UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE *MATHEMATIQUES, SCIENCES DE L'INFORMATION ET DE L'INGENIEUR*

THÈSE Présentée par

Imane MALASS

soutenue le : **13 MAI 2016**

pour obtenir le grade de **Docteur de l'université de Strasbourg**

Discipline : Électronique, Électrotechnique et Automatique
Spécialité : Micro et Nanoélectronique

Design of an Integrated Streak Camera based on a Time Correlated Single Photon Counting System

DIRECTEUR DE THESE :
UHRING Wilfried — Professeur, Université de Strasbourg, France

CO-DIRECTEUR DE THESE :
LE NORMAND Jean-Pierre — MCF HDR, Université de Strasbourg, France

RAPPORTEURS :
CALMON Francis — Professeur INSA de Lyon, France
CHARBON Edoardo — Professeur EPFL, Lausanne, Suisse

AUTRES MEMBRES DU JURY :
NORMAND Stéphane — Chercheur Ingénieur HDR, CEA, Saclay, France
DUMAS Norbert — MCF, Université de Strasbourg, France

MEMBRE INVITE:
LEONARD Jérémie — Chargé de recherche HDR, IPCMS, Strasbourg, France

Imane Malass

Logo partenaire

Design of an Integrated Streak Camera based on a Time Correlated Single Photon Counting System

Résumé

Nous présentons une caméra à balayage de fente intégrée basée sur un système de comptage de photon unique résolu en temps (TCSPC-SC) employant l'architecture linéaire « streak » pour surmonter la limitation de l'espace inhérent aux systèmes TCSPC bidimensionnels. Cette solution permet l'intégration de fonctionnalités électroniques complexes dans les pixels sans l'inconvénient d'un faible facteur de remplissage conduisant à une faible efficacité de détection. Le TCSPC-SC se compose de deux blocs principaux: une photodiode à avalanche (SPAD) et un bloc de mesure de temps, les deux blocs sont intégrés en technologie 180 nm CMOS standard. La structure de la SPAD utilisée a été sélectionnée parmi 6 structures différentes après un processus de caractérisation précise et approfondie. Le bloc de mesure du temps se compose d'un TDC hybride capable d'atteindre des résolutions de temps élevées et ajustables avec une large dynamique de mesure grâce à un système de conversion de temps (TDC) hybride qui combine l'approche analogique basée sur un convertisseur de temps vers amplitude(TAC), et les approches numériques utilisant une boucle à verrouillage de retard (DLL) et un compteur numérique 12 bit-. Le TDC hybride a été spécialement conçu pour être utilisé dans un système TCSPC qui intègre une ligne de TDC nécessitant ainsi une conception appropriée pour limiter la consommation d'énergie et la surface d'occupation et parvenir à une architecture flexible et facilement extensible. Suite à la conception et la réalisation de ces deux blocs dans une technologie180 nm CMOS standard, une structure de test de la caméra à balayage de fente (TCSPC-SC) qui englobe 8 unités a été réalisée dans le but final de mettre en œuvre un modèle TCSPC-SC complet et plus large.

Résumé en anglais

In this work we present a TCSPC Streak Camera (TCSPC-SC) that takes advantage of the streak mode imaging to overcome the space limitation inherent to 2D TCSPC sensor arrays. This cost-effective solution allows the integration of complex functionalities in the pixel without the inconvenience of low fill factor that leads to low detection efficiency. The TCSPC-SC consists of two main building blocks: a SPAD and a time measurement block both integrated in 180 nm Standard CMOS technology. The SPAD was selected among 6 different SPAD structures following a thorough characterization process to fully determine its performance figures. The time measurement block consists of a hybrid TDC capable of achieving high adjustable time resolutions with large dynamic range owing to a time conversion scheme that combines traditional Analog Time to Amplitude Converter (TAC), Digital DLL-based and counter-based TDC. Furthermore, the hybrid TDC was especially designed to be used in a TCSPC system that incorporates an array of TDCs which required a careful design to limit power consumption and occupation area in order to achieve a flexible and easily scalable architecture. These two building blocks were both fabricated in a 180 nm standard CMOS technology and employed to demonstrate a TCSPC Streak Camera (TCSPC-SC) test structure that englobes 8 units in order to demonstrate the system's operation principle with the final aim of implementing a complete and bigger TCSPC-SC model in the near future

Acknowledgement

The work presented in this document would not have been possible without the help and the guidance of several individuals who in one way or another contributed to this project.

First and foremost, my deepest and utmost gratitude goes to my advisor Wilfried Uhring without whom I wouldn't have had the chance to do this thesis. His guidance, vast knowledge and support were essential for the completion of this work and his great personal qualities added considerably to my PhD experience. It was a true pleasure to witness up-close the remarkable joy and enthusiasm he has for his research and I cannot find words to express my gratitude for his kindness and the friendship he extended toward me since I started working with him.

My sincere gratitude also goes to my co-advisor Jean-Pierre Le Normand for his guidance, his patience and his vital assistance especially during the long nights preceding a deadline. I enjoyed sharing the same office with him and I am profoundly grateful for his kindness and the friendliness we shared.

I would like to take this opportunity to express my earnest gratitude to Professor Luc Hebrard. Pr. Hebrard was not directly involved in my PhD project but this achievement is in part a result of his guidance and excellent pedagogical skills. I have nothing but great admiration for Pr. Hebrard and I consider it a true privilege to be one of his students.

I would like to sincerely thank Foudil Dadouche, Norbert Dumas and Jean-Baptiste Schell for offering their help whenever it was needed. My thanks also goes to Pr. Daniel Mathiot, Freddy Anstotz, Frédéric Antoni, Jean-Baptiste Kamerer and Nicolas Collin.

I would like to thank the Microelectronic group in the Institut Pluridisciplinaire Hubert CURIEN (IPHC), I especially thank Claude Colledani, Frédéric Morel and Gregory Bertolone for kindly providing me with their much appreciated help and technical support.

To my fellow PhD students Fitsum Aweke, Julien Zelgowski, Laurent Osberger, Marwa Garci, Octavian Maciu, Rémi Bonnard, Simon Paulus, Thomas Regrettier, and Vincent Wlotzko, thank you for all the pleasant discussions and the fun times.

Last but not least I wish to express my deepest and utmost gratitude to my parents, Moustafa and Ghada, without whom none of this would have been possible. I am forever indebted to them for all their hard work, dedication and support over the years and to them I dedicate this thesis.

Abstract

Streak cameras have been traditionally based on the vacuum tube technology, a mature and ultrafast technology that, despite great performance figures, suffer from bulkiness, fragility and expensive cost. In recent, Many efforts have been made to find a solid state alternative to the vacuum technology and several integrated streak cameras have been realized in (Bi)CMOS technology allowing a temporal resolution close to 1 ns and most of these devices are specifically designed to operate in single shot mode, their sensitivity represents a real issue for measurements of weak optical signals.

This problem can be resolved in the case of recurrent optical phenomena by employing the time correlated single photon counting (TCSPC), a technique that takes advantage of high gain photon detectors such as Photo-Multiplier Tubes (PMTs) and (MCPs) to measure weak optical signals. These mature technologies are capable of achieving very good performances but they are also expensive, cumbersome, fragile, and require high operating voltages (~kV) which makes them unsuitable for the fabrication of miniaturized portable TCSPC imaging systems. Following the emergence of Silicon Photo Multiplier (SiPM) based on single photon avalanche diodes (SPADs) integrated in standard CMOS technology in the 2000's; several integrated 2D-TCSPC systems have been demonstrated. These 2D systems consist of SPAD arrays integrated in 2D with their associated electronics resulting in a tradeoff between high photon detection efficiency and advanced electronic functionalities. The use of 3 dimensional (3D) heterogeneous integration with deep-submicron CMOS readout electronics represents a good solution to the previously mentioned limitations but the 3D technology is still new, immature and highly costly.

In this work we present a TCSPC Streak Camera (TCSPC-SC) that takes advantage of the streak mode imaging to overcome the space limitation inherent to 2D TCSPC sensor arrays. This cost-effective solution allows the integration of complex functionalities in the pixel without the inconvenience of low fill factor that leads to low detection efficiency. The TCSPC-SC consists of two main building blocks: a SPAD and a time measurement block both integrated in 180 nm Standard CMOS technology. The SPAD was selected among 6 different SPAD structures following a thorough characterization process to fully determine its performance figures. The time measurement block consists of a hybrid TDC capable of achieving high adjustable time resolutions with large dynamic range owing to a time conversion scheme that combines traditional Analog Time to Amplitude Converter (TAC), Digital DLL-based and counter-based TDC. Furthermore, the hybrid TDC was especially designed to be used in a TCSPC system that incorporates an array of TDCs which required a careful design to limit power consumption and occupation area in order to achieve a flexible and easily scalable architecture. These two building blocks were both fabricated in a 180 nm standard CMOS technology and employed to demonstrate a TCSPC Streak Camera (TCSPC-SC) test structure that englobes 8 units in order to demonstrate the system's operation principle with the final aim of implementing a complete and bigger TCSPC-SC model in the near future.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 The Origin of Photography

The history of the camera can be traced much further back than the introduction of photography to the *camera obscura*. The term "*camera obscura*", meaning dark chamber in Latin, was used by the German astronomer Johannes Kepler in the early 17th century but the device was first used much earlier. Indeed, the earliest recorded mention of this type of device was by the Chinese philosopher Mo-Ti (5th century BC) who formally recorded the creation of an inverted image formed by light rays passing through a pinhole into a darkened room and called this darkened room a "collecting place" or the "locked treasure room." Aristotle (384-322 BC) understood the optical principle of the *camera obscura* and made practical use of it to observe the shape of a partially eclipsed sun projected on the ground through the holes in a sieve and the gaps between leaves of a plane tree. The Islamic scholar and scientist Alhazen (Abu Ali al-Hasan Ibn al-Haitham) (c.965 - 1039) gave a full account of the principle noting that reversed images were formed only by means of small holes. The Song Dynasty Chinese scientist Shen Kuo (1031–1095) also experimented with a *camera obscura*, and was the first to apply geometrical and quantitative attributes to it without proclaiming that he was the first to experiment with such a device. In 1490, it was Leonardo Da Vinci who published the first clear description of the *camera obscura* in *Codex Atlanticus* (1502). The image quality was later improved with the addition of a convex lens into the aperture in the 16th century by Giambattista della Porta who compared the shape of the human eye to the lens in his *camera obscura*, and provided a readily comprehensible example of how light forms images in the eye. He also promoted the use of this device for painting. In 1685, Johann Zahn published *the Oculus Artificialis Teledioptricus Sive Telescopium*, which contains many descriptions, diagrams, illustrations and sketches of the *camera obscura*. Early models of the *camera obscura* were large, consisting of either a whole darkened room or a tent as employed by Johannes Kepler but by the 18th century, more easily portable models became available owing to developments made by Robert Boyle and Robert Hooke; these cameras were later adapted creating the first photographs.

The first photographic image was made by the French inventor Joseph Nicéphore Niépce using a *camera obscura*. Niépce formed the image using the light passing through a hole on a polished metal sheet of pewter with a bitumen coating. After reportedly 8 hours of exposure, the bitumen on the lightened sections of the plate was hardened proportionally to its exposure to the light. The unhardened regions were removed after being washed by a solvent to create a print of the observed scene (Figure 1).

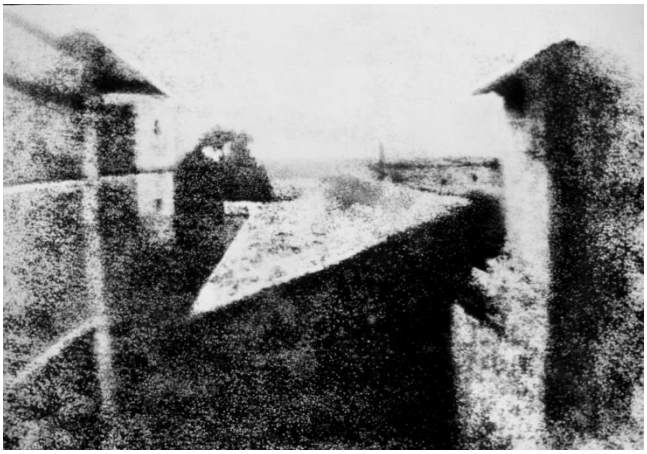


Figure 1 - "Point de vue du Gras", 1826. The oldest surviving camera photograph taken by Niépce

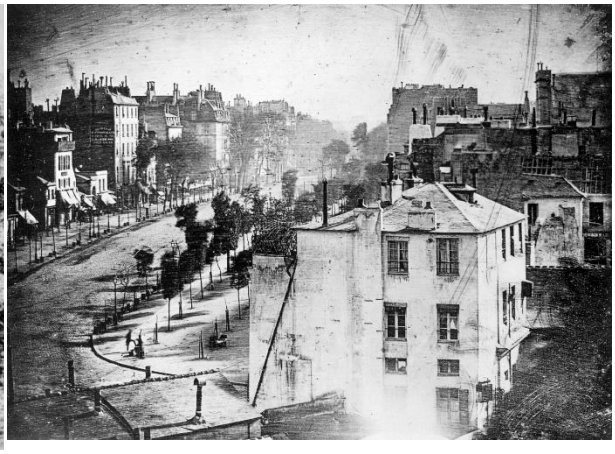


Figure 2 - "Boulevard du Temple". First ever photograph of a human taken by Daguerre in Paris, 1838.

After Niépce's sudden death in 1833, his notes were handed to Louis Daguerre who worked to improve the chemical process involving the interaction of the plate with light. In 1839 he announced the invention of a new process using silver on a copper plate. The process, dubbed Daguerreotype after the name of its inventor, led to the reduction of the exposure time to 30 minutes and the image was no longer progressively disappearing afterwards. This invention denotes the birth of modern photography and it gained popularity very quickly as an alternative to expensive oil paintings that required long hours to be completed..

1.2 From Photography to Cinematography

In 1878, the English photographer Eadweard Muybridge took the first high-speed sequence of 12 photos known as *The Horse in Motion* (Figure 3) in a bid to find out if there was moment mid-stride where horses had all hooves off the ground. The pictures of the sequence were taken by a dozen cameras sequentially triggered with a set of strings, the resultant pictures were spaced from each other by around 400 ms and the exposure time was less than 500 µs. A year later, Muybridge presented the *Zoopraxiscope*, a simple device that projects images drawn on a rotating glass disks in rapid succession to give the impression of motion. The first roll film was patented in 1882 by George Eastman, which led to the acquisition of the first motion pictures. Four years later, Louis Le Prince, patented *Method of and apparatus for producing animated pictures*. He filmed what is believed to be the first moving picture sequences using a 16 lens *receiver* (Figure 4), which is what he called the

camera, and an Eastman Kodak paper film. The film, known *the Round hay Garden Scene* was shot at a speed of 12 frames per second (fps) and lasted less than 2 seconds.

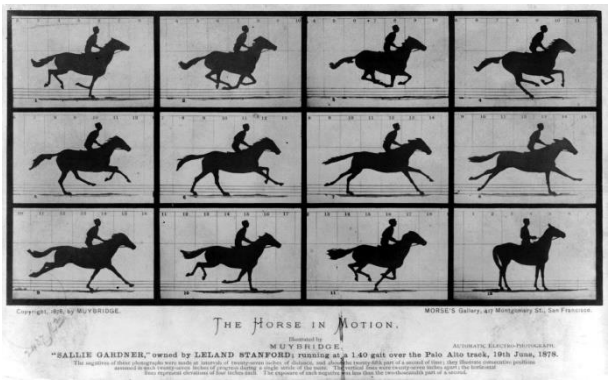


Figure 3 - 'The horse in motion', first high speed sequence" taken by Muybridge in 1886

Figure 4 - Le prince 16 lens receiver

Two years later, Thomas Edison presented *The Kinetoscope*, a motion picture device inspired by the *Zoopraxiscope, capable* of acquiring sequences at a speed up to 40 fps. The *Kinetoscope* created the illusion of movement by conveying a strip of perforated film filled with sequential images over a light source through a mechanical shutter. In 1904, the Austrian physicist August Musger patented the *Kinematograph mit Optischem Ausgleich der Bildwanderung*. The instrument, shown in Figure 5, was capable of recording fast transients and projecting them in slow motion using a mirrored drum as a synchronizing mechanism. Musger's high speed film shooting principle was used during the First World War by the German company Ernemann-Werke A.-G to develop the *Zeitmikroscop,* a 500 fps camera used mainly for ballistic purposes. In 1926, Heape and Grylls constructed the Heape and Grylls Machine for High Speed Photography where the film shifting mechanism was replaced by a film lining the inside of a rotating drum driven by an 8 horse power engine (Figure 6). The 4 tons instrument [1] was designed for the British Army; it was capable of reaching a frame rate of 5000 fps and was mainly used for studying the action of machine guns and puncturing of armor plates by projectiles. This instrument was later replaced by more compact versions using a drum powered by an electric motor [2].

Drum cameras are still in use until these days and their fundamental principle is still the same (Figure 7). The fastest drum cameras produce frame records at up to 200 000 fps using a drum rotating in an air evacuated housing to reduce friction. Higher speeds are hard to achieve with drum cameras due to mechanical constraints imposed by the need of a synchronous rotation of the mirror and the drum at very high speeds.

