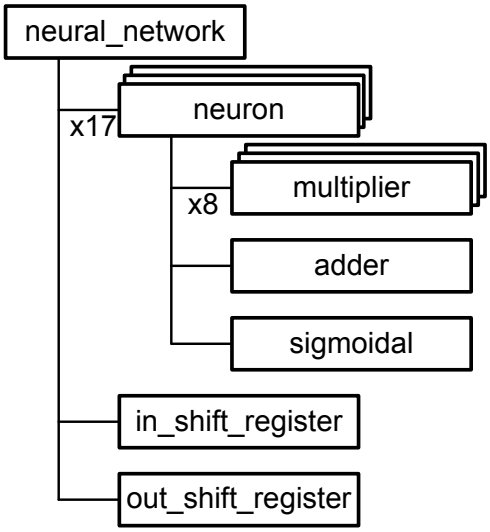


Figure 6.11: Structure of VHDL code, each rectangle represents an entity in VHDL. The "xN" multiplier in front means the number of times that module needs to be instantiated.

Performance of the net as function of the number of bits: the digitization of the net is the number of bits used as fixed-point operations in the modules to represent the magnitudes. The higher this number, the better the net will perform [6]. On the other hand, the higher this number is, the more area the components will occupy in the silicon; running therefore into a difficult trade-off. This analysis leads to the question: How many bits are enough to carry out all the multiplications

and additions required by the net for a given problem?. This a-priori easy question has no easy answer. Since neural networks are highly non-linear and quantized neural networks have an additional constraint, it is very difficult to tackle the problem in an analytical fashion. Furthermore, the problem to solve has direct impact in the topology and complexity that should be chosen for the net. In this section we analyze the particular implemented topology with a particular problem to see this effect.

In this exercise, the neural network was used to digitize a continuous sinusoidal signal with a 20% of noise into 4 levels. Fig. 6.12 shows the input and the output of the net in production phase.

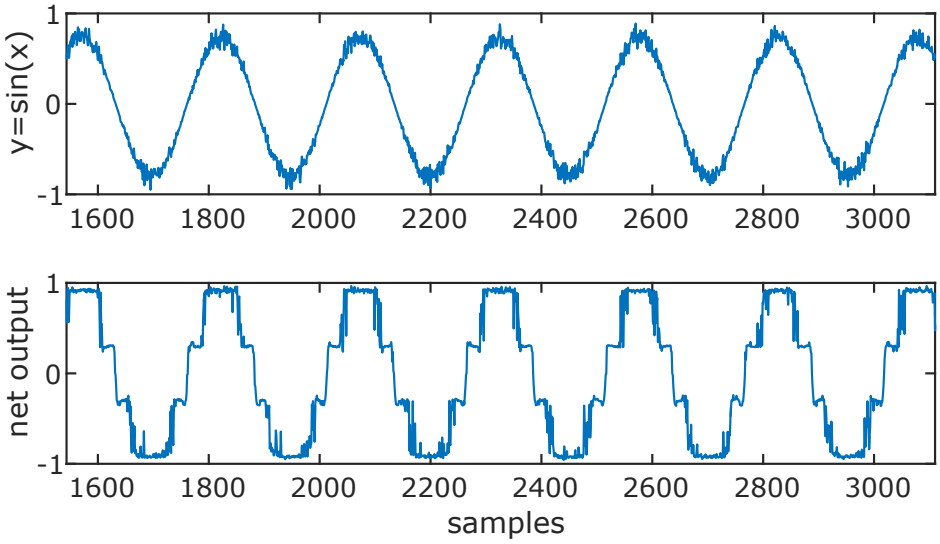


Figure 6.12: Sinusoidal signal with 20% noise (top). Output of the net after the training phase (bottom). The net learnt to digitize the sinusoidal into 4 voltage levels.

The output of the net is shown for the network that has 10 bits resolution. The neural network was trained with 10k samples and tested with 10k new samples. The performance P is measured as $P = \frac{h}{h+m}$, where h is the number of hits and m is the number of misses. Once the neural network has been trained, its number of bits was swept from 1 to 10 to see how that performance is affected. Multiplication and addition modules are reduced at the number of bits specified and the transfer function is reduced as well, as it can be seen in Fig. 6.13.

The performance was measured for each case according to the bit reduction of the neural network. The results are shown in Fig. 6.14.

The result is very interesting to analyze. It is observed that the performance is poor (below 50%) for very low number of bits and it is high (around 85%) for high number of bits; it is however not necessary to use the full 10 bits of the net to achieve a performance of 86%. In fact, only 5 bits achieve almost the same

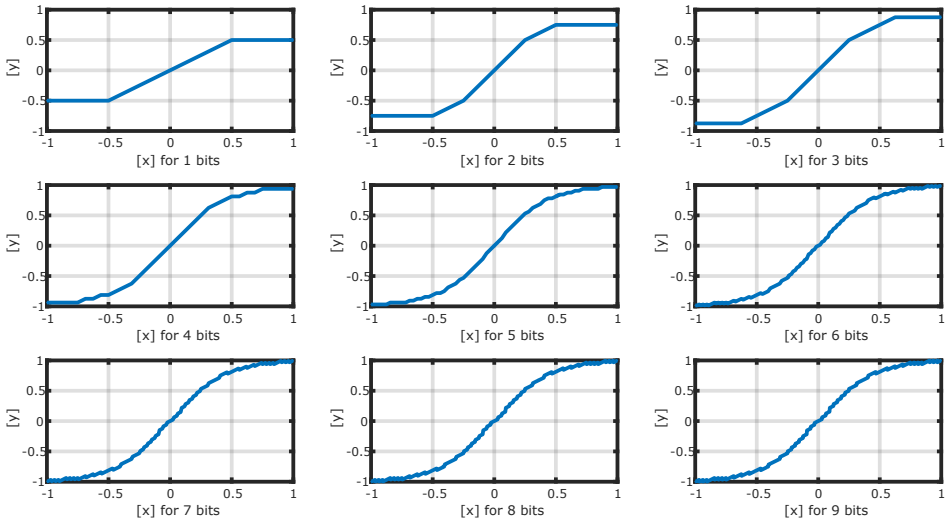


Figure 6.13: Transfer function utilizing different number of bit for the representation. The actual LUT has 5 bits for the function and 5 bits for the interpolation.

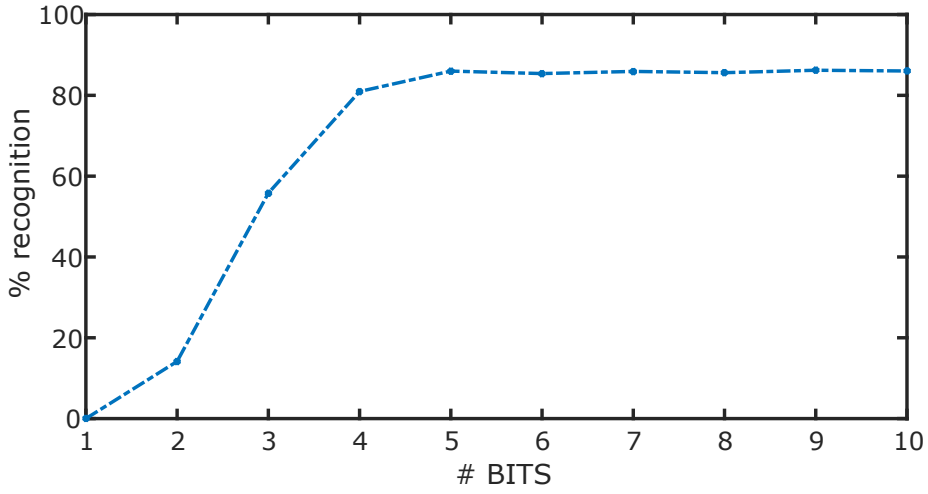


Figure 6.14: Performance or percentage of recognition vs number of bits used in the net. The number of bits affects the adders, the multipliers and the transfer function.

performance for this particular problem. In this case, the implemented network has 10 bits because the problems that it needs to solve are completely unknown as this was part of the goal.