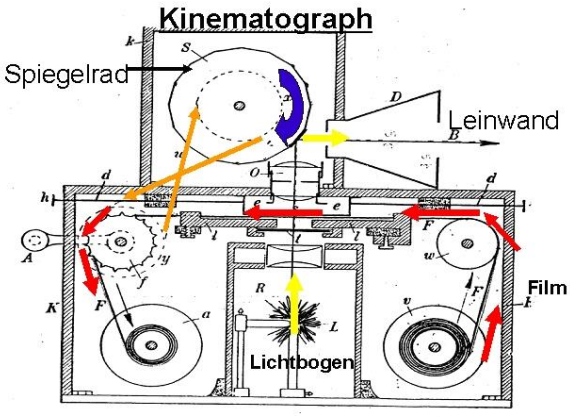


Figure 5 - Operation principle of a Kinematograph

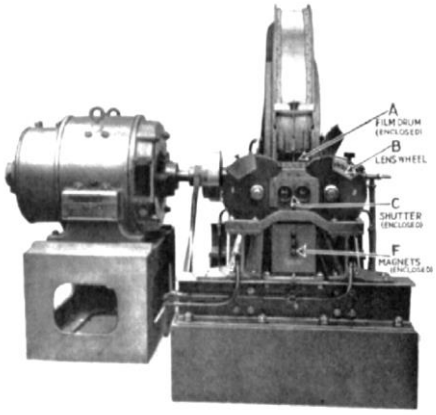


Figure 6 - Photograph of the Heape and Grylls's Machine for High-speed Photography [1]

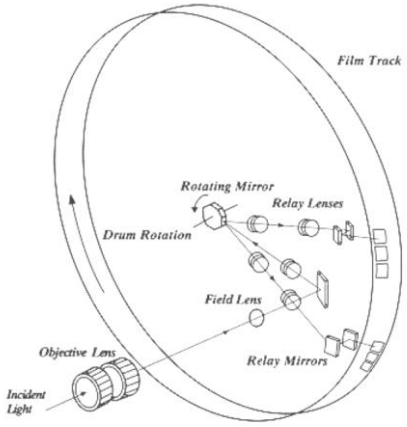


Figure 7 - Operation principle of a drum camera

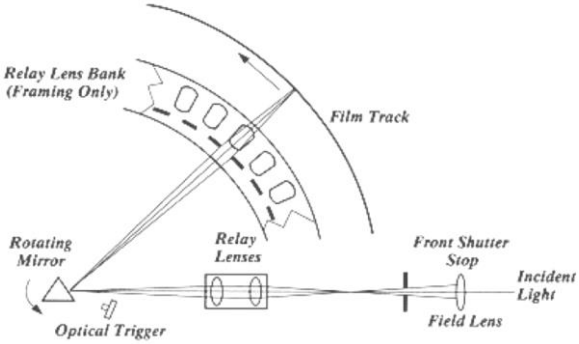


Figure 8 - Operation principle of a rotating mirror camera

Toward the end of World War II, a new type of cameras called the rotating mirror camera was being developed in the aim of photographing atomic explosions. The image formed by the objective lens is transmitted to a rotating mirror which sweeps the reflected image focused through an arc of relay lenses and shuttering stops on a static film (Figure 8). Overwriting was avoided by using an input capping shutter or by quenching the light source. The first rotating mirror camera was built by Miller in 1939 and the concept was patented in 1946 [3]. Miller's original rotating mirror camera reached a speed of 500 000 fps. This speed was doubled several years later when Berlin Brixner built a camera with a speed of 1 million fps using the same principle [4]. The current commercial Cordin's Model 510 rotating mirror camera reached 25 million fps, by driving the mirror at up to 1.2 million rpm in a helium environment using a gas turbine [5]. Rotating mirror cameras are capable of achieving frame

rates as high as 25 million fps, with typical speed in the millions of fps but this technology have reached the speed limit fixed by the maximum rotation speed of the rotating mirror and the drum which hasn't improve for more than 40 years [6]. The speed barrier imposed on mechanical cameras due to physical constraints was transcended using the image conversion tubes, a novel technology based on the successive photon–electron–photon conversion. The first video camera using this technology was Zworykin's *Iconoscope* in 1935. The basic principle of these cameras is to project the image on the photosensitive cathode of a vacuum tube. The photocathode emits a number of electrons proportional to the number of photons. The electrons are then accelerated through the tube due to the existence of a high electrostatic field until hitting a screen at the opposite end of the tube. The electrons are thus converted back into photons and the screen is projected on a film or a digital camera. During the 50's, image converter cameras achieved previously unattainable temporal resolution and sensitivity through the electrical manipulation of the exposure duration and the use of internal image intensifiers. The maximal frame rate achievable by this technology is limited by the shuttering and deflection electronics to about 100 million fps. Next the *Orthicon*, the *Vidicon* and the *Plumbicon* tube video cameras were introduced respectively by Rose in 1937, Weimer in 1950, and Philips in 1970. A schematic diagram of an *Orthicon* converter tube is shown in Figure 9. The image is projected on an input window coated with a fine metal layer with photoelectric characteristics. The electrons emitted by this photocathode are extracted and accelerated towards a target screen by a high electric field and directed through a fine wire mesh, acting as a screen grid. As a result, a positively charged electronic image of the scene appears on the target. This image is then scanned by an electron-beam guided by a set of deflection coils, which absorbs the positive charge on the target in proportion to its magnitude. The remaining electrons from the beam are then reflected and their number is a linear measure of the target's original charge. The reflected signal is amplified through a multistage electron multiplier network and conveyed to an output screen where the electrons are converted back to photons. Image converter tube cameras were capable of achieving very performance figures but they are fragile, cumbersome and relatively expensive and as a result they were rapidly replaced by the emerging solid state optical sensors.

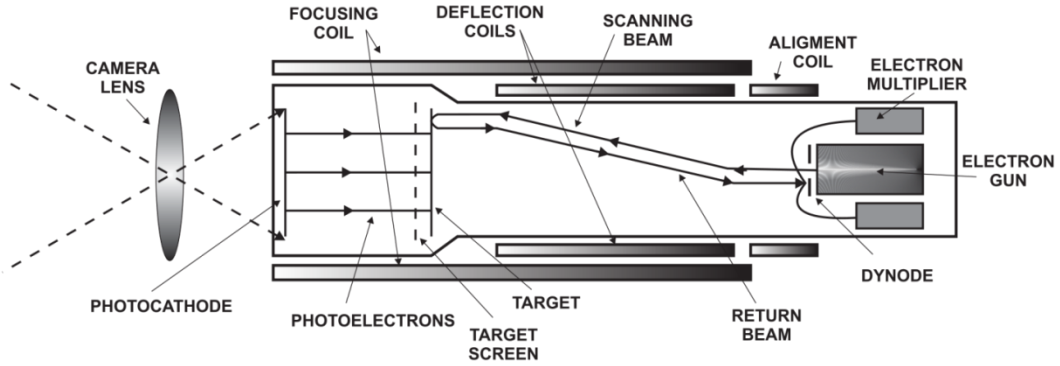


Figure 9 - Operation principle of an Orthicon tube

1.3 The Emergence of CMOS Imagers

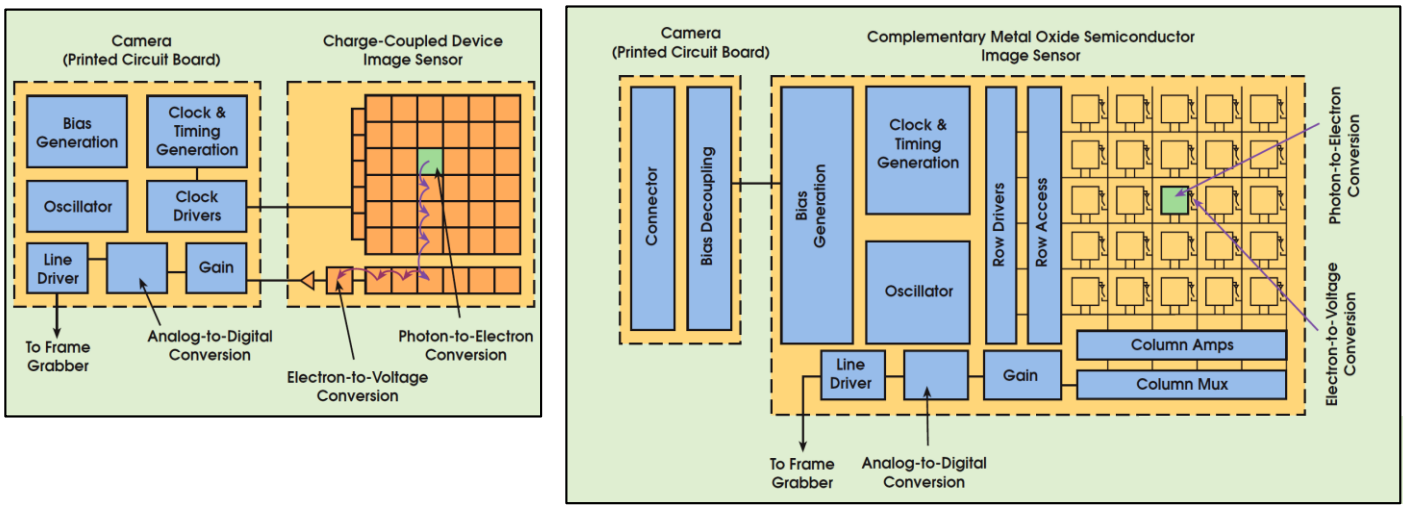


Figure 10 - Typical architecture of a CMOS camera (left) and a CCD camera (right) [7]

The basic operation of both CCD and CMOS image sensor is to convert light into electricity by exploiting the photoelectric effect. Both sensors must perform the same basic functions: generate and collect charge, convert it into voltage or current and output the signal. The different between the two technologies is in how these basic functions are carried. The functional blocks of CCDs are similar to those of a CMOS camera (Figure 10) but only the light detection and the signal transfer are integrated on the sensor while the other functionalities needed to operate the camera are implemented on an external Printed Circuit Board (PCB). In the case of a CMOS image sensor, all these functionalities are implemented on the same chip. The history of Complementary Metal-Oxyde-Semiconductor (CMOS) image sensors started in the 1960's with the release of the first solid-state imagers aiming to propose a more flexible alternative to photographic films and image converter tubes. The demonstration of a functioning CMOS image sensor was promising but the performances obtained were quite poor compared to the existing imaging technologies at that time [8], [9]. This was mostly due to the novelty of the CMOS technology and the used fabrication process. Soon after the first CMOS image sensors were introduced in 1970, the first charged coupled device (CCD) imagers [10].

1.3.1 CCD Cameras

The first CCD was invented in 1969 by Boyle and Smith at the Bell Laboratories [11]. Initially developed for semiconductor bubble memory purposes; this technology was then adapted to be used in image sensor applications and gradually replaced the photographic film [12]. CCDs consist of light sensitive cells or pixels that are capable of producing an electrical charge proportional to the amount of light they receive. The electric charge in each pixel is

transferred to the output stage of the sensor where it is converted to a voltage level, buffered, and sent out as an analog signal which is then amplified and converted to numbers using an off-chip amplifier and A/D-converter (Figure 10- left). Depending on the application, the pixels can be arranged in either a single line or in a two-dimensional grid with each CCD pixel comprising a MOS capacitor used simultaneously as a light detector and data storage device. MOS capacitor exists in two types: surface channel and buried channel. The two differ only slightly in their fabrication however buried channel capacitors offer major advantages and as a result almost all of today's CCDs employ the buried channel structure. A schematic cross section of a buried channel capacitor is shown in Figure 11. When under illumination, negative carriers are created by photogeneration and accumulated in the depletion region beneath the MOS capacitor gate known as potential well. Individual sensing elements assembled in CCD array are electrically isolated from their neighboring pixels by isolating barriers or channel stops in the substrate. Once collected, the charge packets held in the potential wells are subsequently shifted from pixel to pixel under the influence of the cycling clock pattern applied to the gates (Figure 10 - left). The N-type buried layer is employed to form a PN junction structure, which localize the potential wells beneath the silicon–silicon dioxide interface. This technique significantly lowers the charge trapping during the charge transfer thus increasing the charge transfer efficiency. Outputted charge is converted to voltage by an off-chip output amplifier. The rate at which data is transferred depends on the bandwidth of the output amplifier and the required charge transfer efficiency.

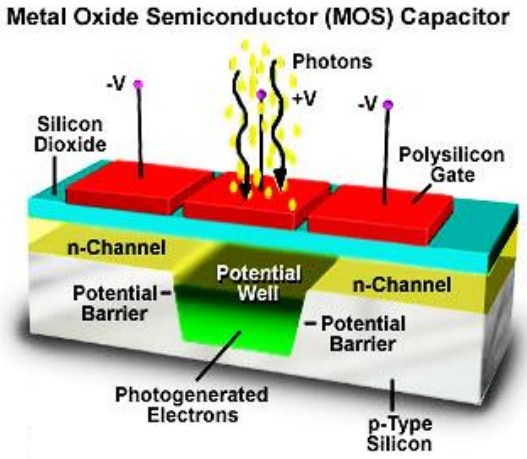


Figure 11 - Schematic cross section of a buried channel MOS capacitor, the basic CCD sensing element in CCDs [13]

The first working CCD was invented in 1969 at the Bell Laboratories and researchers were able to capture images with simple linear CCDs in 1970. Four years later Fairchild Semiconductor demonstrated a linear 500-element device and a 2-D 100 x 100 pixel device and in 1975, the first CCD camera was invented using a Fairchild 100 x 100 CCD. The first *KH-11 KENNAN reconnaissance* satellite was launched equipped with an 800×800 pixels CCD array technology for imaging in 1976 [14]. Soon after, CCD was used in a medical imaging application (Laparoscopy) in 1982 and in 1983 CCD cameras were employed in telescopes

for the first time. Up until the early 2000s, CCDs were widely manufactured and preferred over CMOS imagers mainly because of their superior image quality. But the progress made in CMOS technologies led to improvement in CMOS imagers making them less expensive, more accessible and more complex due to improvement in their designs and fabrication process.

1.3.2 CMOS Cameras

In the 1960's solid-state image sensors were being developed by numerous research groups with varying degrees of success using NMOS, PMOS, and bipolar processes. In late 1960s, Noble and Chamberlain invented separate sensor arrays that utilized MOS output amplifiers. The pixels were arranged two-dimensionally and each column had an amplifier, but the designs were faulty with a high-level of white noise as well as a low transmission rate [14]. No significant improvement were made on the design until the early 1990s when engineers demonstrated the first CMOS sensors with higher output speed and lower noise level. Still, until quite recently, CMOS sensors could not match the image quality of CCDs but in recent years, Improvements in CMOS design and fabrication technology have allowed CMOS sensor image quality to rival that of CCDs. Indeed, CMOS imagers benefits from the multipurpose nature of CMOS processes, which allow the implementation of a camera-on-a-chip integrated with a mixed-signal processing circuit and other complex functionalities within the same die thus creating a smart sensor with advanced features such as noise reduction, blooming suppression or dedicated functions like real time target tracking, three-dimensional range finding, and modulated light detection to be easily implemented next to the sensor, reducing the cost and improving the performance of the imager. Today, CMOS image sensors achieve higher performance and are able to realize versatile functions that cannot be accomplished with CCDs. They currently dominate the integrated circuit imaging market thanks to lowered power consumption, reduced production price and faster time to market. Still CMOS cameras suffer from some shortcomings compared to CCDs. In CMOS image sensors, the integration of electronics in the pixel leads to a lower percentage of photosensitive area in the pixel defined as ***fill factor*** (the fill factor in CCD image sensors is 90%-100 % whereas CMOS image are limited to 60-70 %). The use of additional layers in CMOS process results in a lower quantum efficiency compared to CCDs, The ***Quantum efficiency*** (QE) is the measure of the efficiency with which incident photons are detected as some incident photons may be lost due to optical reflection or absorption where the electrons cannot be collected. CMOS image sensors show lower QE compared to CCD mostly due to additional layers used in the CMOS process. Another inconvenient in image sensors is the ***Dark current noise*** defined as the unwanted charge that accumulates in the pixels due to natural thermal processes that occur while the device operates at any temperature above absolute zero. Again CMOS image sensors exhibits higher level of dark current noise compared to the CCDs specially optimized to limit this factor. Several solutions are possible to improve the performances CMOS image sensors such as the use of micro-lens to increase

the photosensitivity of the pixels, the use of dedicated CMOS processes allow the improvement of QE up to 80% for a fixed wavelength interval at the price of using a more expensive process that allows a better absorption of the light in the chip. Finally the uniformity and the noise levels can be reduced using the Correlated double sampling technic based on comparing the obtained image to a reference measured in the dark.

1.4 Streak Mode Cameras

1.4.1 Streak-Mode Imaging Concept

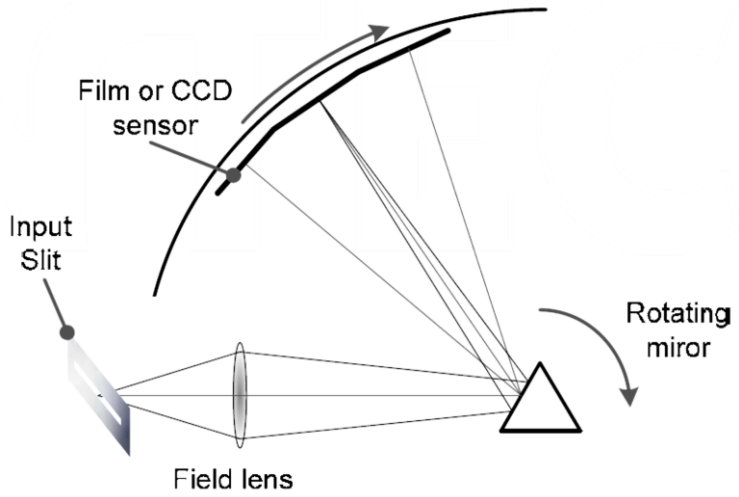


Figure 12 - Principle of the rotating mirror high speed video camera [15]

All the devices presented until now produce 2-dimensional (x,y) images I_f separated by equal time intervals Δt. Consequently, I_f is a function of the two spatial dimensions x, y and a time constant $t_0 + n.\Delta t$, where n is an integer:

$$I_f = f(x, y, t_0 + n.\Delta t) \quad (1)$$

This is known as the framing mode photography. This mode has the advantage of recording the information in the form of two spatial dimensions representation; as a result the recorded image is an easily recognized version of the subject. However, when finer temporal information is required, a considerable gain in the frame rate can be achieved by reducing the spatial information. This is the case in streak-mode imaging where only one spatial dimension is recorded resulting in a picture containing the continuous temporal evolution of this one spatial information. The recorded image I_s contains the spatial information x which crosses the slit, observed at different instants t:

$$I_s = f(x, t) \quad (2)$$

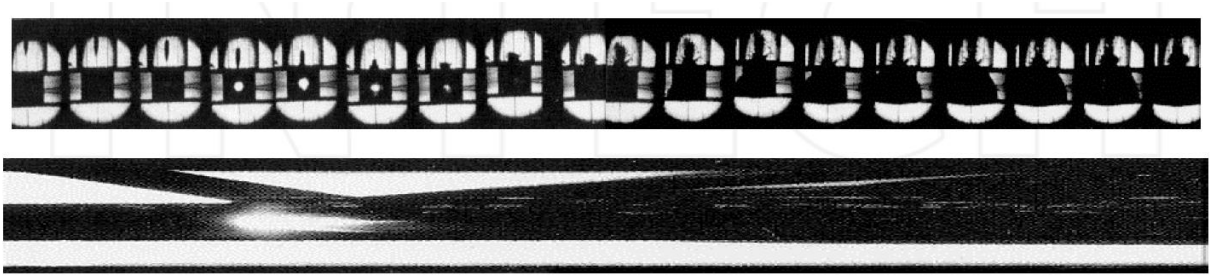


Figure 13 - Illustration of the streak imaging concept: frame record at 100 000 fps (top) and streak record at 0.53 mm/µs (bottom) of a ballistic event acquired by a Cordin Model 330 camera [5]

A streak record is made by placing a narrow mechanical slit between the event and the camera. Next, the temporal evolution of the one dimensional spatial information crossing the slit is swept along a taping material resulting in a continuous record containing the position, time and intensity information. The recording rate, called sweep speed, translates the distance covered on the record surface for a unity of time and is measured in mm/µs or ps/pixel. To operate a rotating mirror camera in streak-mode, the relay lenses are removed and a slit is positioned on the input optical path, (Figure 12). Using this technique, a maximal sweep speed of 490ps/pixel has been reported in [5]. Figure 13 shows a framing record (top) and a streak record (bottom) of a projectile fired through a near-side cut plastic chamber to impact an explosive sample. By using the streak mode, continuous inter-frame temporal information of the slit image is recorded at the price of a single dimension representation. The improvement in temporal resolution using Streak-mode imaging is about 2 orders of magnitude which making it ideal for studying events of uniform growth, e.g., an expanding sphere, where the rate of dimensional change is to be measured. In general, the streak and framing records are often combined to avoid misinterpretation and obtain the maximum amount of information. Thus, framing and streak-mode cameras are not comparable, but rather complementary [16].

1.4.2 The Conventional Streak Camera

The rotating mirror technology described earlier is constrained by the centrifugal force applied to the mirror as a result of the very high rotation speed. Even with the use of light metals like the beryllium does not prevent the mirror from explosion when the rotation speed exceeds 2 million rpm [17]. To overcome this limitation a completely different technology employing vacuum tubes is used. The conventional streak camera is the fastest device for direct light measurement available with a temporal resolution in the picosecond-order. The operation principle of a conventional streak camera [15] is depicted in Figure 14. The very core of the streak camera is a modified first-generation sealed vacuum image converter tube, known as a streak tube, comprising four main sections: a photon-to-electron converter, an electron bunch focusing stage, an electrostatic sweep unit, and an electron-to-photon conversion stage. On some streak tubes, an internal Micro Channel Plate (MCP) is added in front of the phosphor screen for signal amplification. A mechanical slit is illuminated

by the time varying luminous event to be measured and is focalized on the photocathode of the streak tube. The incident photons on the photocathode are converted to photoelectrons with quantum efficiency depending on the type of the photocathode. A mesh is placed in the proximity of the photocathode and a high static voltage is applied between these two components in order to generate a high electrical field, which extracts the photoelectrons from the photocathode, makes their velocities uniform, and accelerates the pulse of photoelectrons along the tube. At this stage, the photogenerated electrons represent a direct image of the optical pulse that has reached the photocathode surface. When the photoelectrons approach the sweep electrodes, a very fast voltage ramp *V(t)* of several hundred Volts per nanosecond is applied at a timing synchronized with the incident light. During the voltage sweep, the electrons which arrive at slightly different times are deflected in different angles in the vertical direction. As far, first a photo-electrical conversion is carried out by the photocathode, and then a translation from time to space is operated through the sweep electrodes. After being deflected, the electrons enter the MCP, where they are multiplied several thousands of times by bouncing on the internal channel walls acting as continuous dynodes and are extracted by the high electrical field applied on both sides of the MCP. At the end of the streak tube the photoelectrons impact against a screen which emits a number of photons proportional to the incident electron density.

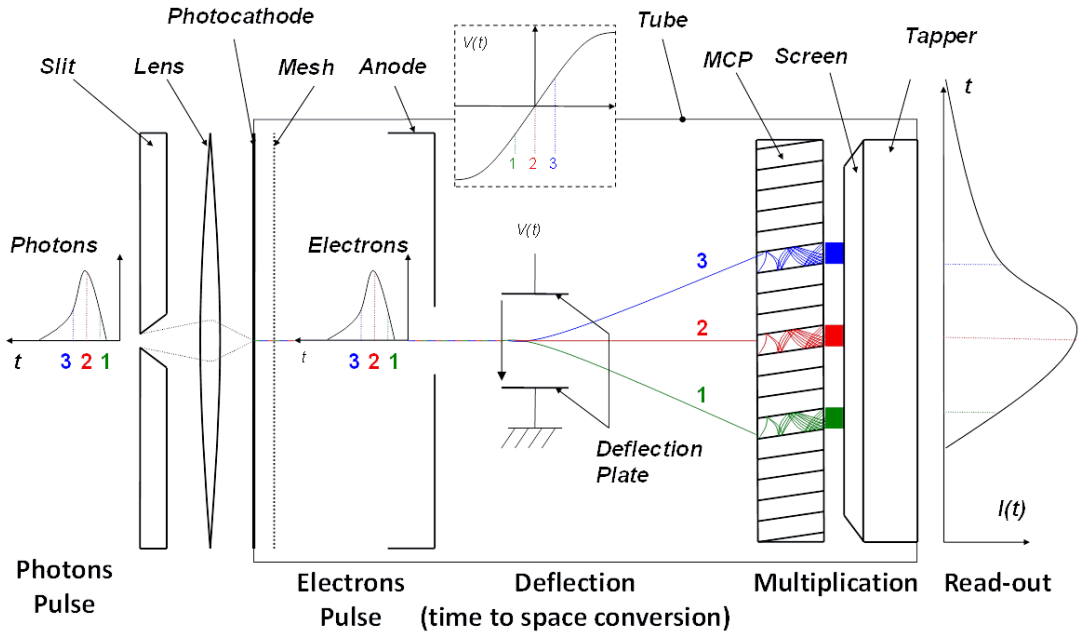


Figure 14 - Operation principle of a conventional streak camera [15]

To prevent dispersion, an objective or a tapper made of a fine optical fiber grid is positioned in-between the phosphor screen and the read-out CCD or CMOS camera to guide the emitted photons. As the time-varying electric field caused by the voltage ramp between the electrodes is assumed to be spatially uniform, the spatial distribution of the light is directly obtained on the phosphorus screen without temporal modification. Finally, the temporal axis is found

along the vertical direction of the screen, the position in the horizontal direction corresponds to the location of the incident light, and the brightness is proportional to the intensity of the respective optical pulses. The relationship between a given vertical position x and the time t depends on the slew rate $S_R\ [V/s]$ of the sweep voltage $V(t)$ and the deflection sensitivity $D_s\ [V/mm]$ of the streak tube is:

$$t = t_0 + \frac{D_s}{S_R}.(x - x_0) \quad (3)$$

Where x_0 is the position in mm of the slot image on the phosphorus screen when $V(t_0)=0$. As for the rotating mirror camera operating in streak-mode, the sweep speed of the camera is found as the ratio D_s/S_R, generally expressed in ps/mm. The conventional streak cameras are very versatile devices with high-end performances (Table 1). The wavelength detection spectrum ranges from the X ray [18] up to the far infrared [19]. Their very high sensitivity allows the detection of a single photon event with the repetition rates ranging from single shot operation up to several hundred of MHz. The typical spatial resolution of a conventional streak camera is between 25 µm and 100 µm. Specific tube designs using magnetic solenoid lens can reach a spatial resolution of 10 µm [20]. Finally their temporal resolutions are very close to the physical theoretic limitation, about 100 fs [21] but with a poor signal to noise ratio whereas 1 ps can be reached with a high signal to noise ratio [22]. Besides their extreme performances, conventional streak cameras and rotating mirror cameras have drawbacks: they are bulky, fragile, expensive (~100 k€) and delicate to manufacture. Moreover there are many applications in which a temporal resolution of about 1 ns is sufficient. For these applications, solid-state technologies can offer interesting alternatives.

Table 1 - Performance summary of a conventional streak camera [15]

Spectral range	[41 pm, 10 µm] depending on the photocathode type
Sensitivity	Single photon detection capability (with MCP or image intensifier)
Repetition rate	From single shot up to 250 MHz (synchroscan)
Spatial resolution	From 25 µm (ns resolution) to 100 µm (ps resolution)
Temporal resolution	From 200 fs (single shot) or 2 ps (synchroscan) to 300 µs
Sweep speed	From 10 ps/mm (fast sweep unit) up to 5 ms/mm (slow sweep unit)
Observation time	From 60 ps up to 175 ms (phosphorus screen size ∈ [9–35 mm])

1.4.3 Integrated Streak Camera

Conventional streak cameras are based on vacuum tube technology a mature technology that has reached its intrinsic limit with a temporal resolution value of 10 fs, a limit predicted 50 years ago by Zavoiski [23] [24] [25]. However, many applications such as bacteria identification by fluorescence [26], laser-Doppler velocimetry [27] or the study of random transitional phenomena in MicroElectro-Mechanical Systems [28], require a temporal resolution with a nanosecond order of magnitude [29] in which case solid-state technologies can offer interesting, more compact and above all less expensive alternatives. The idea of using classic pixel array CCD or CMOS sensors in order to obtain a tubeless streak camera was first reported in [30] and [31]. The concept is based on using fast moving opto-mechanical parts in a rotating mirror high speed streak camera developed during the 1940's. The highest temporal resolution reported using this approach is 490 ps/pixel, it was obtained for a rotation speed of 300000 rpm which required a very complex and delicate mechanical assembly [32]. In 2001, a study on a new concept of tubeless SC based on a time-resolved pixel array CMOS imager was published [33]. In this solution, only a mechanical slit and an appropriate optical objective are required for reproducing the functionality of a CSC on a single CMOS chip. The optical setup needed for a correct operation of the so called Matrix Integrated Streak Camera (MISC) is schematized in Figure 15. The image of a light source illuminating a mechanical slit is uniformly spread along the temporal axis (the rows) of the 2D CMOS sensor through a cylindrical lens. Thus, each pixel of the same row is subjected to the same optical event. The pixels of the sensor operate in the integrating mode. The circuit temporally sweeps the columns by shifting the beginning of their integrating phase. Consequently, each pixel of a given row is subjected to the same illumination, which is measured at different moments. The first generation of MISC features a 64 × 64 pixel array and a fixed sampling rate of 1.25 GS/s. In [34], a second generation of MISC allowing signal shuttering, analog on-chip accumulation and a sampling rate of 1.56 GS/s has been reported. In both prototypes, the pixels operate in the

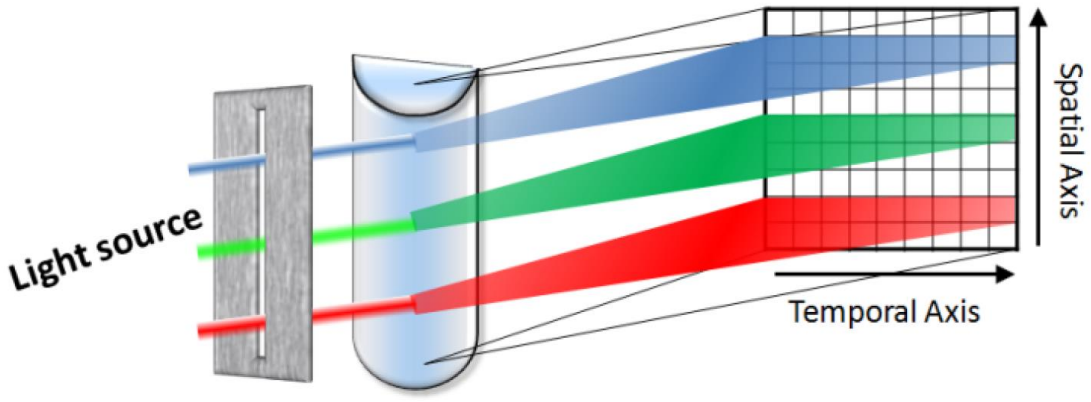


Figure 15 - Optical setup of a MISC [35]

In both prototypes, the pixels operate in the integrating mode [36] an appropriate signal processing of the raw output data is necessary to recreate a record of the luminous phenomenon recorded. A more direct approach has been also reported with the Single Vector ISCs (VISC). In [37], a solid state streak camera consisting of a 1-D linear array (Figure 16) of 150 photodiodes operating in the integrating mode with regular reset between samples and a 150-frame analog storage array is presented. The maximum electrical sampling rate achieved was 400 MS/s. In [38], another VISC prototype featuring a 12 photodetector column, a 128-deep analog memory block, and a 7.14 GS/s sampling rate is presented. The photodetectors in this design are operated in the direct current mode and the photogenerated current is converted to voltage via wideband transimpedance amplifiers (TIA). The TIA VISC prototype was fabricated in standard BiCMOS technology and reached a temporal resolution of ~500 ps, an overall bandwidth of 1 GHz and a global sampling rate of more than 700 GS/s [39]. This architecture allowed a sensitivity 100 to 1000 time better than the MISC in [34] as all the optical signal was focalized on a single pixel with a fill factor close to 100% instead of a total line of pixels but the improvement came at the price of a more complex architecture and higher power consumption. The use of a standard (Bi)CMOS technology to implement the MISC and VISC devices ensures quick prototyping with very low production cost estimated to be about 10 % of the price of a traditional streak-mode imaging devices. Moreover, the integration possibilities offered by CMOS technologies allow the use of the in-situ storage approach in the streak-mode CMOS imagers. However, these devices were designed to operate in single shot mode and their sensitivity represents a real issue for all phenomena where the optical power to be measured is weak. This limitation can be solved in the case of recurrent phenomena using another approach that allows the observation of very low power events. This approach is known by the Time Correlated Single Photon Counting (TCSPC).

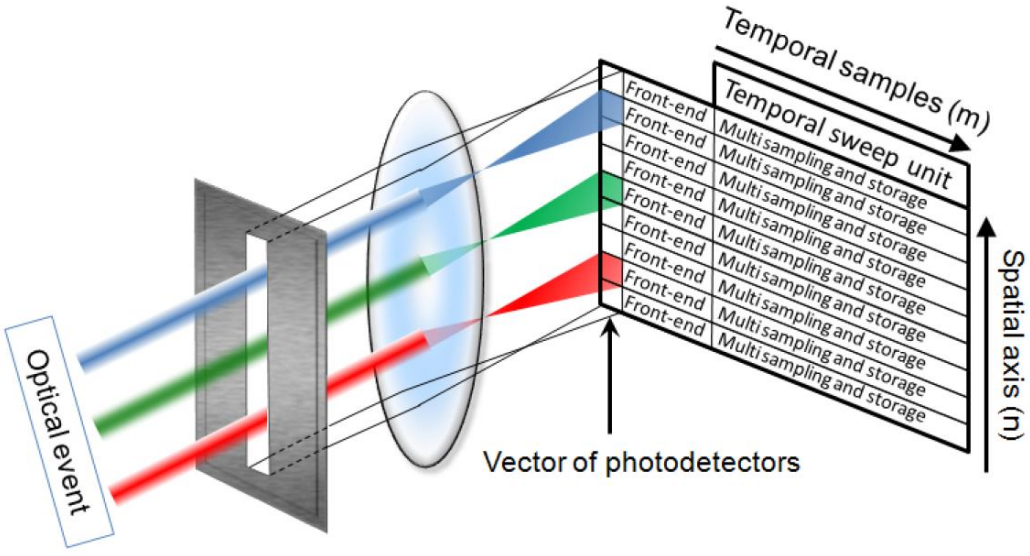


Figure 16 - Optical setup and architecture of a VISC [40]

1.5 Time Correlated Single Photon Counting

1.5.1 Fundamentals and History of TCSPC

Time-Correlated Single Photon Counting (TCSPC) is a mature and extremely accurate low light signals measurement technique that uses single quanta of light to provide information on the temporal structure of the light signal. The method was first conceived in nuclear physics [41] and was for a long time primarily used to analyze the light emitted as fluorescence during the relaxation of molecules from an optically excited state to a lower energy state [42].Today TCSPC is widely used in many applications that require the analysis of fast weak periodic light events with a resolution of tens of picoseconds such as diffuse optical tomography (DOT) [43] [44], fluorescence life time imaging (FLIM) [45] and high throughput screening (HTS) [46]. TCSPC is based on detecting single photons of a periodical light signal, measuring the detection times within the light period and reconstructing the light waveform from the individual time measurements after repeating the measurements for enough times (Figure 17).