A real case in the implemented neural network: one of the most likely problems to be solved by this type of networks in artificial vision is Optical Character Recognition (OCR). In several systems as LiDAR and self-driving cars, reading signs on the street can be of particular interest. In this section we show how the implemented networks in silicon can solve this problem taking the information from the optical sensor.

Training: the network was trained with standard image database for characters called NIST [7]. In this example they were trained using the number database. The training set has 60k images that are divided equally among the 10 digits from 0 to 9, thus totalizing about 6000 images per digit. These 60k were divided in two subsets of 50k and 10k. The first set was used for weights adjustment while the second set was used to measure the performance of the net simultaneously with the training to assess the generalization ability of the net. The images has 28x28 bits that were compressed into 7x7 to feed the network. The actual size of the inputs of the net is 8x8 but a synthesis problem made the biases of the network not be available. At the end of this section, it is explained how this problem was solved by using 1 column and 1 row of the matrix. The transformmation process is shown in Fig. 6.15 where the border of the pictures are highlighted to distinguish the rank and column used for the biases. Fig. 6.16 shows how the training set-up looks like.

Production for evaluation: NIST database has also a set of 10.000 images to test trained networks. This set was introduced to the neural network achieving a very high average performance of 96.47 %. The neural network implemented, as said before, is not omnidirectional; thus meaning that the performance might be different if the digits are rotated by 90 degrees. In Table 6.2 the performance for the particular classes is shown.

Signal to Noise Ratio (SNR): besides the recognition ability of neural networks, there is another strong feature that has to be mentioned which is the tolerance to noise or robustness. The nature of neural networks is such that it enables the system to recognize and solve problems whose inputs are difficult to define. For instance, in this case, the handwriting numbers might differ from person to person. In the same fashion, when the inputs are stained by random noise, in this case DCR or background light, the nets are still able to perform their task. In this experiment, the nets were fed with not only the input but also with random-noise image with a certain level of SNR. The noise was chosen to be Poisson as it is the most common case for these types of sensors. Fig. 6.17 shows the performance or percentage of recognition as function of SNR. No screamers were defined in this exercise.

Figure 6.15: Images from the database NIST are 28x28 pixels. They were transformed into 7x7 images just by adding the light intensity that should be measured by the sensor. The last rank and column are set to 1 to implement the bias.

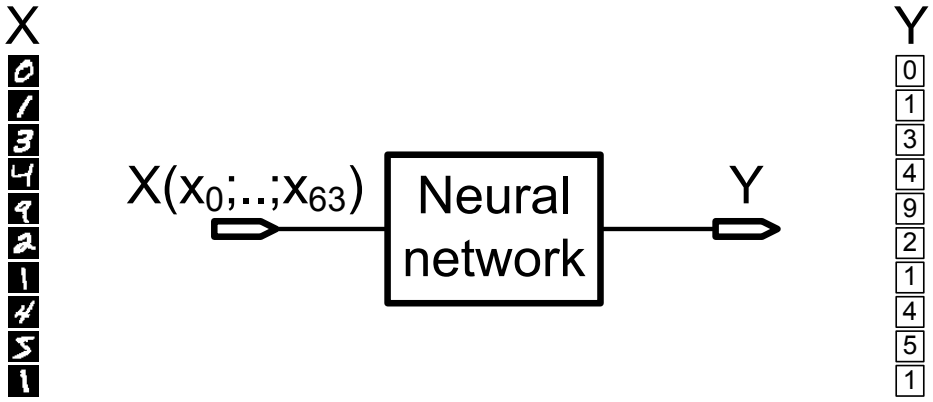


Figure 6.16: Set-up for the training set. All the images from the training set are shown one by one to the network for the algorithm to train it. This takes several iterations where the images are randomly permuted not to bias the net onwards a particular image order.

Digit	Performance [%]
0	98.57
1	97.27
2	96.99
3	97.31
4	96.38
5	95.69
6	96.95
7	96.68
8	92.49
9	96.32

Table 6.2: Performance of the net per digit, given by recognized cases over false-classified cases.

To analyze the results, it is better to split the plot into three regions. For the first region (SNR > 10), it is to notice that the overall performance remains at a level of 96%, hence the network stays unaffected by noise. For values of SNR between 2 and 10 the performance lowers until it reaches 85% when $SNR = 2$. As the SNR continues to lower ($SNR < 2$), the overall performance lowers until a value of 68% when the $SNR = 0.1$. However, it should be noticed that although the overall performance remains at good levels, shorter-distance classes are affected more than larger-distance inputs. This is due to the nature of noise that makes it evenly distributed. For classes with smaller separation distance, the deterioration is much faster. The classes with larger separation distance include the numbers $\{1; 2; 4; 5; 6; 7; 8\}$. The classes with smaller separation distance are the numbers $\{0; 3; 9\}$. A purple line is drawn vertically to mark at SNR = 1.

Biases' compensation

As said in a previous section, unfortunately the biases of the network are not available to include in the equation due to a problem in the synthesis. Biases are very important as they prevent the neurons (as classifiers) from being centered in the origin of the N-dimensional space. In order to overcome this problem, we should remember the equation of a single neuron 6.2. If the equation is decomposed for each input and the bias is replaced by 0, we obtain:

$$f\left(W_0X_0 + W_1X_1 + ... + W_7X_7 + 0\right) \tag{6.5}$$

The effect of the bias, which is not other than a constant added to the main function, can be replaced by a constant input to the net. In the test, the 8th rank was used for this purpose. The equation can be rewritten as follows.

$$f\left(W_0X_0 + W_1X_1 + ... + W_7 1 + 0\right) \tag{6.6}$$

$$Y = f\left(\sum_{i=0}^{6} W_iX_i + W_7\right) = f\left(\sum_{i=0}^{7} W_iX_i + b\right) \tag{6.7}$$

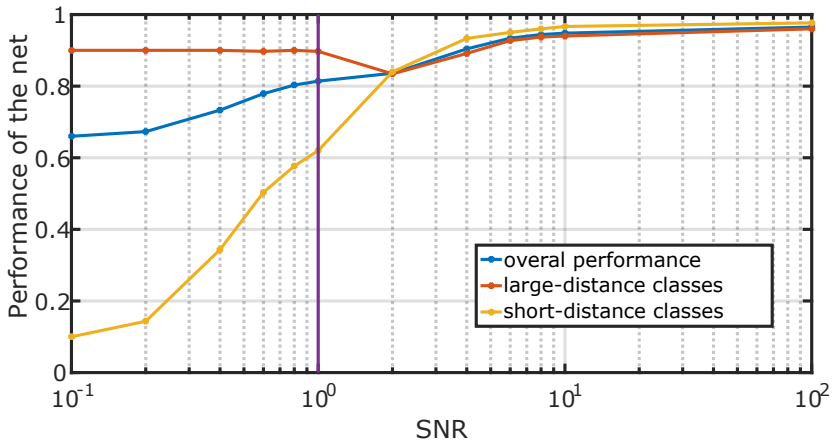


Figure 6.17: Performance of the net as a function of SNR. The blue curve is the performance considering all the characters. The red curve is the performance considering only the groups of classes with large distance, and the yellow curve shows the performance for classes with short distance. The performance of a net is measured as the ratio between hits and fails in the recognition.