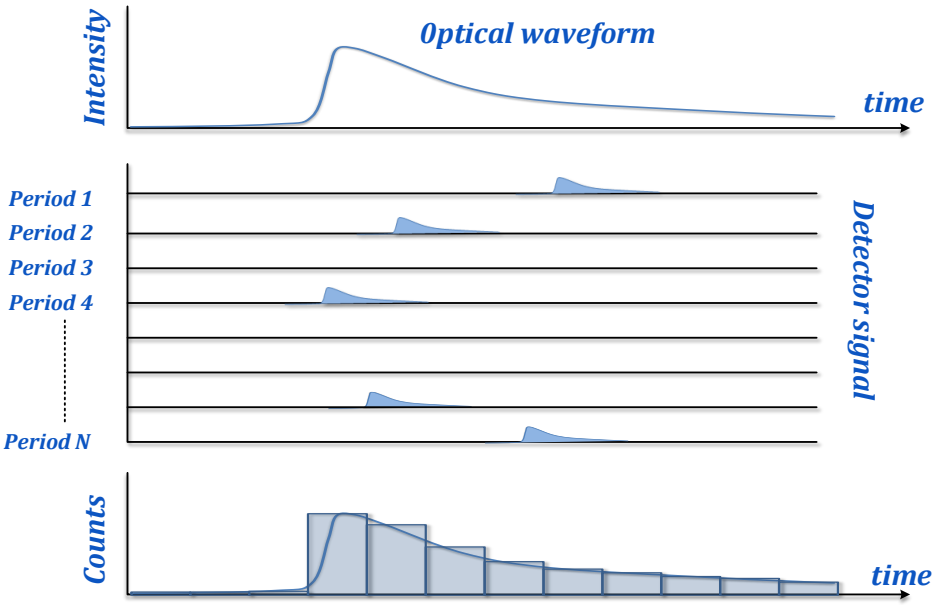


Figure 17 - TCSPC Measurement Principle

The TCSPC technique makes use of the fact that for low level, high repetition rate signals the light intensity is usually so low that the probability of detecting one photon in one signal period is much lower than one. Therefore, the case where several photons are detected within the same measurement period can be neglected and the principle shown in Figure 17 can be used. The detector signal consists of a train of randomly distributed pulses due to the detection of the individual photons. There are many signal periods without photons; other signal periods contain one photon pulse and the periods with more than one photon are

extremely rare. In order to appreciate the merits of TCSPC, one must consider the direct alternative of analog transient recording. Obviously analog recording is well suited for high-intensity signals; however in the case of low intensities the Signal to Noise Ratio (SNR) defined as the square root of the number of detected photons (N) within the impulse response time of the photon detector drops far below 1 [47]. As a result the sensitivity of the system is limited by the baseline instability and the electronic noise which degrades its efficiency for low photon rates. However, in the case of TCSPC, each single photon detection creates a pulse resulting in the registration of a photon count that will contribute to the final histogram. This yields a near-perfect counting efficiency therefore achieving optimum signal-to-noise ratio for a given number of detected photons compared to the analog transient recording alternative. Moreover, the fact that, in TCSPC, it is the number of photon counts that matters and not the detector signal's amplitude results in several advantages for the TCSPC compared to the analog recording technique. In particular, the TCSPC technique yields a higher bandwidth and a shorter instrument response function (IRF) for a given detector than any analog recording technique because the time resolution of photon counting is not limited by the width of the single electron response but by the accuracy at which the arrival time of the individual pulses can be determined. Furthermore, because the TCSPC method counts all photons with the same weight, the SNR is not influenced by the gain noise of the detector and the variation of the photon detector gain due to supply voltage variations or aging effects have only negligible influence on the results compared to the analog transient recording method where the magnitude of the recorded signal is directly affected by the detector gain. Still, it is important to remember that, as mentioned earlier, the advantage of TCSPC over analog recording stands only for periodic low intensity signals which should be such that the system can detect 0.1 to 0.01 photons per signal period. A higher count rate, the dead time of the detector would lead to a "pile-up" effect which translates into a substantial signal distortion due to the non-detection of the late photons hidden by the early one in the case were more than 1 photon should be detected per signal period (Figure 18).

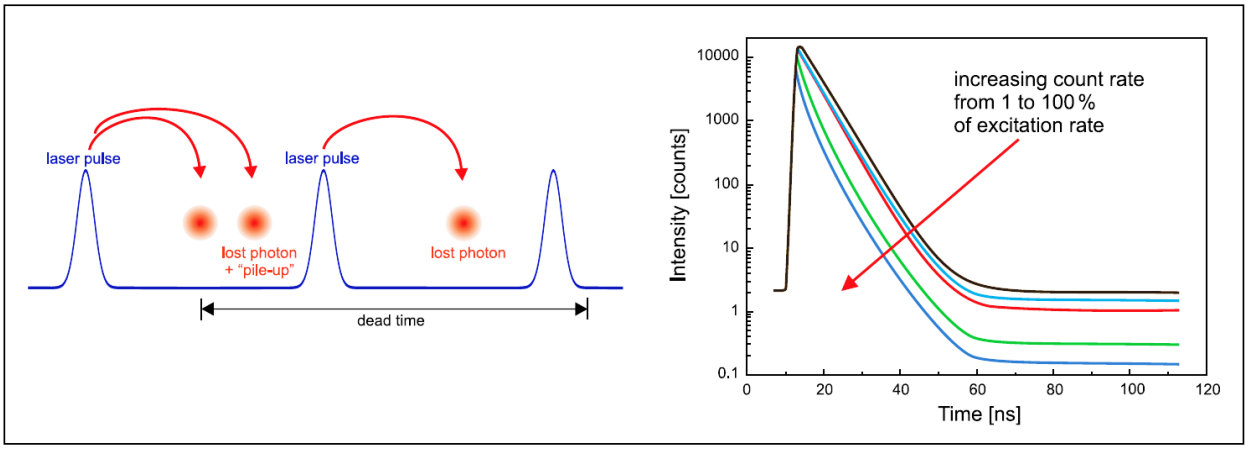


Figure 18 - Distortion of the TCSPC measurement by pile-up effect [48]

1.5.2 TCSPC Building Blocks

The main building blocks of a traditional TCSPC system includes a single photon detector, a discriminator, and a counter. The classic TCSPC setup adds a time to amplitude converter and replaces the counter with a combination of analog-to-digital converters and a memory block with an optional amplifier. A classic TCSPC setup is illustrated in Figure 19; the output of the single photon detector is connected to the discriminator. When a photodetection occurs, the detector delivers pulses to the discriminator that activates the timing electronics when it senses a valid pulse. One specific issue that arises with some types of detectors in single-photon detection mode is the fluctuating pulse amplitudes due to the quantum physical nature of the amplification process in these devices. The use of a simple voltage comparator to trigger the timing electronics would translate these amplitude fluctuations into timing errors resulting in a distorted output. It is therefore important to ensure that the discriminator is not affected by these fluctuations and that it is activated with minimum jitter when it senses a valid pulse. The output of the discriminator is connected to a time-to-amplitude converter and is used to stop a timer that calculates how much time has elapsed between the arrival of the photon and the reference excitation pulse which caused the photon. The other pulse that delivers the stop signal to the TAC comes from the reference signal. The reference pulse signal might be electronic and directly connected from the light source to the TAC input or may come from another detector/discriminator combination that monitors the output of the light reference source. Finally, the output TAC signal is sent to an analog to digital (ADC) converter and the output of the ADC is equivalent to the photon detection time.

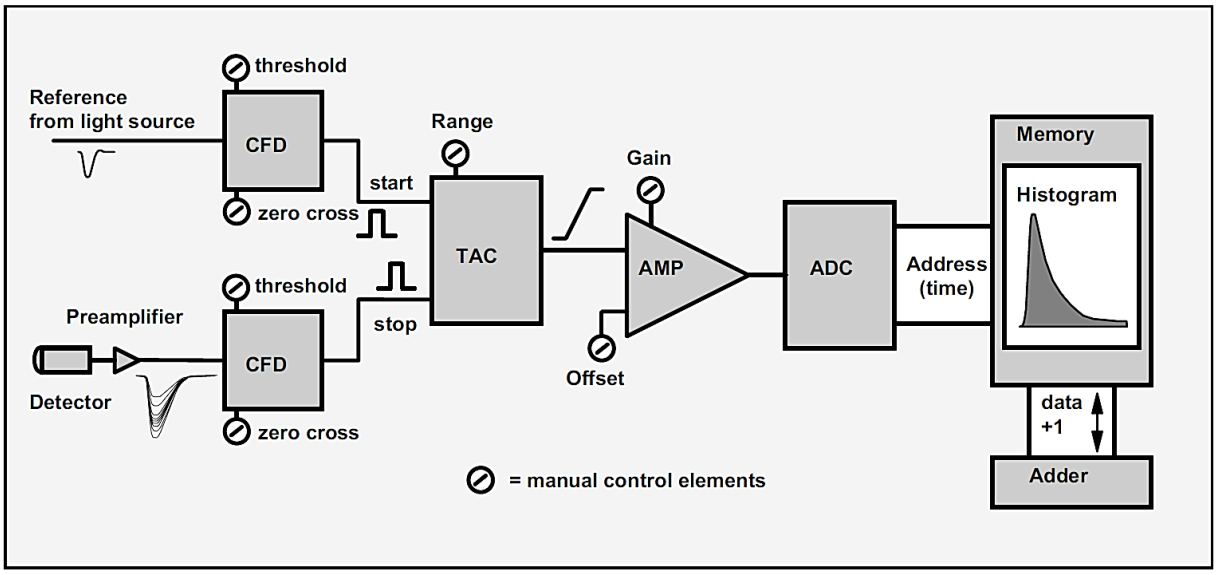


Figure 19 - Classic TCSPC setup [49]

1.5.2.1 Discriminators

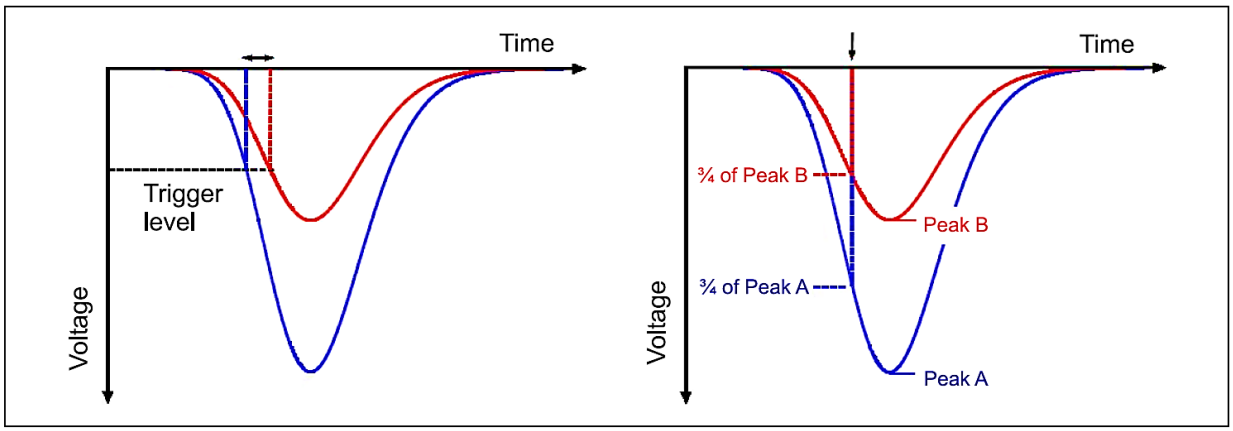


Figure 20 - Comparison between level trigger (left) and CFD operation (right). [48]

There are two main types of discriminators: trigger level discriminators and constant fraction discriminators. A trigger level discriminator is basically a voltage comparator that looks at the leading edge of the detector pulse and outputs a logic pulse if the pulse reaches a user-predefined threshold. The main problem with this type of discriminators is that pulses from some photon detectors like PMTs and APDs exhibit amplitude variations due to their inherent gain mechanism. These amplitude fluctuations are translated into a jitter in the rise time of the pulses thus triggering the leading edge discriminator at slightly different times depending on the amplitude of the pulse (Figure 20-left). This issue is known as time-walk and in the case of TCSPC systems where photon arrival time's statistics are accumulated to build of a temporal profile of the optical signal, these fluctuations are a critical problem that could affect the system precision. The triggering time dependency on the pulse's peak amplitude could be rectified by means of a constant fraction discriminator. A constant fraction discriminator (CFD) is more suited to solve the time-walk problem encountered with the level trigger discriminators as it triggers only when the input pulse reaches a user predefined fraction of these pulse peak value (Figure 20- right). The most common way of implementing a CFD is by comparing the original detector signal with an amplified and delayed version of itself. The signal derived from this comparison changes its polarity exactly when a constant fraction of the detector pulse height is reached. The zero crossing point of this signal is therefore suitable to derive a timing signal independent from the amplitude of the input pulse. This is done by a subsequent comparison of this signal with a settable zero level known as Zero Cross Trigger. Making this level fixable allows an adjustment to the noise levels in the given signal, since in principle an infinitely small signal could trigger the zero cross comparator. It should be noted that modern CFDs detect the vertex of each pulse and trigger on that point which is like applying a constant fraction of 1. Additionally, typical CFDs employs a discriminator level to set the lower limit the detector pulse amplitude must pass

in order to be considered valid thus making it possible to suppress random background noise pulses otherwise mistaken as photon detections.

1.5.2.2 Time Measurement in TCSPC Systems

In addition to detecting a single photon, a TCSPC system measures the time arrival of the photon within the light signal period. Traditionally this was done by means of a TAC followed by an ADC. TAC acts like a stop watch where one input initiates the start of the count while another one stops it. The output of the TAC is a voltage (or current) signal proportional to the time difference between these two events. In the case of TCSPC systems, one of the inputs to the TAC is obtained directly from the CFD while the other input is obtained from the reference. The reference sync pulses may be provided electronically by certain systems such as mode-locked lasers or other laser systems or by feeding the input laser source to a detector and then using a separate discriminator circuit to detect the start of the reference pulse. A basic TAC implementation consists of a current source and a capacitor, the start pulse turns the current source on thus charging the capacitor until the stop pulse switch the current source off. The voltage level retained by the capacitor following this procedure is proportional to the time difference between the start and the stop pulses. This voltage is then fed to an ADC that translates the analog level into a digital representation in order to be saved in a memory and processed later to reconstruct the photon arrival temporal histogram. The ADC must be very fast in order to limit the dead time and should also guarantee a very good linearity to avoid distorting the results. In recent TCSPC systems, the TAC and the ADC group are often substituted by a single fully digital circuit known as Time to Digital Converter (TDC). TDCs measure time intervals by means of delay times in digital logic gates, TDC architectures often employ one or several delay lines that englobes several identical delay elements to measure time intervals, they use the transit time of a pulse through chain of logic gates for time measurement. The most straight-forward technique to do so is based on counting the number of delay elements the start pulse crossed before the arrival of the Stop pulse. The use of a TDC makes it possible to perform time interval measurements with sub nanoseconds resolutions and allows us to directly obtain and process the digital results. They represent a practical solution that allow exceptionally compact and cost-efficient TCSPC systems implemented in Standard CMOS technology as Application Specific Integrated Circuits (ASICs) at low cost and high reliability.

The Time to digital conversion is far too complex to be resumed in this section and we will further elaborate later (Chapter 3) on the various time conversion schemes and architectures used to ensure linear and precise time interval measurements.

1.5.2.3 Detectors

There are essentially two detector classes available [49] detectors based on the external photoelectric effect (i.e., the generation of photoelectrons through a photocathode) such as photomultiplier tubes, microchannel plate photomultipliers, or hybrid photomultiplier tubes and 2) detectors based on the internal photoelectric effect such as single-photon avalanche diodes (i.e., the generation of photoelectrons inside the device). When comparing detectors there are five key parameters that must be considered to find the most suited model for the targeted application:

Sensitivity – The sensitivity represents the probability that a photon generates a measurable electrical pulse. It is usually given in percent and is referred to as the "quantum efficiency" or the "detection efficiency". The sensitivity of the detectors used in single photon counting is not uniform over the complete spectral range, it is therefore necessary to look at the sensitivity of each detector at the targeted detection wavelength (range) in order to make an appropriate choice.

Dark counts – dark counts are the equivalent of noise in single photon detectors. They represent the output pulses generated "inside" the detector in the dark. Dark counts are expressed in counts per second (cps); they are emitted at random times and cannot be avoided or removed. It is important to ensure that the dark count rate of the chosen detector is much lower than the expected signal rate otherwise the effective usable signal count rate will be reduced due to a competition between dark counts and "real" photon counts as they will result in a dead time during which the detector is incapable of detecting additional photon arrivals.

Afterpulsing – afterpulsing is similar to a dark count in that it doesn't result from photodetection, but unlike the dark counts, which are random, the afterpulsing is correlated to a previous detection event. Afterpulsing is usually expressed as a percentage that indicates the probability that one detected photon creates an afterpulsing event after a given time.

Timing resolution – The internal temporal resolution or "jitter" of the detector represents the time uncertainty between a photon arrival and the ensuing electrical output. Jitter is a crucial parameter for time-resolved applications, because a lower jitter translates into a better timing precision and a better overall temporal resolution of the complete TCSPC setup.

Size of the active area –Detectors based on the external photoelectric effect usually have active areas of several mm^2, whereas detectors based on the internal photoelectric effect only have active areas in the range between a few μm^2 to several thousands of μm^2. In the latter case, the optical setup must be designed in way that allows to couple the collected light from the sample effectively on the small active area (e.g., by means of optical fibers or using a confocal setup). Another way of representing this parameter is by means of the **Fill Factor**

which represents the percentage of the photosensitive area compared to the total detector area. A higher Fill Factor would obviously translate into higher detection efficiency as the photon will be more likely to be detected.

In the following subsections, we will concisely describe and discuss typically used photon detector types (photomultiplier tubes, microchannel plate photomultiplier tubes, hybrid photomultiplier tubes, and single-photon avalanche diodes).

Photomultiplier Tubes (PMTs)

Photomultiplier tubes (PMTs) are the traditional detectors for single-photon counting. The first PMT was already demonstrated in the mid-1930s, after intensive studies of the photoelectric effect and secondary emission multipliers (dynodes) [50]. A PMT is basically a vacuum tube that includes three core components; A photocathode, dynode stages and an anode. The photons are converted to electrons in the photocathode by the photoelectric effect and emitted into the vacuum. Depending on the material of the photocathode, PMTs can be effective for detection of light at varying wavelengths. The electrons are then multiplied in the dynode stage through secondary electron emission process. When an electron is emitted from the photocathode, it is accelerated toward the first positively charged dynode. The electron collides with the dynode and generates more electrons, which are then accelerated toward the next dynode, where they collide and creates even more electrons resulting in signal amplification as the increasing number of electrons collides with later dynodes Figure 21. At the end of this process, the multiplied secondary electrons emitted from the last dynode are collected by the anode.

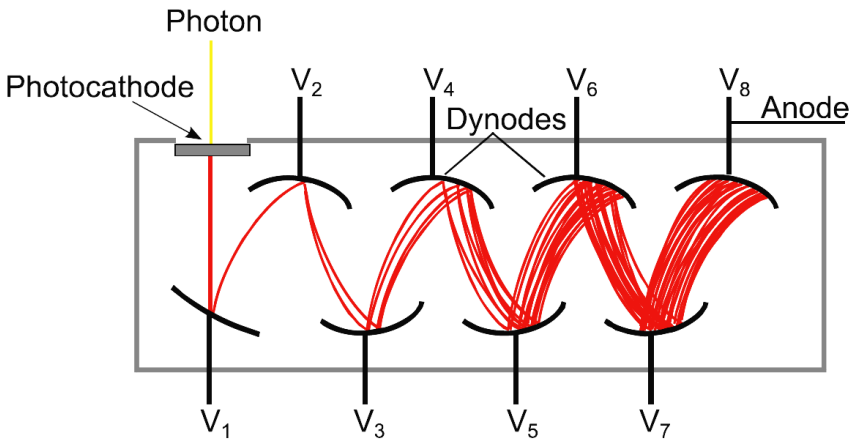


Figure 21 - Operation principle of a photomultiplier tube [51]

There are a variety of dynode types available, and each type exhibits different gain, time response, uniformity, and secondary electron collection efficiency depending on the structure and the number of stages. This amplification process is very effective and typically

leads to multiplication factors between 10^6 and 10^7 [51]. However, the dynode stages typically require operating voltages in the order of 1 kV. This, along with the necessary design of the multiple dynode stages, made PMTs rather large and bulky detectors although in the recent years, PMTs have been successfully miniaturized and made available as small compact units that include the necessary high-voltage power supply [52] [53]. PMTs are analog devices that output a current proportional to the light level on the photocathode. At very low light intensities the PMT outputs individual, well-separated pulses that can be amplified and further processed by the photon counting electronics. However, at high light intensity, the output pulses of individually amplified photoelectrons overlap and can no longer be detected as individual pulses. Another problem with PMTs is that the output pulses suffer from amplitude fluctuations due to variations in the amplification process of the dynodes. These fluctuations lead to a timing jitter of the order of the pulse rise time thus requiring the use of a constant fraction discriminator (Figure 22).

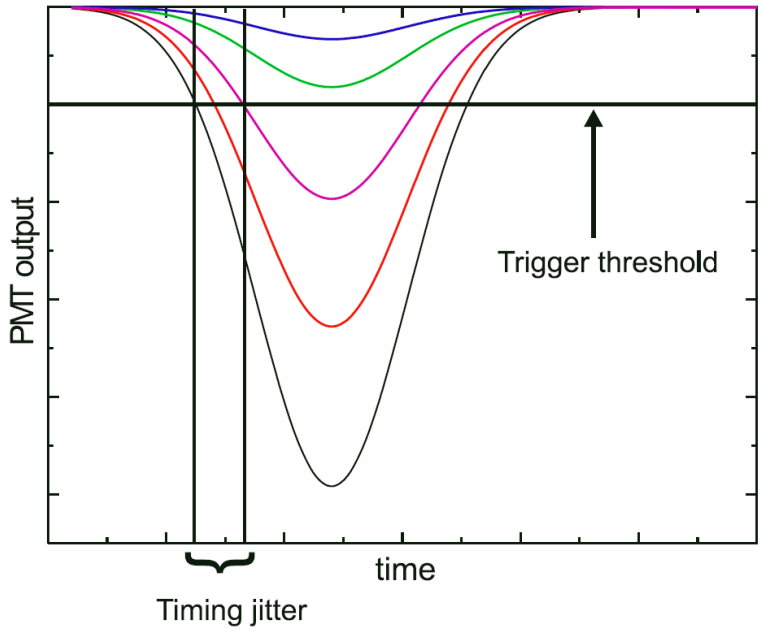


Figure 22 - Timing jitter in time resolved measurements due to output pulse amplitude fluctuation when using a simple level trigger [51]

Microchannel Plate Photomultipliers Tubes (MCP)

A Microchannel plate photomultiplier tube (MCP) is similar to that of a PMT [48]. An MCP comprises a large number of glass capillaries (channels) with an internal diameter between 6 and 20 µm. The inner wall of these capillaries is coated with a photoemissive material and biased at each end such that it acts as a continuous dynode. When a primary photoelectron hit the inner wall of a channel, secondary electrons are released. These electrons will collide again with the inner wall to release even more electrons, resulting in an exponential multiplication of the electron flux (Figure 23). MCPs require rather high operating voltages (~kV) but their gain is lower than that of conventional PMTs. MCPs suffer from many weaknesses: 1) they are fragile and easily damaged. 2) The channels need a certain time (µs to ms) to be recharged which limits the maximum count rate achievable by the device and like PMTs. 3) the output pulse of an MCP suffers from amplitude fluctuations and therefore needs to be connected through a constant fraction discriminator for time resolved photon counting measurements.

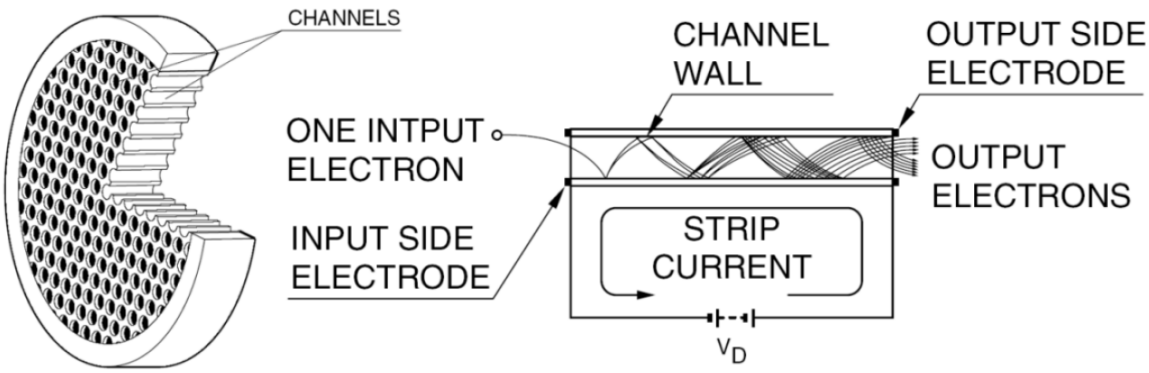


Figure 23 - MCP schematic and operating mechanism [54]

Hybrid PMTs

A hybrid PMT is a photomultiplier tube that incorporates a silicon avalanche photodiode in an evacuated electron tube (Figure 24). When light strikes the photocathode, photoelectrons are emitted and then accelerated by a high-intensity electric field of a few kilovolts applied to the photocathode. The photoelectrons are then "bombarded" onto the silicon avalanche photodiode where they create electron-hole pairs according to the energy of the photoelectron. These carriers are then further amplified by the linear gain of the avalanche diode. The total gain of a hybrid PMT is in the order of 10^5 which is lower than the gain of PMTs or MCPs. Hybrid PMTs require an operating voltage of 8 kV and low-noise amplification to deal with the small amplitude of the single-photon pulses [55]. They also require a carefully designed integrated cooling system in order to controls the temperature of the APD to avoid fluctuations in the gain and dark counts and like conventional PMTs the output needs to be connected through a constant fraction discriminator for time-resolved

photon counting measurements due to fluctuations in its output pulse amplitude. A Hybrid PMT offers a better temporal resolution than the conventional PMT.

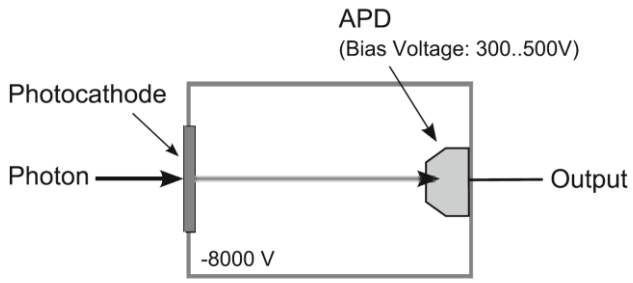


Figure 24 - Schematic of a hybrid PMT [51]

Single-Photon Avalanche Diodes (SPAD)

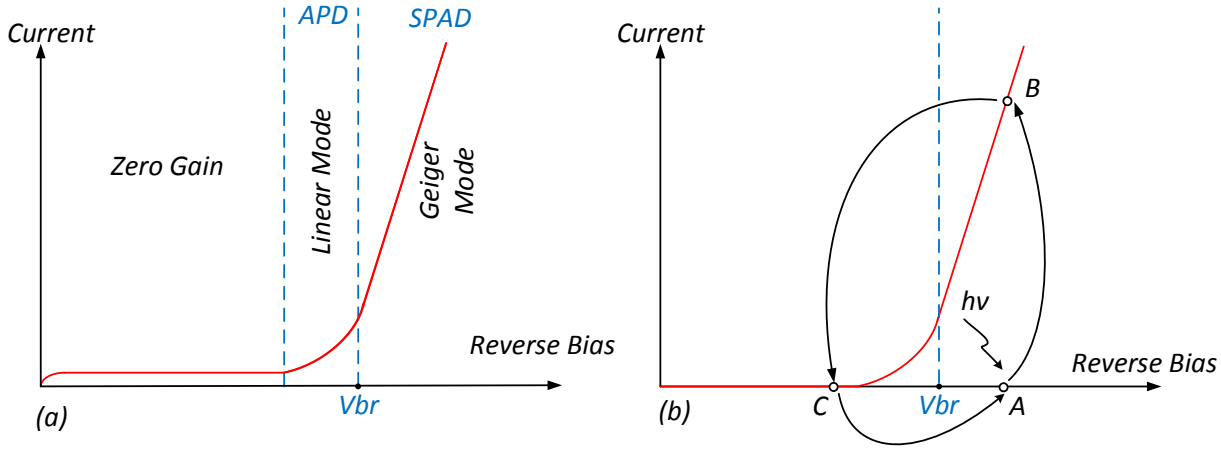


Figure 25 - The photodiode gaining modes (a) and the SPAD biasing cycle (b)

In contrast to photomultiplier tubes based on the external photoelectric avalanche photodiodes are based on the internal photoelectric. Avalanche photodiodes (APD) are PN junctions that are biased near but slightly below their breakdown voltage, as a result a high electric field exists in the multiplication region and a photo-generated carrier can start an internal avalanche by impact ionization with a finite internal gain M defined as the number of charge carriers generated by a single photon. In this mode known as the linear mode (Figure 25-a), the APD operates as a linear amplifier with a finite gain to the optical signal. The detection of single photons in the linear mode is possible although not very practical as linear APDs suffers of many drawbacks: a limited gain of few hundred at best, high noise levels since many factors (temperature, voltage, doping levels etc.…) can alter the internal gain, and long integration times making them unsuitable for ultra-fast low light detection. These limitations can be avoided by using the avalanche multiplication to obtain a virtually infinite gain. Unlike linear APD, a Single Photon Avalanche Diode (SPAD) operates in the "Geiger Mode" (Figure 25-a). In this mode, the reverse bias of the SPAD is well above its

breakdown voltage. As a result, a very intense electric field exists in the multiplication region and the arrival of a single photon can create a photo generated charge carrier capable of starting a self-maintained avalanche. The avalanche leads to a rise of the current to a macroscopic constant level within less than a nanosecond, which can then be easily detected by suited electronics. If the primary carrier is photogenerated, the leading edge of the avalanche pulse marks the arrival time of the detected photon with picosecond time jitter. The avalanche current will keep flowing as long as the applied voltage is kept above the breakdown voltage. In this stage absorption of additional photons will not lead to any change in the signal output, making the device unable to detect additional photon arrivals. The self-sustained avalanche must be stopped and the detector must be reset in order to detect another photon (Figure 25-b). This process is called avalanche "quenching", it requires the use of a "Quenching circuit" to take in charge of detecting the leading edge of the avalanche, then reducing the bias voltage below the breakdown level to stop the avalanche and finally restores the voltage to its initial value so that a new photon could be detected.

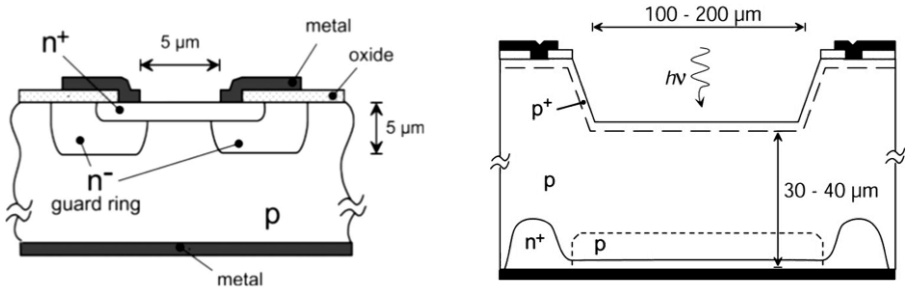


Figure 26 - Cross section of a thin-junction (left) and a thick-junction (right) SPAD devices [56]

In general, SPADs have being fabricated in both Standard and custom CMOS technologies but the performance of standard devices is still far from the one achieved with custom technology devices. Indeed compared to custom technology devices, standard devices show a lower PDE [51], a significantly higher DCR per unit area, and a higher afterpulsing probability, they are often designed with smaller active areas and low excess bias voltages in order to limit the detector noise. SPADs architecture are divided into 2 main types (Figure 26): thick-junction devices and thin-junction devices. The main difference between these two designs is the thickness of the depletion region in which photon absorption takes place with a depletion region of a few tens of µm for thick-junction SPADs, thin-junction SPADs on the other had usually have a few µm thickness. SPADs fabricated in silicon can generally be used in the spectral range between 400 and 1,100 nm and thin-junction SPADs typically have lower detection efficiency than thick-junction devices because they have smaller absorption region (Figure 27). However, Thin-junction SPADs are implemented in planar epitaxial technology which makes them compatible with CMOS circuits and offer the typical advantages of microelectronic devices. They are therefore gaining wide acceptance, and can already advantageously substitute PMTs in many cases

especially with their higher photon detection efficiency (PDE), lower power consumption, and the possibility to be integrated in arrays

.

The subject of SPADs will be further elaborated in Chapter 2 of this manuscript where we will present and discuss their operation principle, types, structures, figures of merit and the associated quenching electronics.

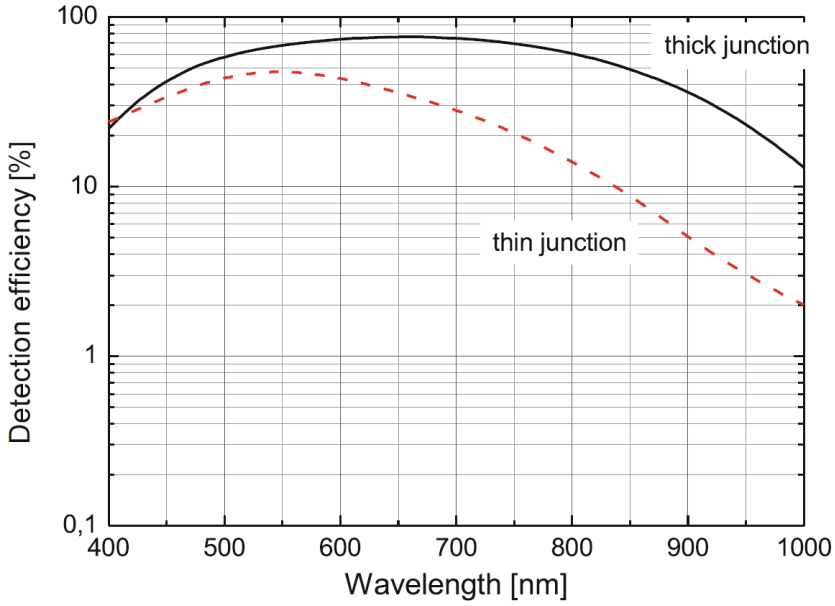


Figure 27 - Typical detection efficiency of a thick-junction SPAD and a thin-junction SPAD [51]

1.6 This Thesis

1.6.1 Aim

Time-correlated single photon counting (TCSPC) is a technique that offers both single photon sensitivity and picoseconds temporal resolution. It is widely used in many fields of science, medicine, engineering, industry and security for applications requiring the precise measurement of single-photon arrival times. These applications include but are not limited to Fluorescence Lifetime Imaging (FLIM) [57] [58] [59] [60] and Time of Flight measurements (TOF) [61] [62] [63] [64] [65]. Traditionally, the TCSPC technique relied on vacuum tube technologies such as PMTs and MCPs. These mature technologies are capable of achieving very good performances but they are also expensive, cumbersome, fragile, and require extremely high operating voltages which makes them unsuitable for the fabrication of miniaturized portable TCSPC imaging systems. In recent years, SPADs have gained a wide acceptance as a less expensive more compact alternative for vacuum tube detectors. High-performance SPAD devices that exhibit high photon detection probabilities (PDE) with high timing resolutions in the range of several tens of picoseconds, have been implemented in dedicated technologies. But these devices are more expensive, they offer lower levels of miniaturization resulting in larger parasitics that can limit their performances. The integration of planar epitaxial SPADs in Standard CMOS technology has significantly improved the level of miniaturization of SPADs and paved the way for SPAD arrays. These devices possess the typical advantages of microelectronic devices, such as small size, ruggedness, low operating voltages, and low cost. Furthermore they can be directly implemented with the necessary associated circuits on the same chip to realize an integrated, ultra-sensitive, high-speed, and low-cost TCSPC imaging system. Many SPAD based TCSPC systems have been successfully demonstrated lately .Nowadays, state-of-the-art imaging sensors integrating thousands of single-photon detectors on the same chip have been demonstrated in standard CMOS technology [66] [67]. Most integrated TCSPC systems consist of 2D-arrays that are limited by their architecture and mode of operation due to space constraints. The work done in the scope of this thesis aimed to design and demonstrate an integrated TCSPC Streak camera (TCSPC-SC) comprising of a 1D pixel array. A 1D-array TCSPC structure makes it possible to reach the physical limits of the technology by releasing the space constraints that arise from the 2D array implementation [65] [62] [68] in order to achieve the best possible sensitivity. Presently, only one single-photon integrated streak camera was reported in the literature [63] but it is designed around a common and single delay line which requires a gating operation mode that limits the overall repetition rate and the sensitivity. Moreover, no TDC were embedded in the sensor. These drawbacks were resolved in the architecture demonstrated in this work.