Therefore, the same form of the full equation of a neuron is obtained by fixing the value of one of its inputs to 1. For this, one macro-cell has to be disabled and be used as a constant. This should be done for each neuron so the full 8th rank has to be used to create the bias. In this way the problem was solved only for the first hidden layer and the problem still remains for the second hidden layer. To provide a bias to the second hidden layer it is used the same concept: one input should be fixed to 1. However, the input of the second hidden layer is not available from the input as every input of the second hidden layer is an output of the first hidden layer. If the output of one of the neurons of the first hidden layer is constant, the problem will be solved. The way to make the output of a neuron constant is to make all its inputs constant. Then:

$$Y = f\left(\sum_{i=0}^{7} W_i\right) \tag{6.8}$$

The training algorithm can set all the weights as if they were the bias of the second hidden layer. For this, a complete column of macro-cells has to be set to 1. In this fashion the problem is solved and the consequence is the reduction of the matrix from 8x8 to 7x7. It should be noticed that any rank or any column of the matrix will do the trick. It was chosen to be the last rank and the last column to keep the sensitive area all together; thus forming a continuous square sensor.

6.3.4. Second-generation Sliding-Scale on-the-fly retriggerable Time-to-Digital Converters with 8.6ps interpolation and time of conversion of 700ps.

As by now the reader is very knowledgeable about TDCs, this section is focused on the aspects that are improved from the previous versions of TDCs. The architecture is shown in Fig. 6.18 and the layout is presented in Fig. 6.19. There is one source of oscillation that is a fully balanced and compensated 3-stages ring oscillator whose phases are distributed along the chip. The positive and negative versions of the phases are compared to a voltage level to generate 12 divisions of the VCO cycle. Then the positive phases are compared with their counterparts to generate 3 more divisions per half cycle, therefore achieving 18 sub divisions of the main VCO cycle. The Ring oscillator, shown in Fig. 6.20 has a maximum frequency of 6.25 GHz, thus the TDCs can achieve a bin size of 8.6 ps. Layout symmetry and current direction to keep matching were pursued in the design, shown in Fig. 6.21. The reference voltages need calibration to lower DNL. The comparisons that do not need calibration are the one made between positive and negative phases. The cycle is in this way split into 6 divisions and the bin size of TDC is 30ps. The reference voltages can be ignored by setting them to ground if a bin size of 30ps is enough for the application.

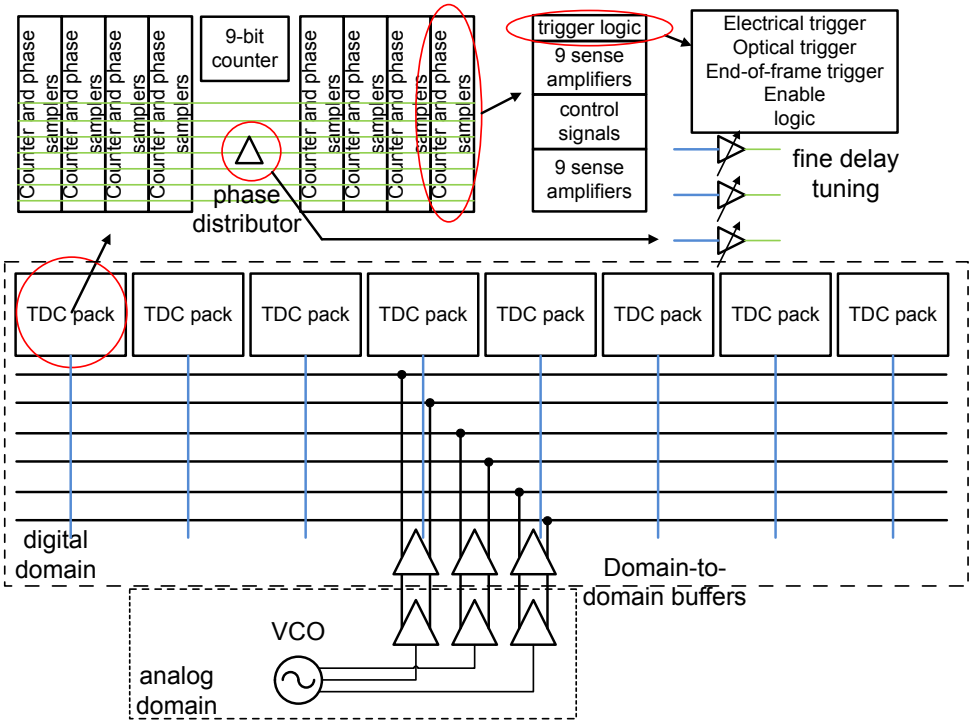


Figure 6.18: TDCs bank of the hive. The phases are distributed horizontally along the columns.

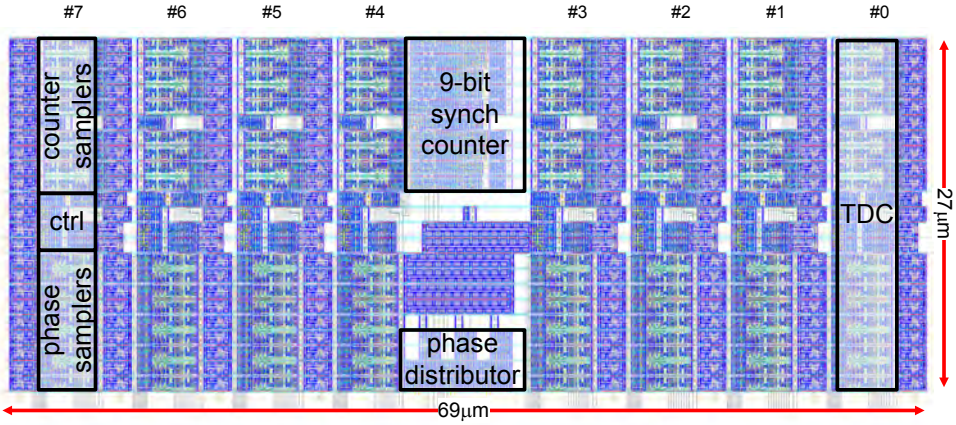


Figure 6.19: Layout of a 8-TDCs pack. Lines have been equalized.

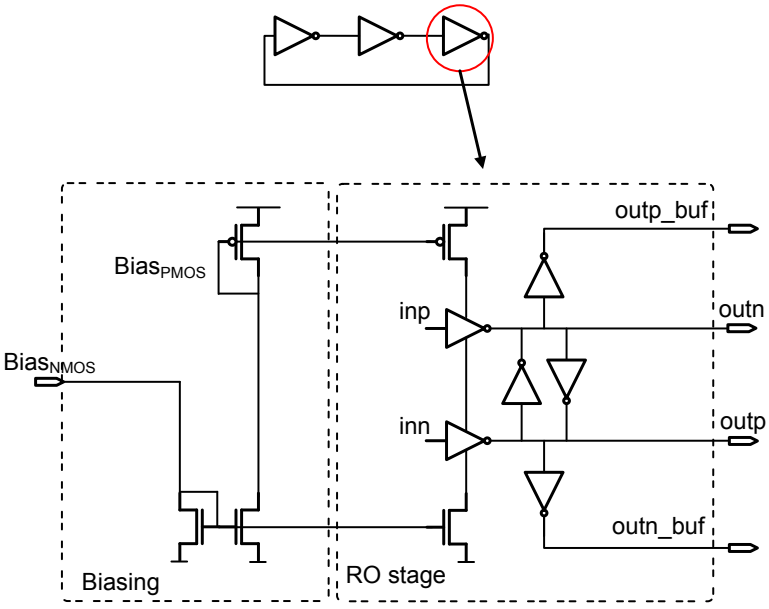


Figure 6.20: Pseudo differential ring oscillator of 3 stages. Stages are starved by a balanced current generated by a current mirror.