1.6.2 Organization

An integrated TCSPC system consists of 2 main parts: the Single Photon Avalanche Diode (SPAD) and a time measurement units by means of Time to Digital Converter (TDC). Hence" this manuscript is organized as follows:

In Chapter 2 we present the SPADs, their operation principle, their figures of merit and their physical implementations in CMOS technology. Next we show the work done in this thesis to design and characterize a SPAD structure in order to be used as detector in the TCSPC-SC and the measurements results of the used SPAD structure.

In chapter 3 we elaborate on the TDCs by presenting and discussing their operation characteristics and the different TDC architecture found in the literature. Then we present the hybrid TDC architecture specially designed to be used as a measurement unit in the TCSPC-SC with the preliminary characterization results.

In chapter 4 the architecture of the TCSPC-SC is presented along with the operation principle and the work done to optimize the system's efficiency based on behavioral and mathematical modeling.

Finally we conclude in chapter 5.

Chapter 2
Single Photon Avalanche Diodes

The detection of weak ultra-fast signals is required in many applications from different fields such as nuclear physics, chemistry, biology and astronomy. Such signals can be detected using Photomultiplier Tubes (PMT), Charged Coupled Devices (CCD) and Avalanche Photodiodes (APD). In the last decades many work have been done in developing Single Photon Avalanche Diodes (SPAD) in particular as replacement to PMT, novel designs and advancement in the CMOS technology process has allowed the fabrication of SPADs with characteristics comparable to those of PMTs however unlike PMTs which are bulky, fragile, expensive and require the use of high voltage, SPADs are less expensive, requires moderate reverse biases and can be integrated with their front-end circuitry on the same substrate making for smaller and easier to manipulate sensors.

2.1 Avalanche Photodiodes in Geiger Mode

Avalanche photodiodes (APD) are PN junctions that are biased near but slightly below their breakdown voltage mode, as a result a high electric field exists in the multiplication region and a photo-generated carrier can start an internal avalanche by impact ionization with a finite internal gain M defined as the number of charge carriers generated by a single photon. M can be theoretically increased by widening the multiplication region however this causes an increase in the noise level and a reduction of the detection bandwidth resulting in the achievable internal gain to be limited by a maximum value defined as [69]:

$$M = {}^{\alpha}/_{\beta} \tag{4}$$

With α is the ionization coefficient of electrons and β the ionization coefficient of holes where the ionization coefficient defined as the ability of an electron or hole to generate a carrier by ionization. In this mode known as the linear mode, the APD operates as a linear amplifier with a finite gain to the optical signal. The detection of single photons in the linear mode is possible although not very practical because linear APDs suffers of many drawbacks: a limited gain of few hundred at best [70], high noise levels since many factors (temperature, voltage, doping levels, etc.) can alter the internal gain, and long integration times making them unsuitable for ultra-fast low light detection.

SPADs are capable of overcoming these limitations by using the avalanche multiplication to obtain a virtually infinite gain. Indeed unlike linear APD, SPADs operate in the "Geiger Mode". In this mode, the reverse bias of the SPAD is well above its breakdown as result a very intense electric field exists in the multiplication region and the arrival of a single photon can create a photo generated charge carrier capable of starting a self-maintained avalanche resulting in a large current flowing through the junction until the avalanche is quenched and the SPAD is returned to its initial state otherwise it will overheat and get destructed. Quantitatively, the photodiode is considered operating in Geiger mode when it meets the condition [71]:

$$1 \leq \int_0^w \alpha . \exp\left[\int_0^x (\alpha - \beta)\, dx'\right] dx \qquad (5)$$

Where *w* is the width of the multiplication region. The bias at which the equation is equal to 1 defines the breakdown voltage.

2.1.1 The SPAD Biasing Cycle

In Geiger mode, the SPAD is biased well beyond its breakdown voltage and unlike linear APDs, where stopping the light signal is enough to stop the avalanche, when an avalanche is triggered in an SPAD the current will continue to increase until the destruction of the component as a result of overheating. Therefore the avalanche must be swiftly quenched by an associated circuitry that senses the avalanche and stops it by reducing the reverse bias below the breakdown voltage so that the avalanche cannot maintain itself then returned it to its initial condition. The SPAD polarization cycle can be described in three main phases (Figure 28):

A. **Charge**: The SPAD is initially biased above its break down voltage Vbr.
B. **Avalanche**: The arrival of a photon triggers an avalanche resulting in a swift appearance of an avalanche current
C. **Quenching**: To stop the avalanche, the SPAD biasing voltage is lowered below its breakdown voltage, the SPAD biasing voltage remains at this level until the avalanche is stopped, a time interval known as the hold-off time, then the biasing is restored to its original value (A).

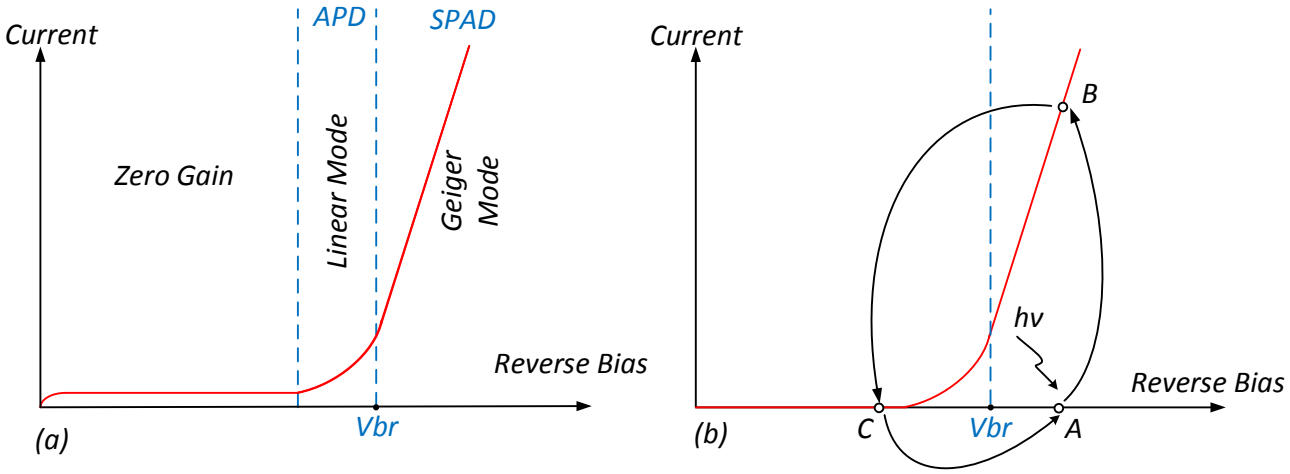


Figure 28 - The photodiode gaining modes (a) and the SPAD biasing cycle (b)

2.1.2 Figures of Merit for Individual SPADs

When measuring the performances of an SPAD, there are several key parameters that must be considered in order to characterize the SPAD in terms of noise, sensitivity, and timing resolution. This is accomplished by means of the following parameters.

2.1.2.1 Dark Count Rate

SPADs have a binary output indicating that a photon was detected (on) or that there is no photo-detection (off). It is possible that an avalanche is triggered by other mechanisms then the photo generation. To quantify this phenomenon we use The Dark Count Rate (DCR) which represents the ratio of faulty detection in the dark expressed by counts/s. The DCR is due to several triggering mechanisms such as thermal generation, trap assisted generation and tunneling. It depends on many factors: Temperature, excess bias voltage, the SPAD structure and the fabrication process [56]. Thermal generation is the main generation factors at ambient temperature and it increases exponentially with temperature. The contributions of tunneling and trap assisted generation on the other hand increase as the temperature decreases and are very dependent on the doping levels and the fabrication process (Generation/Recombination centers and impurities concentration). In brief, DCR increases with the excess bias, temperature and the SPAD active area and although it's impossible to determine the exact contribution of each factor in the global DCR, a high breakdown voltage usually results in a high contribution of thermal generation while a low breakdown voltage correlated with a high tunneling effect.

2.1.2.2 Afterpulsing

Primary DCR is random and statistically uncorrelated but SPADs display also secondary avalanche pulses statistically correlated to primary avalanches known as afterpulsing. Afterpulsing is due to trapping and de-trapping of charges flowing through the

multiplication region during the primary avalanches. It is affected by the avalanche current's density, the avalanche duration, the associated electronics as well as temperature and can be reduced by increasing the SPAD hold-off time to allow the releasing of the trapped charges before activating the SPAD again. Afterpulsing is usually expressed as a percentage that indicates the probability that one detected photon creates an afterpulsing event and it can be determined by measuring the inter-arrival times of individual dark pulses for different hold-off times.

2.1.2.3 Photon Detection Probability

Photon Detection Probability (PDP) represents the ratio of the number of detected photons over the number of incident photons. PDP is the result of the fill factor, the absorption probability and the triggering probability [56] [72]. The ideal 100% PDE is unattainable in reality due to self reflection, absorption and self-quenching [73]. The PDP is a function of the signals wavelength, it is usually non null over a wavelength range between 350nm and 950nm with a peak value that depends on the junction structure as deeper junctions are more sensitive to the longer wavelengths as opposed to shallow junctions which more suitable for near UV signals. PDP depends on the active area, the noise and the doping levels [71]. It increases with the excess bias but such improvement can be mitigated by an increase of the DCR.

2.1.2.4 Timing Resolution (Jitter)

Timing resolution or jitter represents the timing incertitude due to statistical variation of delays between photon absorption and the subsequent avalanche detection by the SPAD's associated electronic. The timing resolution is improved with higher excess biases as a higher electric field implies that carriers drift faster from the absorption point to the multiplication region and triggering probability increases. Furthermore SPADs with narrower and smaller depletion region offers better timing resolution as they represents better timing due to lateral and vertical dependence of timing uncertainty on photon absorption location [74]. Timing resolution is characterized using a TCSPC setup to measure the photon arrival time distribution, it is measured at full-width-half-maximum (FWHM) of the distribution that usually shows a sharp Gaussian peak from photons absorbed in the multiplication region with a tail that is due to photons absorbed in the neutral regions below the multiplication region and that is more marked for higher wavelengths.

2.1.3 Figure of Merit for SPAD Arrays

2.1.3.1 Fill Factor

The Fill Factor (FF) represents the ratio of the photosensitive area in the pixel to the total pixel area. The Fill Factor depends on the SPAD structure for instance the use of

Shallow Trench Isolation (STI) guard rings results in better fill factor compared to other guard ring structures. Complex pixels with additional functionalities results in a lower Fill Factor compared to pixel with basic electronics limited to a passive quenching resistor another approach is putting the pixel's associated electronics around the SPAD array to maximize the Fill Factor without sacrificing the pixel complexity.

2.1.3.2 Pixel Pitch

Pixel pitch represents the distance between the centers of neighboring pixels. Pixel pitch is directly related to the detector's spatial resolution, a larger pixel pitch results in a higher spatial resolution.

2.1.3.3 CrossTalk

Cross talk represents the probability that an avalanche triggered in a pixel is correlated to an avalanche in another neighboring pixel. Cross talk is due to two main mechanisms:

Electrical cross talk: A charge carrier generated in a pixel can travel to another nearby multiplication region and start another avalanche. Electrical Cross talk can also be triggered by electromagnetic coupling between the SPAD signal lines and the power supply or in some cases switching noises. Electrical cross talk can be eliminated using appropriate physical and electrical isolation mechanisms such as guard rings and the proper decoupling of signal lines [75]

Optical cross talk: Optical cross talk occurs when secondary photons are emitted by electroluminescence during an avalanche then captured by nearby SPADs resulting in secondary avalanches correlated to the same primary photo detection event. Optical cross talk is greatly affected by the avalanche current density; it can be reduced using proper electronic circuitry to limit the parasitic capacitance and the charge flowing during the avalanche. The use of trench isolation can also reduce optical cross talk but it results in a lower Fill Factor thus decreasing the Photodetection efficiency of the SPADs.

2.1.4 Quenching Electronics

In Geiger mode when an avalanche is triggered, the avalanche current rises swiftly and continues to flows through the device making it impossible to detect new photons. It is therefore important to (i) stop the avalanche by lowering the breakdown voltage so that the avalanche can no longer sustain itself and (ii) restores the photodiode to its original biasing state so that a new photon can be detected. The first task is the quenching process and the time duration between the avalanche start and its quenching is the quenching time. The

second task is the reset process and the time needed to restore the device original bias is the reset time. The sum of the quenching time and the reset time represents the dead time of the SPAD meaning that the SPAD is blind to any other photon arrivals. The circuit used to accomplish these tasks is the quenching circuit and the selection of such circuit is not a trivial task as it directly affects many of the SPAD performance metrics (Maximum count rate, afterpulse, timing resolution, etc.). It is therefore important to choose a suitable quenching circuit for the desired application so it won't limit or deteriorate the SPAD characteristics. It's also important to have a quenching circuit with a fast response, a short dead time and a reduced parasitic capacitance. Quenching circuits are divided into 3 types: passive, active and mixed based on the components they employ to accomplish the Quenching and reset process.

2.1.4.1 Passive Quenching

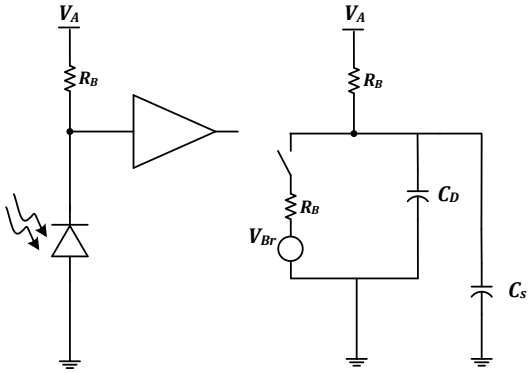


Figure 29- A basic Passive Quenching Circuit and its equivalent circuit [76]

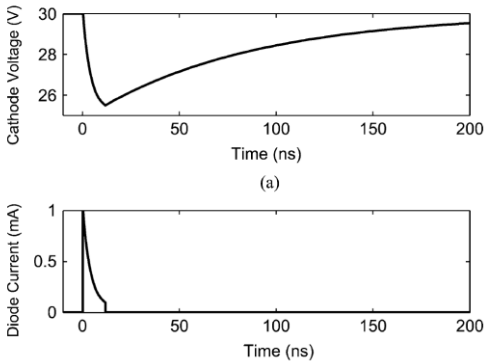


Figure 30 - (a) Cathode voltage and (b) Diode current waveforms for a SPAD connected to a Passive Quenching Circuit [76]

The simplest quenching circuit is the Passive Quenching Circuit (PQC) shown in Figure 29 with its equivalent circuit [76]. The passive quenching scheme makes use of the fact than an avalanche current self-quench by developing a voltage drop on a high value impedance load. The SPAD is polarized through a ballast resistance R_B of 100 kΩ or more, C_D is the intrinsic capacitance and C_s is the stay capacitance to the ground. R_D is the SPAD series resistance equal to the sum of the space charge resistance and the ohmic resistance of the neutral region passed by the avalanche current. V_{Br} is the breakdown voltage and V_A is the bias voltage. When an avalanche is triggered, the current (Figure 30) quickly rises to the maximum value of the avalanche current defined as:

$$I_{0=}{}^{V_{Ex}}/_{R_D} \quad (6)$$

Where V_{Ex} is the excess bias and R_D is the SPAD series resistance. As this current discharges the parasitic capacitance C_D of the SPAD cathode the voltage bias decreases exponentially with a time constant:

$$\tau = C_D \times (R_D // R_B) \quad (7)$$

It remains however above the break down voltage, consequently the avalanche is not directly quenched and the current reaches a final value:

$$I_{F=}{}^{V_{Ex}}/_{R_B} \quad (8)$$

If I_F is small enough the avalanche will self-quench otherwise it will be self-sustaining, I_s=100 µA is considered as the quenching threshold a value that defines if the avalanche is self-sustaining or self-quenching. The quenching time defined as the time elapsed until the current reaches the quenching threshold is determined as:

$$T_Q = \tau . ln\left(\frac{I_0 - I_F}{I_S - I_F}\right) \quad (9)$$

And the overall quenching charge is defined as [76]:

$$Q = V_{Ex} . C_D . \left[1 + \frac{I_F . ln\left(\frac{I_0 - I_F}{I_S - I_F}\right) - I_S}{I_0}\right] \quad (10)$$

The avalanche charge is directly proportional to the parasitic capacitance C_D which needs to be reduced in order to limit the avalanche charge this can be done by directly integrating the ballast resistor with the detector. PQCs are simple, low-cost and compact however they suffer from a long and not well-fixed quenching time, as well as a long reset time which limits the counting rate and increases the possibility of an avalanche triggering during the reset phase. This leads to high afterpulsing probability, increases optical crosstalk and lowers the timing resolution due to the uncertainty of the quenching process.

2.1.4.2 Active Quenching

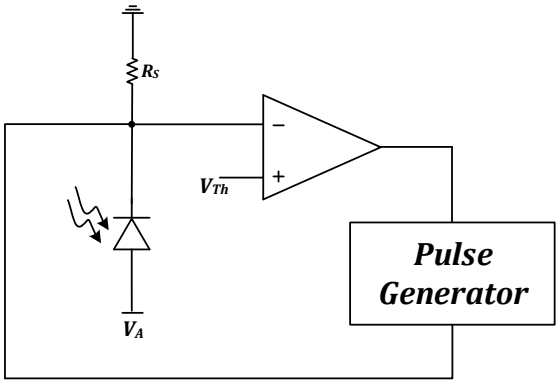


Figure 31 - Basic Active Quenching Circuit [76]

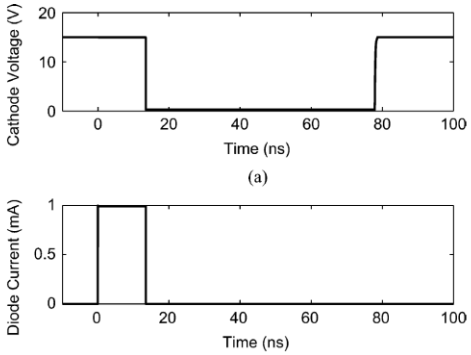


Figure 32 - Cathode voltage (a) and diode current waveforms (b) for a SPAD connected to an Active Quenching Circuit [76]

To overcome the limitations of the PQC in particular the long and variable dead time and the slow voltage reset, S. Cova introduced a new Quenching scheme [77] [78] based on sensing the avalanche pulse and forcing the quenching and reset transitions in short times using active components such as pulse generators or fast active switches. A figure of an Active Quenching Circuit (AQC) using a pulse generator is represented in Figure 31, the cathode voltage and the diode current waveforms are shown in Figure 32. When an avalanche is sensed the bias voltage is actually lowered below breakdown therefore the quenching time is not affected by fluctuations due to the statistical process of avalanche multiplication. Once the avalanche is quenched, the bias voltage is kept below breakdown for a fixed period known as the hold-off time and then restored to its original value. These circuits offer quicker quenching and reset phases resulting in shorter overall dead time and considerable reduction in the probability of avalanches occurring during the reset transition but this comes at the price of more complex circuitry and larger occupation area resulting in a lower fill factor. Furthermore, AQCs suffer from an intervention delay which is the time needed until the circuit senses the avalanche pulse and starts the quenching process. During this period the avalanche current continues to flow at maximum value (I_0) resulting in a high Avalanche Charge given by the maximum current value times the quenching time (T_Q):

$$Q = I_0 . T_Q \tag{11}$$

This leads to high power dissipation and increases the afterpulsing probability due to trapped charges. The integration of the AQC directly with the detector helps reducing the intervention delay and as a result the avalanche charge, it also limits the parasitic capacitance thus reducing the time required for voltage reset. The reset duration must be carefully chosen so as to be the shortest time needed for voltage reset, the hold-off time should also be long enough for the trapped charges to be released in order to limit the afterpulsing probability.

2.1.4.3 Mixed Active-Passive Quenching

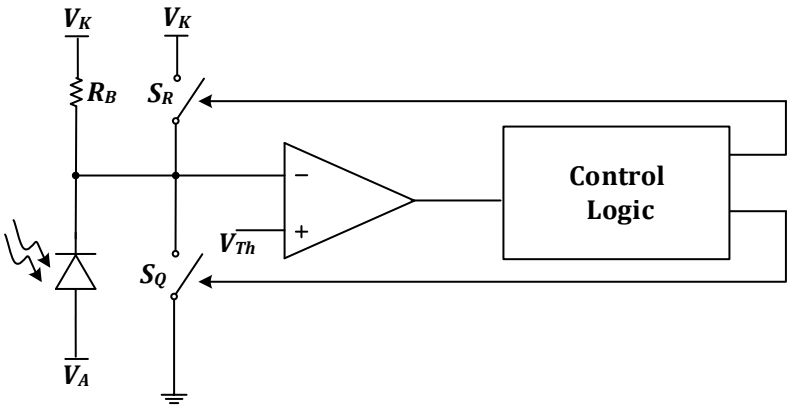


Figure 33 - Basic diagram of a mixed active-passive quenching circuit [76]

The mixed active-passive quenching approach combines the PQC and the AQC schemes to overcome their limitations, it includes many subtypes but the best results are obtained using mixed quenching active reset. A basic diagram of a mixed quenching active reset circuit using switches is represented in Figure 33, in this approach when an avalanche is triggered the avalanche current flows through the ballast resistor thus the quenching starts as fast as in the PQC then once the active circuitry enters in action the circuit operates like an AQC: the avalanche is actively quenched by keeping the bias voltage below its breakdown value then an active reset restores the initial biasing value (Figure 34). In this scheme the dead time is reduced and well-defined, the voltage reset is short limiting furthermore the avalanche charge is limited thus improving the power consumption and the afterpulsing probability. Its main drawback is a large area of occupation making it unsuitable for large SPAD arrays.

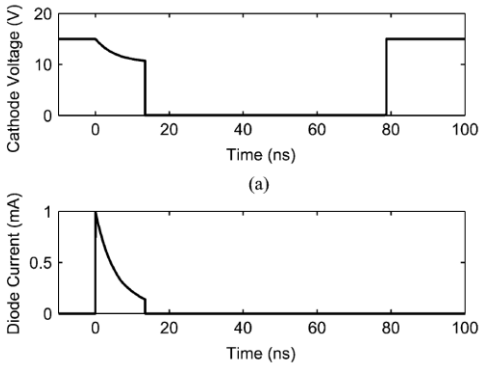


Figure 34 - (a) Cathode voltage and (b) Diode current wave forms for a SPAD connected to a Mixed Active-Passive Quenching Circuit [76]

2.1.5 SPAD progress and physical implementation

2.1.5.1 Early SPADs

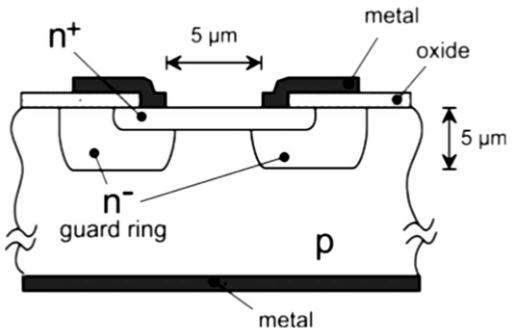


Figure 35 - Cross section of the first planar SPAD structure devised by Haitz [56]

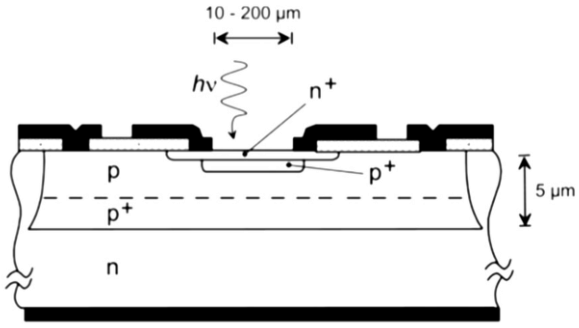


Figure 36 - Schematic cross section of the double epitaxial SPAD device structure [56]

Detection of single photons by means of PN junctions was first achieved in the early 60's at Shockley laboratory and the SPAD implementation realized by Haitz is considered to be the ancestor of modern planar SPAD devices. The device (Figure 35) was fabricated with a planar technology suitable for monolithic integration with circuits and other detectors; it operated at low voltage resulting in limited power dissipation during the avalanche. Nonetheless the fabricated device performance was limited by the existing technology and it represented many drawbacks in particular a low photon detection efficiency and a time response curve showing a long wavelength dependent tail due to minority carriers' diffusion from the neutral region. An improvement in the photon detection efficiency was achieved with the realization of the first epitaxial SPAD structure [79] fabricated in a fairly thin P-silicon epitaxial layer on an N-silicon substrate resulting in higher quality silicon (Figure 36). The timing response was also improved due to the epi-substrate PN junction that limits the depth of the neutral region from which photogenerated minority carriers can diffuse into the depletion region [56]. The first epitaxial device inspired the design of a new SPAD with a double epitaxial structure forming 2 diode junctions. The top epitaxial layer was used to form an ultra-clean n+/p junction, a p+ implant in the center of the structure served as a guard ring to reduce edge break down and a p+ implant below the top epitaxial layer was used to limit the large resistivity of the structure. The timing uncertainty was improved by means of the buried diode that prevented photogenerated electrons to enter the multiplication region. The double epitaxial device had a breakdown bias voltage of 50 V and a jitter performance of 55 ps at FWHM, it was later revised and tested [80] [81] and a better timing response with no diffusion tail was successfully obtained at the price of a more complex fabrication process.

The next big step was the demonstration of the first SPAD fabricated in a standard CMOS technology by Rochas in 2003 [82], the device was implemented in a commercial High Voltage 0.8 µm process by Austrian Micro Systems (AMS). Rochas HV-CMOS SPAD was based on Haitz structure, it showed good performance with a 50 ps timing resolution but had a low fill factor. This finding opened the way to compact and cost effective large SPAD arrays with the first 8×4 [83] [84]and 32×32 [65] SPADs image sensors as well as linear sensors [85] [86].

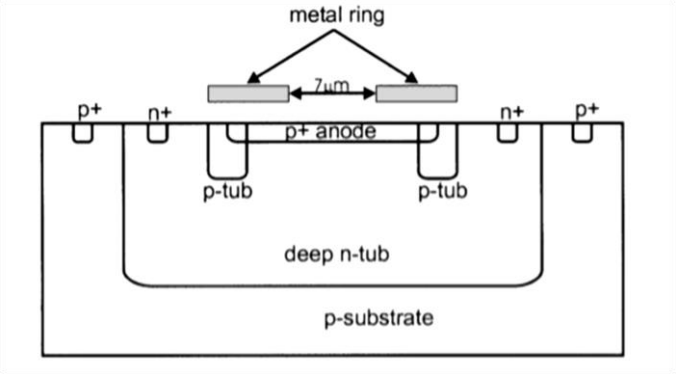


Figure 37 - Schematic cross section of the first CMOS SPAD devised by Rochas [82]

2.1.5.2 Technological requirements

In order to be used as a SPAD, the device structure must fulfills some basic requirements [87] : (1) The DCR and the afterpulsing probability must be as low as possible, this means that the thermal carriers generation and trapping should be minimized which requires a clean fabrication process to guarantee a very low density of defects and impurities that creates deep levels in the silicon gap acting as generation/recombination and afterpulsing centers. (2) The structure must show a uniform breakdown over the whole active area this means that the breakdown voltage and the electric field in the multiplication region must be as uniform as possible. (3) The electric field in the depletion region should be low enough to avoid band to band tunneling and limit premature edge breakdown, according to [56] this means that the device breakdown voltage should be higher than 20 V. These requirements are dependent on the structure used for the SPAD implementation where a guard ring is often commonly used to isolate the PN junction, guarantee a uniform electric field and prevent premature edge breakdown. Five commonly used structures for premature edge prevention [86] are shown In Figure 38; the grey line in each schematic represents the limit of the multiplication region. In structure *a* the n+ layer is used to maximize the electric field in the center of the diode, structure *b* and *c* [88]aims to limit premature edge breakdown by lowering the electric field at the edge of the multiplication region. The use of an STI guard ring in structure *d* results in an intense and uniform electric field but this approach suffers from a very high noise due defects introduced by the STI fabrication process. A possible solution to limit the noise is by inserting several layers of doped semiconductor material with

decreasing doping levels from the trench to the multiplication region thus isolating the junction from the STI by forcing carriers generated there to recombine before reaching the multiplication region [88]. Finally, in structure *e*, a deep p-well and deep n-well establishes a PN junction with an implicitly defined lightly doped guard ring [89]

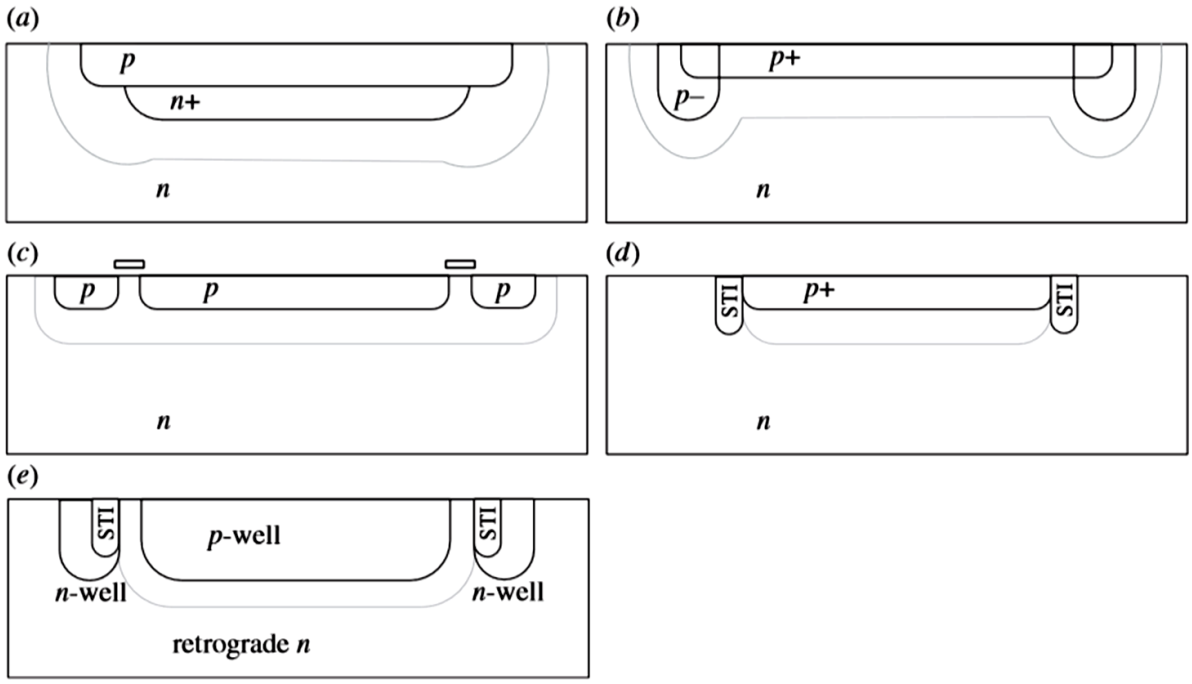


Figure 38 - Prominent SPAD structures for premature edge breakdown prevention [86]

2.1.5.3 Fabrication technology: CMOS vs Custom technologies

In the last decade many work have been done to optimize SPAD technology and numerous SPADs have been implemented in both dedicated and standard technologies. SPAD fabricated in custom technologies [75] have been able to reach great performances in terms of noise and very good efficiency from the ultra-violet to the near infra-red wavelength ranges. However SPAD fabricated in custom technologies cannot be monolithically integrated with their associated electronics. This results in a large module size, high power dissipation, and long dead time. SPAD fabricated in standard CMOS technologies are limited by the used implementation process, the implemented structures cannot be specially tailored to optimize the SPAD performance figures. Nevertheless the implementation of SPAD in standard CMOS technologies led to the reduction of the avalanche charge thus lowering the afterpulsing probability and optical and electrical crosstalk. Furthermore it allowed the reduction of dead time values from micro to nanoseconds thus increasing the achievable maximum count rates and operation frequencies. The first CMOS SPAD introduced by Rochas [82] in 0.8 µm technology was followed by a second generations of SPADs implemented in 0.35 µm HV CMOS technology. The scaling trend continued by implementing SPADs in Deep sub-micrometer (DSM) technologies with the demonstration of a SPAD implemented in

180 nm [90] [91], in 130 nm [92] both by Niclass et al. and in in 90 nm CMOS SPADs [93] [94]. The Scaling trend is especially beneficial because it allows the fabrications of larger models and lowers the avalanche charge thus improving some related performance figures such as cross talk, afterpulsing and dead time. Design of SPADs in DSM technologies results in additional as a result of higher doping levels that increase DCR due to increased band to band tunneling and a lower depletion width that limits the PDP.

2.2 SPAD Structures and Characterization Set Up

A high quality SPAD should offer a low DCR and a high PDP, this requires a structure design that allows a uniform high electric field in the multiplication region and limits the premature edge breakdown effect [88]. In order to find the best SPAD structure to be used as a detector in the TCSPC-SC, 6 different SPAD structures were fabricated with 8 active area diameters ranging from 5 µm up to 40 µm. The SPADs had an octagonal shape in order to ensure a uniform electric field distribution and to limit the premature edge breakdown phenomenon. These 48 SPADs were integrated in a standard 180 nm CMOS technology they were implemented and characterized using a versatile characterization circuit, the architecture of the circuit as well as the characterization results will be presented in the following section.

2.2.1 SPAD Structures

2.2.1.1 Structure A (P+ Nwell with STI guardring)

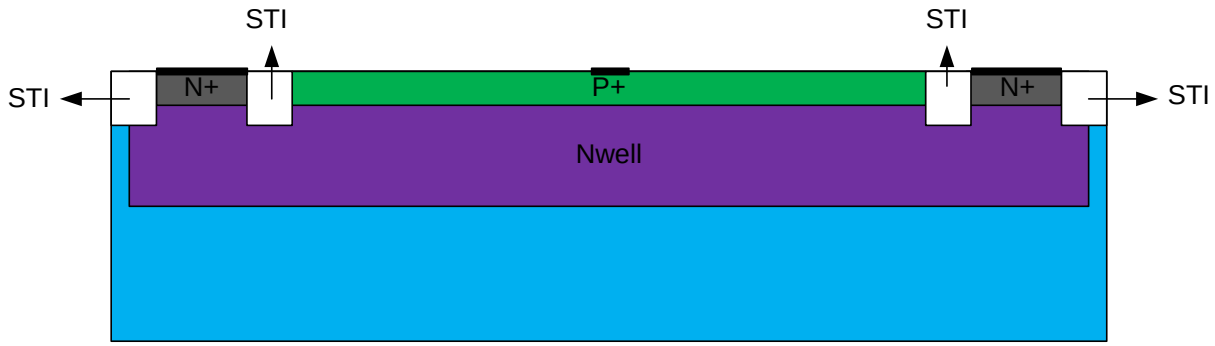


Figure 39 - Cross-section of the P+/NWELL structure with STI guard ring.