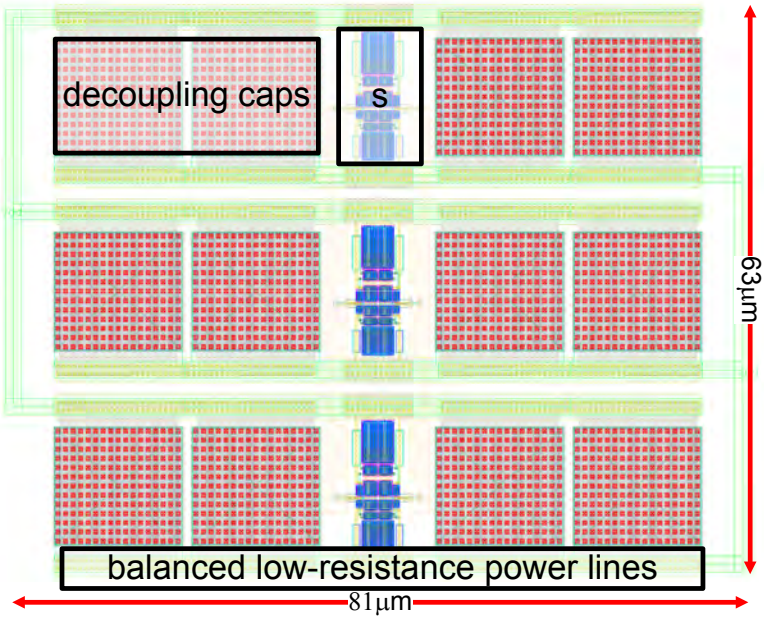


Figure 6.21: Layout of the ring oscillator. The stages of the RO have been laid out with the same orientation to account for process variations. Decoupling caps and power lines are symmetrically placed.

The waveforms of the post-layout simulation of the RO are shown in Fig. 6.22.

The phase samplers use the same sense amplifiers used in TFIFO to assess the differential lines of the SRAM cells. These sense amplifiers are proven to work at frequency in the order of $6GHz$. For more details, refer to chapter 2. The counter of the TDC-pack is totally different from the counter used in Concolor. The first-generation sliding-scale TDCs use asynchronous counters that count the time between an event until the end of the frame. Asynchronous counters do not have glitches and are the easiest to implement. However, they can count only one event per frame. For MindHive, the counters of the TDCs are fully synchronous and can be sampled many times during the same frame and stored into a FIFO. Unlike in the other case, the counters are sampled while they run at full speed, thus making the system possibly incur into metastability if the counters are sampled at the time the are changing. The design of these counters was a big challenge for this new generation of TDCs. The purpose was to gain the possibility to sample multiple events taking advantage of the super fast recovery of the TDC whose time of conversion is around $700ps$.

MindHive has two internal clock dividers in order to enable external access to the frequency generated by the ring oscillator. As it was done in Concolor, the ring oscillator can be controlled externally as a part of a PLL. The first counter is an asynchronous counter that divides the clock by 256, and the second counter is a programmable synchronous counter that can divide the frequency by a range that

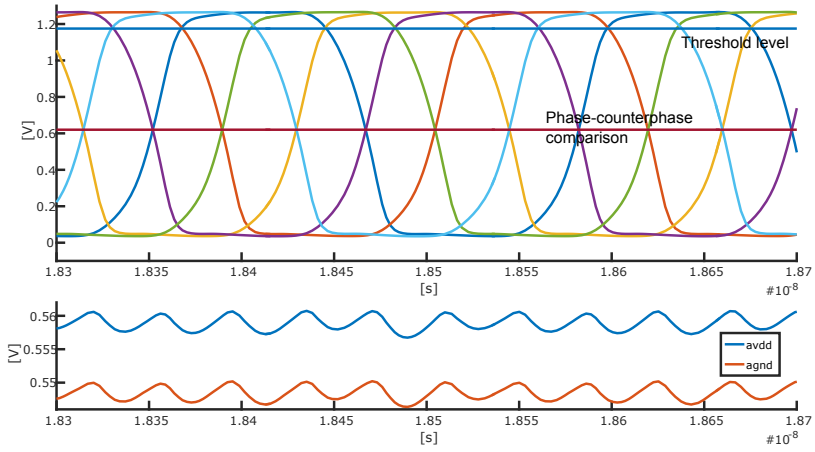


Figure 6.22: For the post-layout simulation, inter-domain buffers, inductance of bonding wires, PAD capacitance, clock distribution buffers and resistance of the power lines were taken into account. Phases on the top. Ripple of power lines and ground on the bottom , referenced to (avdd+agnd)/2.

goes from 2 to 32. Measured jitter of the VCO is 32.3ps. It was calculated in a period of $T = 1\mu s$. The frequency of the synchronous counter is 112.08MHz and the number of cycles within T is 116. The number of VCO cycles that are within T is 3711 and its frequency is 3.586GHz. The stability of the ring oscillator is then $32ppm$. For the jitter measurement, 60.10^6 sets of 116 cycles were considered.

Design of synchronous counters: the counters have 9 bits so they can count until 512 cycles of the VCO. The realization of a 9-bit synchronous counter running at 6GHz is very challenging as the combinational logic needed to set the next value of each of the nine flip-flop is too big and too slow. This frequency, in this technology is close to the maximum operating frequency of the flip-flops themselves, thus making this task nearly impossible. In order to overcome this problem, since synchronous counters are fundamental for this design to achieve multiple events captures, we purposed a design that tackles this issue from three different approaches. The architecture of such counter is shown in Fig. 6.23.

a. Divide-and-conquer strategy: the counter was split into 2 smaller 4-bit synchronous counters plus an independent flip-flop to generate the ninth bit.

b. Slowing down strategy: the frequency of operation was halved by using a flip-flop. In this way the 4-bit counters can work at around 3GHz.

c. Dual phase of the VCO: taking advantage of the multiplicity of phases of the VCO, the negative phase was used to generate the zeroth bit.

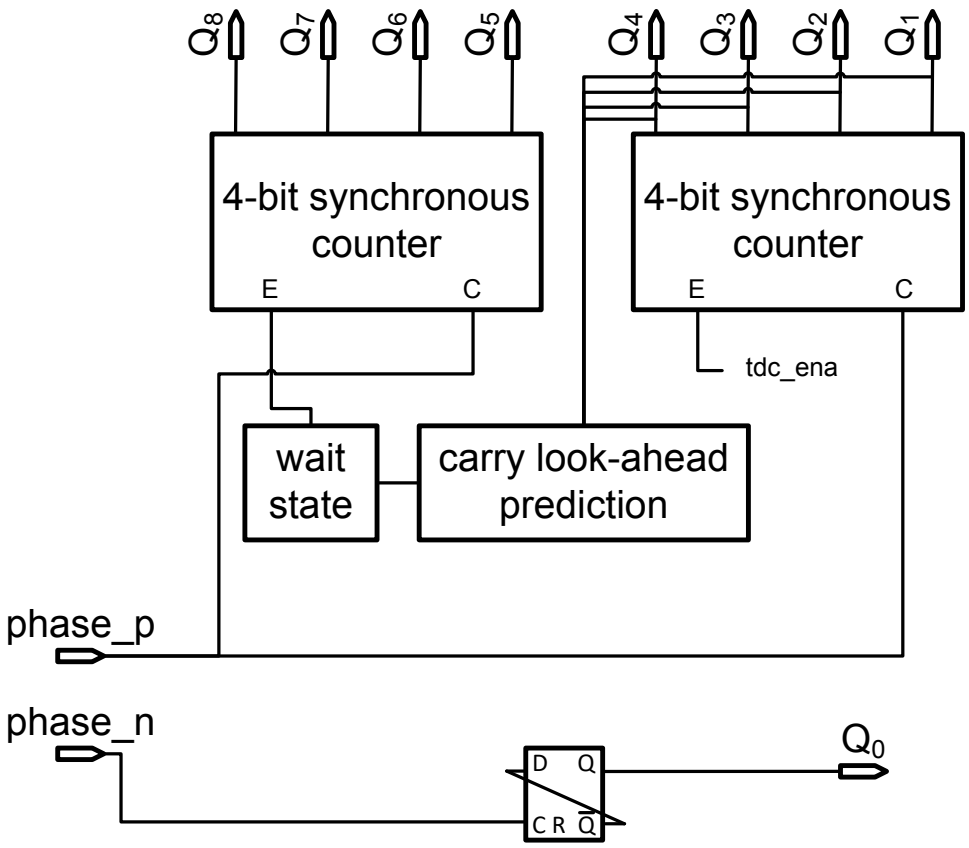


Figure 6.23: The counter has 2 smaller synchronous counter connected through a stage that computes the carry-look-ahead one cycle in advance.

Operation: Every two cycles of the clock, the first counter increments its value by 1 and overflows automatically and freely. A module predicts the overflow 1 cycle before it happens, and after one cycle, the second counter is incremented by 1; this continues until another overflow happens again and again until the second counter overflows and the total system has reached the full range limit. The reason to calculate this overflow one cycle before is because a very well-known problem that synchronous counters suffer from, which is the propagated delay from the stage 0 to the stage N. The independent flip flop, by using the negative phase of the VCO, can provide the bit 0. The reader might have noticed that both Q_0 and Q_1 even if they are fed by different phases, they still change every two clocks in counter phase, so they never change at the same time thus they cannot generate the sequence ″00″ ″01″ ″10″ ″11″ as this requires both bits simultaneously changing. This effect of counter-phase counting actually generates Gray counts leading to a sequence ″00″ ″01″ ″11″ ″10″ and so on. Therefore this 9-bit synchronous counter has 7 bits in binary and 2 bits in Gray code. It needs a posterior de-codification to convert it in fully binary. A pair of XOR gates can do the decodification.

Architecture of the 4-bit synchronous counter: it is necessary to remember why we are doing such a complicated synchronous hybrid binary-Gray design with carry-look-ahead for a 9-bit counter. The goal is to sample the counter at any time during the frame and as many times as needed while the counter runs, there is no time to pause the counter, sample and start it again. It means it will lead to metastability if the sampling happens at the time the bits are changing; consequently getting a wrong value. In order to achieve this, it is required a synchronous counter with its outputs compensated for both the flanks and for parasitics capacitance. The architecture is shown in Fig. 6.25. The 4-bit counters are constituted by fully-balanced flip flops that have same delay for both the flanks to generate changes at the same time and minimize the chance of metastability. Every flip-flop has a MUX connected to a NAND gate and it has only two inputs. The idea behind is that the flip-flop either has to remain in the same value if the previous flip-flops have not reached the maximum count, or, the flip-flop has to change in the contrary case. The inputs of the NANDs were sorted from the fastest to the slowest to compensate the propagated delay of different stages. The post-layout simulation is shown in Fig. 6.24. With this architecture, the maximum frequency of operation achieved was 6.2GHz at nominal voltage and 7.7GHz a 10% over-voltage. The metastability of the counter occurs when any of the flip flops is changing its state. A post-layout simulation showed it has a value of 4.3 ps maximum. The frequency that the counter has normally to run at is 5GHz; the metastability therefore represents 2.15% of the total running time. This is an acceptable value considering that 4.3ps is the worst case and the reference voltage can be finely tuned. Additionally, the sense amplifiers used in the samplers can resolve voltage differences as small as 30mV considering local variations.

Full compensation of the VCOs: since 8 flip-flops are hanging from the positive phase of the VCO and only 1 flip-flop is hanging from the negative phase, this

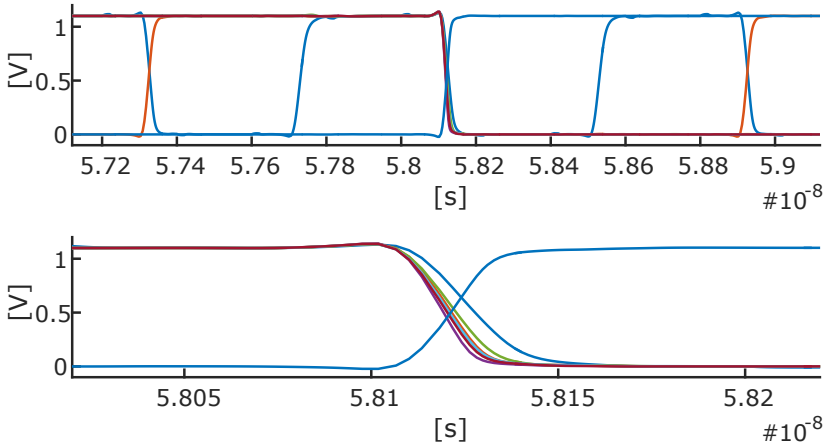


Figure 6.24: Post-layout simulation of the counter (top) and zoomed interested area (bottom). Worst case where all the outputs are changing at the same time. The metstability is 4.3 ps. The most deviated phase is Q_0 but it is not an issue since it works in Gray mode with Q_{int}

would lead to different loads; thus, worsening DNL and increasing delays. The phases were compensated by interleaving the positive and negative phases across the 64 TDCs in such a way that all the phases see the same impedance. The traces that distribute the phases were also compensated by dummy traces to equalize capacitance along the chip.

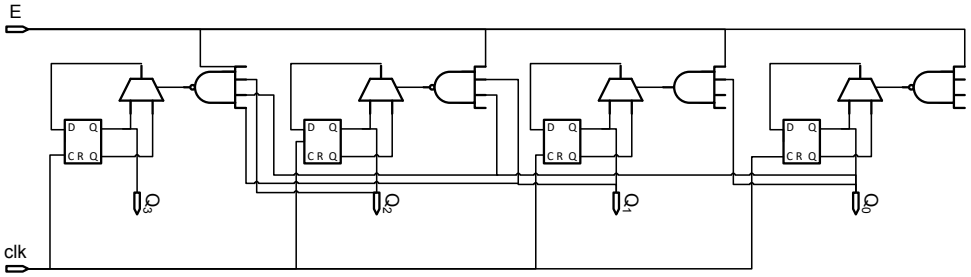


Figure 6.25: Diagram of the 4-bit synchronous counters. The structure is the same for every flip-flop of the chain.

6.3.5. The hive:

In the final subsection of the architecture, after going through all the parts that compose the sensor, we can present the full layout of MindHive in Fig. 6.26.

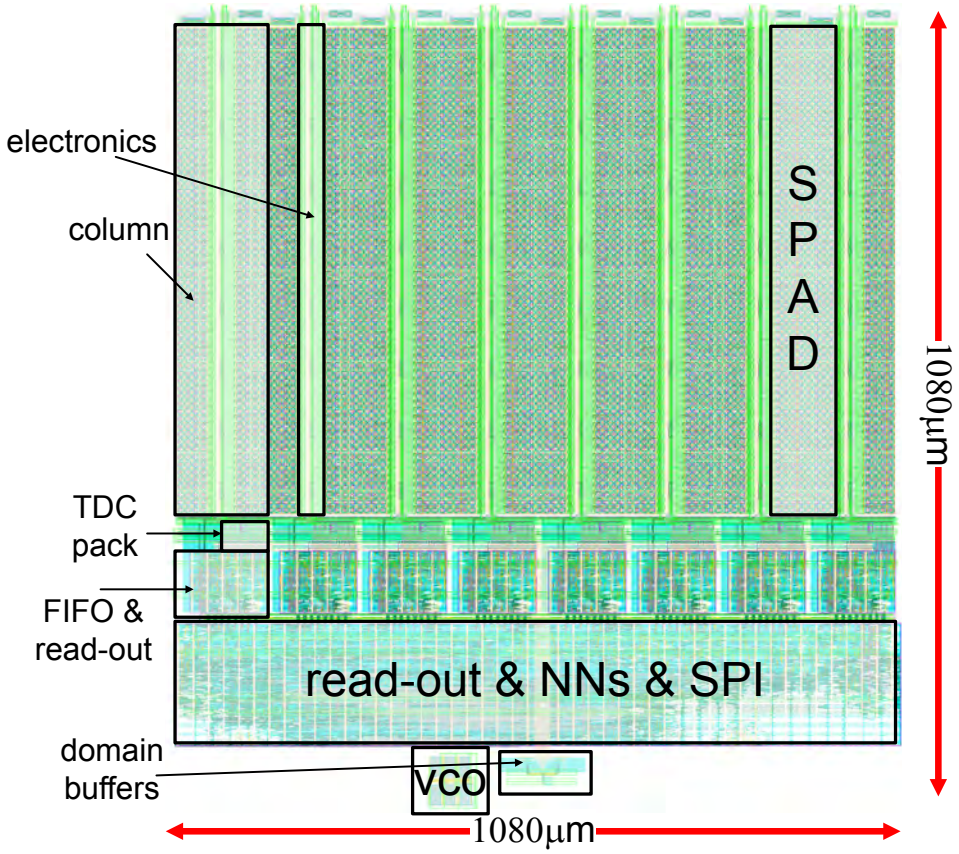


Figure 6.26: The final layout of MindHive.

6.4. Conclusion

A new concept for image sensors has been introduced in this chapter along with a design of an image sensor that implements a reduced version of the whole concept. It was successfully simulated. All digital and analog components have been simulated in post-layout mode that used full extractions from the layouts. Though the system is still under test, many modules were measured and fully characterized: digital modules, SPI, Self-generated clock system, VCO, bias circuits, high-voltage distribution, domain-to-domain buffers, and sense amplifiers. Good agreement between measurements and simulations has been observed.

On the data throughput: the system was synthesized for a high-speed clock of 100MHz. The total processing time for a neuron is therefore $T = 32.1/f = 320ns$. Considering that the neural network has three layers, the total time it takes to process the inputs is 960 ns. However, the network as by-nature pipelined system can perform more operations during that time, meaning the throughput of information is that of the neuron (320 ns) with a latency of 960 ns which was one of the main objectives of the design. For instance, this prompt reaction can be used for driving assistance. Post-layout simulations showed the neural networks can work at this speed and eventually the system could be optimized for higher speeds and bigger neural networks.

On the neural networks: an image sensor with in-silicon implemented neural networks was introduced to bring image sensors to the next step. Two applications utilizing the nets were shown: Optical character recognition and signal digitization. In the future, more analysis should be done about neural networks with memory and self-organized networks that it is most likely to help in self-driving applications.

On storage: For this version, TFIFO and MindHive were designed in parallel. MindHive could then not benefit from the features of TFIFO explained in chapter 2. For this implementation, a regular flip-flop-based synchronous FIFO was designed to store the multiple hits on the TDCs. The disadvantages of this type of memory was fully explained in the same chapter. Next step is to make use TFIFO for this application, possibly reducing its depth to leave more silicon for the sensitive area.

References

[1] S. R. Cherry, A. Y. Louie, and R. E. Jacobs, The Integration of Positron Emission Tomography With Magnetic Resonance Imaging, Proceedings of the IEEE **96**, 416 (2008).

[2] S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf, and E. Charbon, A High-PDE, Backside-Illuminated SPAD in 65/40-nm 3D IC CMOS Pixel With Cascoded Passive Quenching and Active Recharge, IEEE Electron Device Letters **38**, 1547 (2017).

[3] S. S. Haykin, *Neural networks and learning machines*, 3rd ed. (Pearson Education, Upper Saddle River, NJ, 2009).

[4] J. Heaton, *Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks*, Artificial Intelligence for Humans (Createspace Independent Publishing Platform, 2015).

[5] B. Moons, K. Goetschalckx, N. Van Berckelaer, and M. Verhelst, Minimum energy quantized neural networks, in *2017 51st Asilomar Conference on Signals, Systems, and Computers* (2017) pp. 1921–1925.

[6] R. Author Overwater, *Cryogenic Hardware Considerations Of Neural Network Decoders For Quantum Error Correction Using Rotated Surface Codes. (not published) in records at TU Delft* (2019).

[7] P. J. Linstrom and W. G. Mallard, *Eds., NIST Chemistry WebBook, NIST Standard Reference Database Number 69, National Institute of Standards and Technology, Gaithersburg MD, 20899, https://doi.org/10.18434/T4D303* ((retrieved March 18, 2020).

7

Conclusion and future work

7.1. Conclusion of this work

Time-of-flight applications, specially Positron Emission Tomography and LiDAR which are the two discussed in this work, have four different aspects that need to be addressed.

Firstly, they comprise the optical component that in this case is constituted by the SPAD. There is an entire research field on SPADs that tries to develop better SPADs to fulfill the ever-increasing demands of the applications. Those ones include fill factor, time jitter, quantum efficiency, DCR, etc..

Secondly, image sensors have interface electronics that captures events and pulses from the SPADs giving photon and time-of-arrival information. This work has fully explained these modules that include quenching, masking and reset electronics, timing lines, amplifiers, TDCs, etc.. It demonstrated how different architectures for time-stamping and photon-counting work. The methods and circuits implemented and tested in this work will pave the way for the new generation of image sensors with the ever-increasing demand for higher speed, more capabilities, and advanced features. We also presented Sliding-Scale TDCs for image sensors achieving DNL in the order of 0.12 LSB, extended digital TDCs and skew-less timing lines that will boost LiDAR applications in resolution, range and uniformity.

Thirdly, we described very complex and comprehensive computing systems that take the vast information collected by the optical components to process it. These processors are needed by the designs described in this thesis. The processing modules to be included in this type of sensors tightly depends on the application. This work addresses at PET applications in chapter 4 with the design of Concolor and LiDAR applications in chapter 5 with the design of Panther. With help from techniques, such as self-reset, event-driven read-out, full integration of the control digital system and on-chip DCR calculation among many others, we handled the sortcomings of the use of CMOS for SPADs. Concolor was the first chip made in 40-nm technology proved to work for Positron Emission Tomography, also showing results for LiDAR operation. On the other hand, Panther has address-decoders and extended TDCs to provide long ranging for LiDAR applications. It also has on-chip adders and multipliers to calculate the first two momenta of the photons used for PET. Nonetheless, despite the tight dependence of the processing modules to the type of application, we introduced a new concept to design a new generation of image sensors that can potentially address any ToF application. This includes an on-chip reprogrammable unit that can compute all the information and it is not committed to solve only one kind of problem. The concept, fully explained in chapter 6, was implemented in MindHive. Due to technology constraints, the implementation is only part of the full idea. The programmable unit in this case was implemented as a 4-layers neural network intended to be trained for any type of problem. Its results were proved to work for two concrete examples given in the same chapter.

At last, the fourth aspect to take into account is the read-out. This is also critical as all the information that is processed on chip has to be transmitted off the chip. The design performance can be highly compromised if the read-out module cannot deliver data at the pace is required. In this regard, this thesis advocated the

importance of read-out systems for image sensors and showed concrete solutions and implementations with the concept of general and versatile modules that can be re-utilized in the next generation. A particular sub-set of read-out systems is the memory that is used to store TDC captures. This memory has very specific constraints due to the random nature of the events captured by the sensor. In this essay we purposed a new type of memory (TFIFO), built and tested a solution to address this peculiar problem in chapter 2. As part of the same idea, memories designed for FIFO operations, RFIFO memory was presented along with results.

To conclude, this work has pushed the boundaries of knowledge with regard to the electronics involved in the full process from the moment a photon is detected by the optical components (SPADs) until the information is out the sensor. All the topics have been thoroughly covered preparing the ground for a new generation of image sensors.

7.2. Future work

The results of the studies and implementations of Concolor, Panther and MindHive are very encouraging to keep working towards a smart image Sensor capable of processing light towards full integration for computer vision. Self-driving cars are a hot topic at the moment and A.I. is fully capable of pushing it forward [1]. LiDAR started to be used in many other fields mostly to map structures, even space bodies [2]. Fast processing and high resolution and accuracy will boost results for these techniques. The purposed concept of a true FIFO memory was demonstrated and proved to work, and there is plenty of room for improvement in terms of speed and power. Furthermore, a desirable characteristic for memories nowadays has been the ability to process data [3]. This is due to the increasing integration between the processing units and the memories to palliate thermal problems, to boost speed and to reduce latencies [4]. These memories can be used in very complex systems that use A.I., neural networks, etc..

We envision a 3D-image sensor that has two tiers that work together to deliver very detailed, quantitative and qualitative information about the scene the sensor is exposed to. The top tier should be fabricated in the best SPAD technology to maximize fill factor, detection efficiency and time resolution. On the bottom tier, hit and time information is processed by a recurrent neural network connected to a in-situ memory capable of advance mathematical operations like scaling, zooming, roto-translation to preprocess the image before being introduced to the networks. This whole image sensor should be connected to a supervisor system that can train the networks by means of genetic algorithms, machine learning or any other method used in A.I. This system would give qualitative information at an unprecedented speed with outstanding implications for real-time applications. There are many aspects that need special attention. Works are always pursuing more pixel and time resolution for image sensors. However when the systems be supervised and trained by an A.I., the high resolution in time or pixel might not be longer required. Open-loop and close-loop amplifiers are a perfect analogy for this. When amplifiers work in open-loop, the restrictions of gain, bandwidth and biasing are much more demanding than when the same amplifiers are working in close-loop mode. The

structure of the neural networks should be fully analyzed and studied in order to define the best topology that accounts for all the trade-offs described in this work.

References

[1] Self-Driving Cars Could Change the Auto Industry, https://readwrite.com/2020/02/04/self-driving-cars-could-change-the-auto-industry/, accessed: 2020-02-04.

[2] LiDAR, https://en.wikipedia.org/wiki/Lidar, accessed: 2020-03-20.

[3] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory, in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)* (2016) pp. 27–39.

[4] A. Agrawal, J. Torrellas, and S. Idgunji, Xylem: Enhancing Vertical Thermal Conduction in 3D Processor-Memory Stacks, in *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)* (2017) pp. 546–559.

7.3. Glossary

A.I.	Artificial Intelligence
AE	Almost Empty
AF	Almost Full
AND	AND gate
ANN	Artificial Neural Network
APD	Avalanche Photon Diode
BSI	Back-Side Illumination
BW	Bandwidth
CMOS	Complementary Metal Oxide Semiconductor
CNN	Convolutional Neural Network
DAQ	Data Acquisition
DCR	Dark Count Rate
DDR	Double Data Transmission
DNL	Differential Non Linearity
DRC	Design Rule Check
EoC	End of Count
EoF	End of File/Frame
EoT	End of Time
FF	Flip Flop
FIFO	First Input First Output
FLIM	Fluorescence-Lifetime Imaging Microscopy
FPGA	Field Programmable Gate Array
FSI	Front-Side Illumination
FSM	Finite State Machine
HS-RAM	High-Speed RAM
INL	Integral Non Linearity
LFSR	Linear Feedback Shift Register
LoR	Line of Response
LP-RAM	Low-Power RAM
LSB	Least Significant Bit
LUT	Look-Up Table
LVDS	Low Voltage Differential Signal
MD-SiPM	Multi Digital SiPM
MSB	Most Significant Bit
MUX	Multiplexer
NAND	NAND gate
NMOS	N-channel MOS
NN	Neural Network
NOR	NOR gate
OCR	Optical Character Recognition
OR	OR gate
PDE	Photon Detection Efficiency
PDP	Photon Detection Probability
PET	Positron Emission Tomography
PLL	Phase Locked Loop
PMOS	P-channel MOS
PMT	Photon Multiplier Tube

ppm	parts per million
RAM	Random Access Memory
RFIFO	Register-alike FIFO
RO	Ring Oscillator
SDR	Single Data Transmission
SiPM	Silicon Photon Multiplier
SLAM	Simultaneous Localization And Mapping
SNM	Static Noise Margin
SNR	Signal-to-Noise Ratio
SPAD	Single Photon Avalanche Diode
SPI	Serial Port Interface
SR Latch	Set Reset Latch
SRAM	Static Random Access Memory
TDC	Time-to-Digital Converter
TFIFO	True First Input First Output
VCO	Voltage Controlled Oscillator
WD	Word Driver
XOR	XOR gate

Acknowledgements

Getting a PhD degree abroad has undoubtedly been one of the hardest things I had ever done in my life. Obviously it is a challenge from the technical point of view: tons of things to learn, new techniques, new people to work with, etc.., AND there are also many other things at personal level that can be overwhelming at times. Interpersonal relations, being far away from my family, learning a new language, coping with living in a new culture and making friends from any part of the world are some of the things that make you either enjoy life as never before or cry like there is no tomorrow. Nonetheless, which by the way I think it is the most important word in a thesis, this journey is not pulled alone.... on the contrary, there are many, many people that helped me and made this possible too.

I couldn't stress enough how important personal things are. Our own motivation can be our best ally or be our worst enemy. Ups and downs occur, and when we are in the latter (this word I don't like but my professor loves it along with first-ever something), we get support of our beloved ones. We are not machines, forasmuch as we might or might not like to be, interpersonal relations, friendship and love are substantial parameters of the our life equation (what a phrase!, write it down).

Ok, time to thank all those ones who helped in a way or another; that help being technical or supportive.

First of all, I would like to thank Edoardo Charbon, my promoter, who has been with me throughout all this journey going through all the ups and downs and from whom I learned lots of things.

Founders of all the activities carried out during my PhD, PicoSEC project, along with all the wonderful people I met and worked with which I mention here.

Many thanks to Etiennette Auffray who led the project and did everything at her power to push for this project and for the students. Thanks also to Erika Garutti for having me at DESY Institute for 3 months as the first part of my PhD. Special thanks to the Technical University of Delft (TU Delft) which I will remember forever. It has been the place where everything was cooked and done. Thanks to my former university: Universidad Tecnológica Nacional (UTN) where I got my diploma of Electronics Engineer and my master of Systems Optimization and Control jointly with the university Université de Technologie de Troyes (UTT), France. I would like to thank Comisión Nacional de Energía Atómica (CNEA) in Argentina, where I worked for 5 years before diving into my PhD and I learned about PET systems.

There are many people that I work with in a large collaboration that I would like to give my gratitude to.

Many thanks to Sara Pellegrini, Bruce Rae, David Poyner and Neale Dutton for helping with fabrication, discussions and hosting me at STMicroelectronics for our co-joined work. Thanks to Tarek Abbas for helping us on the final stage of the tape-out.

I want to also thank the scientists from LETI that helped me with radiation measurements, Eric Gros D'Aillon and Laurent Cortella.

I want to thank my colleagues and friends of my group and university. I will write their names in the order I met them to prevent any bad feeling, as most of them are very sensitive, you wouldn't believe it.

Chockalingam Veerappan, captain on the bridge! We immediately became friends, his kindness and spirit are only comparable to his knowledge about SPADs (which is extremely high if you are wondering).

Shingo Mandai, an inspiration of work and consistency, smart, nice and funny. We quickly became friends and he was the first person I worked with in this PhD. I kinda latinized him and hugged him the last day, I swear it was not awkward.

Ting Gong, we forged a close and an amazing friendship during these 6 years. He is always open for technical discussions and personal talks. Fully latinized by now, refer to the previous paragraph. We also worked together at the beginning of my PhD with amazing and out-of-this-world outcomes.

[0] Chao Zhang, the super engineer, you are never bored when next to him, got very famous for phrases like "good" or "just finish!". He also helped me with the first steps with Encounter tools.

Michel Antolovic, an authentic powerhouse of talent, we slowly became friends to now be old and supportive friends. He has also been a inspiration for the way I understand the approach to science. This was decisive for my last project.

Scott Lindner, as soon as we met, we started a really nice and long friendship, although at first I didn't understand anything of what he would say... maybe that's why, kidding. We worked, enjoyed, suffered and relaxed together in this journey. He also helped me at work in one of the most stressful moments in my life. Thanks Scott.

Bishnu Patra, it took some time to become friends but he couldn't escape in the end. We did a trip to China together that was memorable. Thanks for the discussions on LC-tank oscillators.

Arin Ulku, he helped me a lot during a critical tape-out that I was fully overwhelmed by. We did trips together and he also hosted me when I went to Switzerland in the last part of my PhD. He is an amazing friend, and BTW, I didn't take the coffee machine, Arin.

Harald Homulle, it took a while to get close with the dutchest Dutch, but eventually I managed or he did? I don't recall. Thanks for your friendship.

Jeroen van Dijk, brilliant and innocent at the same time, we spent many good moments together and trips that will stay in my memory forever.

Rosario Incandella, partner in crime for the Chinese expedition.

Augusto Ximenes, my namesake or simply AX, we worked together in many designs, learning, discussing and suffering together, and pulling a working chip! not bad hu?. That is worth a big ....oh waaaaw!.

Preethi Padmanabhan, Prit and I shared a lot of things and our friendship grew strong along all these years. We support and we are always there for each other. Coffees, trips, long working evenings made us go through similar situations that reinforced our empathy. Thanks for your friendship.

Ramon Overwater, I would like to thank him for all the discussions and exchange about Neural Networks that helped me to carry out some of the last experiments.

Antoon Frehe, he works at the informatics department of the university, and I can't express how much he helped with licenses, missing files, buggy applications and so many other things to help me carry out my experiments and simulations. Thanks a lot.

Joyce van Velzen, she is in a league of her own, helping and solving problems left and right, I thank you a lot Joyce!

Minaksie Ramsoekh, Hozan Miro many thanks for the long days working in the basement fixing the mistakes on my chips!

Zu Yao Chang, I want to thank you for all the help with the bonding. Your technical skills can't be matched by any human.

I would like to thank Msc. Ing. Lucio José Martinez Garbino for his stone-made friendship and for his help on specific topics where no entity, be it human or electro-mechanic, is on a par with him: neural networks and FIR filters.

I want to thank Dr. Msc. Ing. Franco Ferrucci, expert in Statistics and Control Systems who helped me to find the equations for the saturation of MD-SiPMs. Thanks for your friendship as well.

Gerd Kiene, one of the last friends I made, but definitely not less because of that, we started a wonderful friendship which we forged during a crazy tape-out that I was on the verge of a total collapse, many times just about to go to a corner, crunch and suck my thumb like a baby. We set hourly 5-min breaks to destress a bit. What I love about Gerd is the fact that he is highly realistic. He won't say things just to make you feel good, instead he accepts reality and carries with it.

Thanks a lot to Ashish Sachdeva for his wise friendship. I wish you the best.

Pascal 't Hart, I had very good moments with all of them.

I also want to thank colleagues of the cool group, Pinakin, Jian Gong and Job van Staveren.

My recent colleagues of EPFL:

Andrea Ruffino, though old so new! I have great memories with you my friend.

Pouyan, my maaan!, thanks for your friendship, we need to resume our runs!.

Milo, very cool and smart guy that I instantly built a friendship with. He is also an extremely good chess player, searching for his first title. Not sure if he can become GM (Grand Master), but maybe he can lead a chess club like BlunderFest or e4ThenResign.

I also made friends at EPFL, thanks to make my stay in Switzerland very enjoyable: thanks to Francesco, Andrada, Yatao, Bedo and Andrei.

Mamá, my beloved, all-supportive, all-mighty mom, thanks for all your love, knowledge and experience that you gave from when I was a child. The values learnt and the love helped to be the man I am now. Thanks a lot, I love you with all my heart.

mah sista, my second mom has been always next to me from the moment I was born. Our bond knows no distances and gets stronger with as the time goes by. Love you with my heart.

Milagros Serrano, my dear niece, thanks for having come to our lives, I wish you the best for your bright future with all my love.

Marcelo Serrano, thanks for all the love you give to my sister, my mom and my niece too, you made us happy as soon as you walked in in our lives. Best wishes.

Vanesa Carimatto, mi adorable prima, siempre estamos juntos, no importa la distancia, los problemas o las circunstancias. Te quiero mucho, y te agradezco por todo el aguante y paciencia de todos estos años.

Santigo Camba, hermoso del padrino, me encanta ver cómo estás creciendo, sano, bueno y un chico de bien. Muy orgulloso de ser tu padrino, te amo muchísimo.

Muchas gracias a mis primos que los adoro Jésica, Diego, Nati, Pablo, Lucas, Flor, a mis tíos Mónica y Norberto, Quique y Piedad, a mi tía abuela Amelia.

Quiero agradecer a mis amigos de toda la vida!

Pablo Naddeo, Mauro Spagnuolo, Andrés González Andújar, Guillermo Juárez, Guille Sanchez, Ignacio Camba, Anita Ibarrola, Martín Belzunce, Emiliano Achigar, Soledad, Alan, Thomas, Marisa Calello, MariCe Velásquez, Laura Bianchi.

También a los que siempre van a estar en mi corazón, mi abuelo Coco, mi abuela Alicia y mi padrino Carlos.

I also would like to thank friends that I made in Delft, who not only gave me all their love, but also support and help whenever I needed it.

Thanks to María del Rocío Arroyo Valles for her unconditional friendship. We did many trips together, we are always there for each other. She is an extremely experienced chancellor, moreover when it comes to academics.

Reyes Menendez González, famous for her phrase: "first year, everything is new, you are adapting, second year you are working as there is no tomorrow, third year you start preparing for the defense, and there goes by your fourth year", this has been said 6 years ago, actually. Hats off, thanks for your friendship.

Pía González, she is a friend of those who show up when you are in need. Lovely and a wonderful person. Thanks for being my friend.

Cristina Palacios Camarero, her practicality is second to none. If you are doubting how to solve a problem, just call her. Thanks for your friendship and help.

Mathieu Blanke, the least-dutch of the Dutches, he likes warm weather, good food, has a relative small distance bubble, I didn't have too much work to do to latinize him. Bedankt, mijn vriend.

Laura and Fije, "neighbours are your first relatives" is a well-known phrase in Holland, never more true with these two. Thanks for your friendship.

Emilio and Andrés Kenda, my very last friends during my PhD, we were quarantined together in the same house during the pandemic while I was wrapping up my thesis. They made my life easier and more enjoyable.

If I may, I will try to give some small pieces of wisdom that I amassed during these years:

Never be shy before new experiences, everything is learnable, always move forward.

No matter how many times you stumble, you get up and fight.

Learn from your mistakes, but truly do it! Otherwise the experience would be wasted. Think how you will not fall into similar things again and that also will give

you the chance to forgive yourself, because the new-you wouldn't make the same decisions and mistakes.

If you are in love with someone, tell them no matter what. You might not have a second chance. Love is inexplicable and unidirectional; for that matter, if it happens to be requited, it is a divine miracle that deserves any risk.

Never let third parties assess you, moreover if those parties have plans for you.

In despite of how bad things are going, you can always start over from scratch, erasing everything that happened and starting with new spirit.

Declaimer: I hope I have not forgotten anyone to thank. In case you don't find yourself in this thesis, send me an email, you will get a prompt response apologizing for the unwanted omission or, alternatively, you could get a confirmation that should not be here, kidding.

Thanks again to everyone... and on that note, I will call it a PhD [0].