In this structure (Figure 39), the active area is a P+/Nwell junction, a STI guard ring was used to limit premature edge breakdown and ensure a high electric field in the multiplication region. The use of the STI guard ring allows for a more compact SPAD and as a result in a better fill factor but it will most likely results in a high DCR count due to defects introduced by the STI implementation process.

2.2.1.2 Structure B (P+ Nwell Deep Nwell with STI guard ring)

This structure (Figure 40) is similar to the structure A but an additional deep Nwell implant was introduced to isolate the active area from charge carriers diffusing from the substrate and causing faulty detections. The additional deep implant should result in a better PDP but it will lead to a higher jitter.

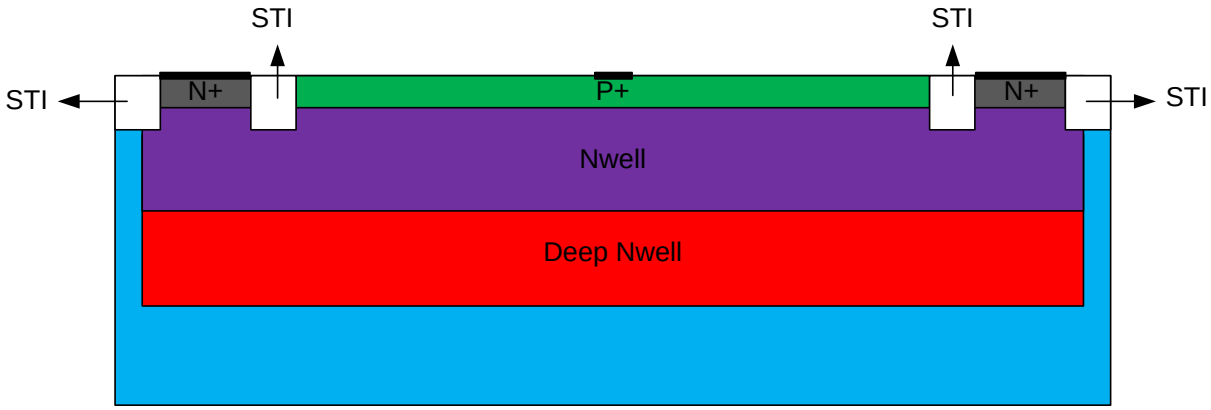


Figure 40 - Cross-section of the P+/NWELL Deep Nwell structure with Pwell guard ring

2.2.1.3 Structure C (P+ Nwell Deep Nwell with Pwell guard ring)

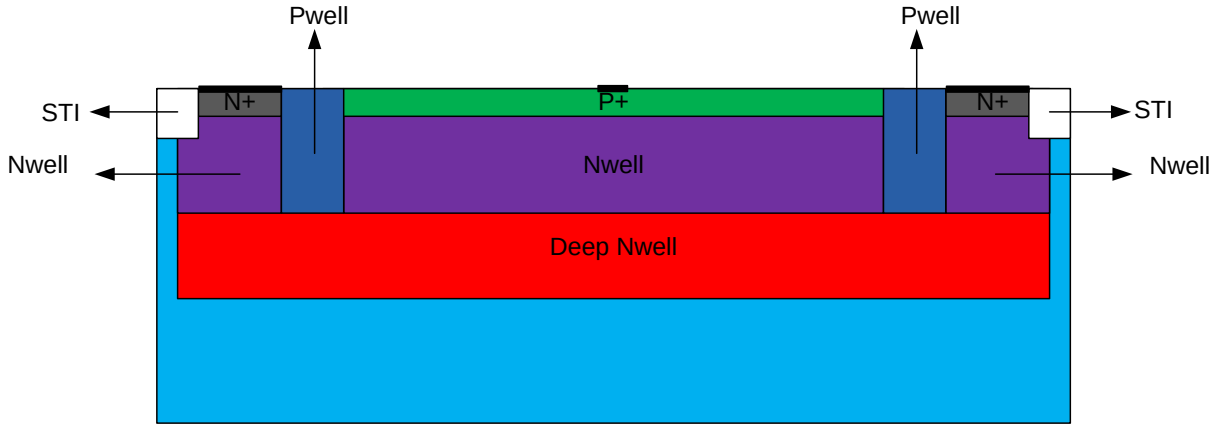


Figure 41 - Cross-section of the P+/NWELL Deep Nwell structure with Pwellguard ring

Like the structure B, the active area in this structure (Figure 41) is a P+/Nwell junction and the Deep Nwell was used to limit the charge carriers' diffusion from the substrate. However this structure employs a Pwell implant as a guard ring. The use of the Pwell implant help reduce the DCR due to defects introduced by the STI implementation process but it results in a lower electric field in the multiplication region and the presence of a low electric field outside the guard ring which increases the active area and as a result lead to a timing jitter with a more marked exponential tail. Another inconvenient of the Pwell guard ring is the limitation on the minimum active area diameter as it should be higher than 5 µm otherwise the depletion regions around the Pwell guard ring will connect together.

2.2.1.4 Structure D (P+ Pwell Deep Nwell with no guard ring)

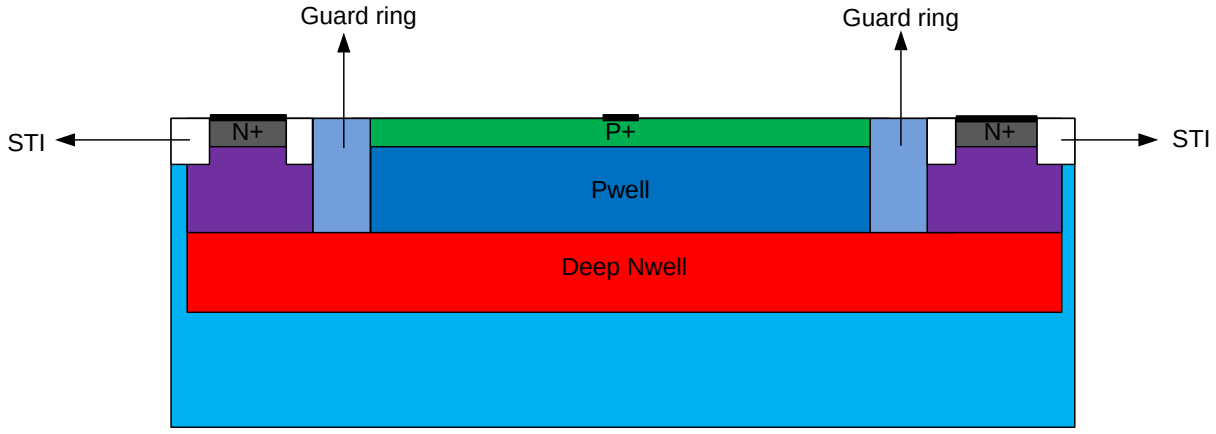


Figure 42 - Cross-section of the P+/PWELL Deep Nwell structure with no guard ring

In this structure the active area is P+/Pwell implant but the multiplication region is localized at the overlapping between the pwell and Deep Nwell implants thus allowing the confinement of a high electric field between the 2 implants (Figure 42) In this structure a virtual guard ring was formed by blocking the generation of any implant and an Nwell implant was used to connect the deep Nwell to N+ contact.

2.2.1.1 Structure E (P+ Pwell Deep Nwell with STI guard ring)

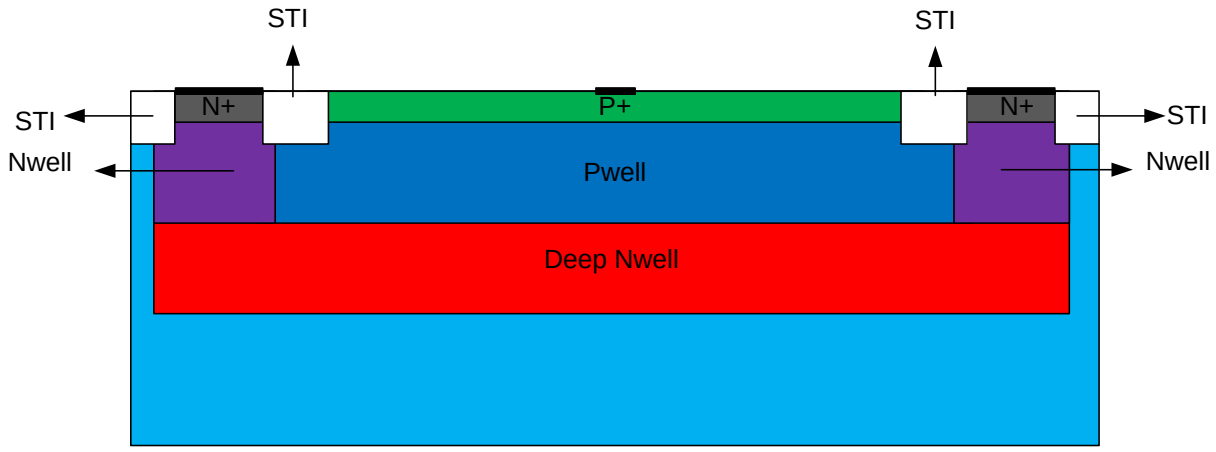


Figure 43 - Cross-section of the P+/PWELL Deep Nwell structure with STI guard ring

This structure is similar to the structure D but in this variation an STI implant was used to create a guard ring around the active area (Figure 43).

2.2.1.2 Structure F (P+ Nwell Deep Nwell with a double Pwell-STI guard ring)

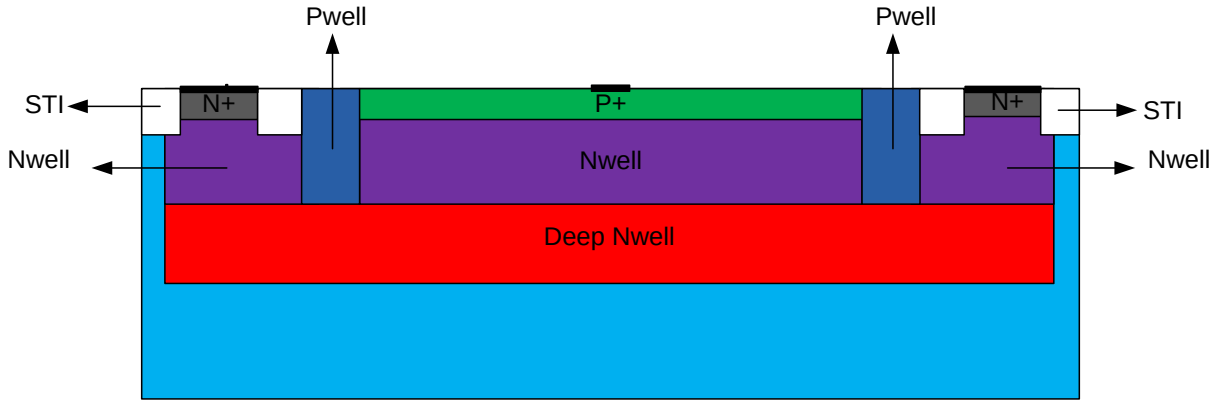


Figure 44 - Cross-section of the P+/NWELL Deep Nwell structure with a double Pwell-STI guard ring

The active area in this structure is a P+/Nwell junction in a retrograded Deep Nwell on a high resistivity P-substrate (Figure 44). To avoid premature edge breakdown a double Pwell/STI guard ring was implemented. The STI guard ring allows a high electric field in the multiplication region and the low doped Pwell guard ring was added to separate the multiplication region from the STI implant to reduce high dark noise due to faulty detections triggered by carrier injections from the STI interface into the multiplication area.

2.2.2 SPAD Versatile Characterization Circuit

In this chapter, we present the characterization circuit used to test and characterize the developed SPAD structures. The demonstrated circuit allows the quick prototyping and characterization of the SPAD structures and can be implemented in any specific CMOS process. The circuit consists of a large array that makes possible to test a large number of different structures in a single chip, thus limiting the number of iteration to find the best SPAD. Using this approach all the major parameters of the SPAD such as the DCR, the timing jitter, the photon detection probability and the after pulse probability can be measured through the versatile quenching circuit driving the SPAD.

2.2.2.1 Circuit Structure

The overall circuit is depicted in Figure 45 where the term “pixel” refers to the different tested architectures as opposed to the classical definition of pixel. This circuit is rather a versatile test bench for SPAD than an image sensor. It consists of an array of 8x8 pixels which allows the testing and the characterization of the 64 different SPADs, 4 quenching modes are available: SPAD off, passive, active reset, active quenching and reset. Furthermore quenching and reset duration can be adjusted by tuning the analog voltage applied to *T_quench* and *T_reset* (Figure 46) the digital input *Stop_Quench* is used to stop the quenching process in case it gets stuck, this could happen sometimes when the quenching mode is changed at the same time an avalanche is triggered causing the blockage of the quenching process. Each

pixel$_{ij}$ of a single line shares the same reverse bias voltage for the SPADs called *Alim<j>*, thus 8 alimentation lines are used for the complete structure. This is particularly useful when testing some SPAD structures that have various breakdown voltages and therefore require different biasing voltages. Furthermore, design rules violations can be deliberately done to obtain a specific but consequently unwarranted architecture that can result in faulty device and in the worst case, a short circuit. The later default will be limited to the line including the defective device. The 3 digital outputs *num_out*, *SPAD_rst* and *Quench* are shared by all the pixels however a designated pixel can be read through those outputs via a series of multiplexing process. This makes it possible to select one column among the total 8 columns of the structure. The 8 digital 3 bit outputs of the pixel of the selected column are then connected to another 8 to1 multiplexer to select the outputs of 1 pixel and transmit them through the 3 external outputs *num_out*, *SPAD_rst* and *Quench*. In order to configure and select pixels, a 198 shifting register (daisy_chain) is used, *Daisy_in* is its input and *Daisy_clk* is the input of the clock signal shared by all the registers of the daisy chain. The first 6 registers of the daisy_chain are used to configure the multiplexers in order to select a pixel to be read out. The remaining registers are used to configure the 64 pixels. Each pixel requires 3 bits to be configured: 2 bits are used to set the quenching mode, the remaining bit is used to control the analog outputs *Ana_out<j>* of the pixels in order to connect its output through the line shared with all the pixels of the same column *j.* Using an analog output for each column makes it possible to monitor the SPAD voltage while minimizing the added capacitance to only the path and the pad capacitance. For these specific outputs, a passive and low capacitance pad has been designed. The resulting 8 analog outputs *Ana_out<j>* are buffered by an external wideband and low input capacitance operational amplifier.

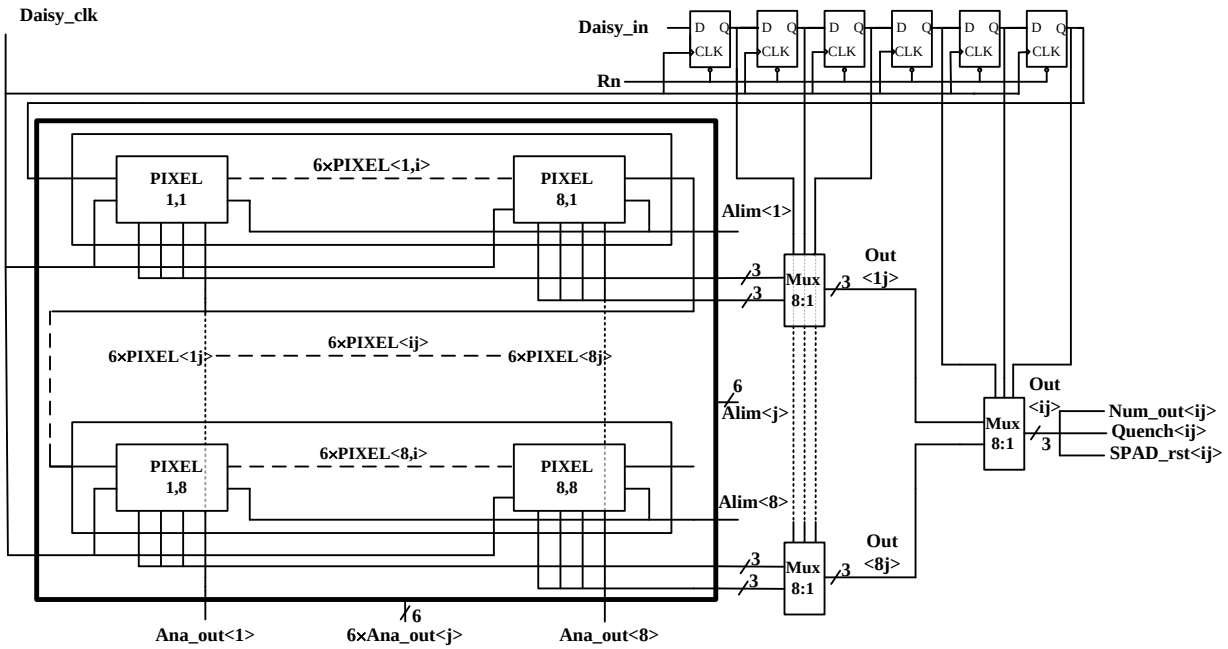


Figure 45 - Illustration of the proposed 8x8 pixels array along with the daisy_chain register and the multiplexing system.

2.2.2.2 Pixel Structure

The basic structure of the proposed pixel is shown in Figure 46. It combines a SPAD, a quench resistor, the quench control bloc and the logic control used to configure the pixel. The used quenching resistor is a voltage controlled PMOS transistor used as a passive quenching resistor which allows more flexibility in setting the value of the quench resistor. An active quenching is provided by the NMOS transistor M_Q while the PMOS transistor M_R provides the active reset. Each pixel contains a 3 bit shifting register which is a part of the general 198 shifting register (the daisy_chain), those 3 bits are used to control the pixel's mode and analog output, the last bit Q_2 controls the transmission gate (T- gate) which connects or disconnects the analog output *Ana_out<j>* of the pixel to the shared external analog output. The 2 remaining bits Q_1 and Q_0 are used to configure the logic control bloc and set the quenching mode via the 3 control block inputs: *Sel_Quench*, *Disable* and *Sel_reset* among 4 possible quenching modes resumed in Table 2.

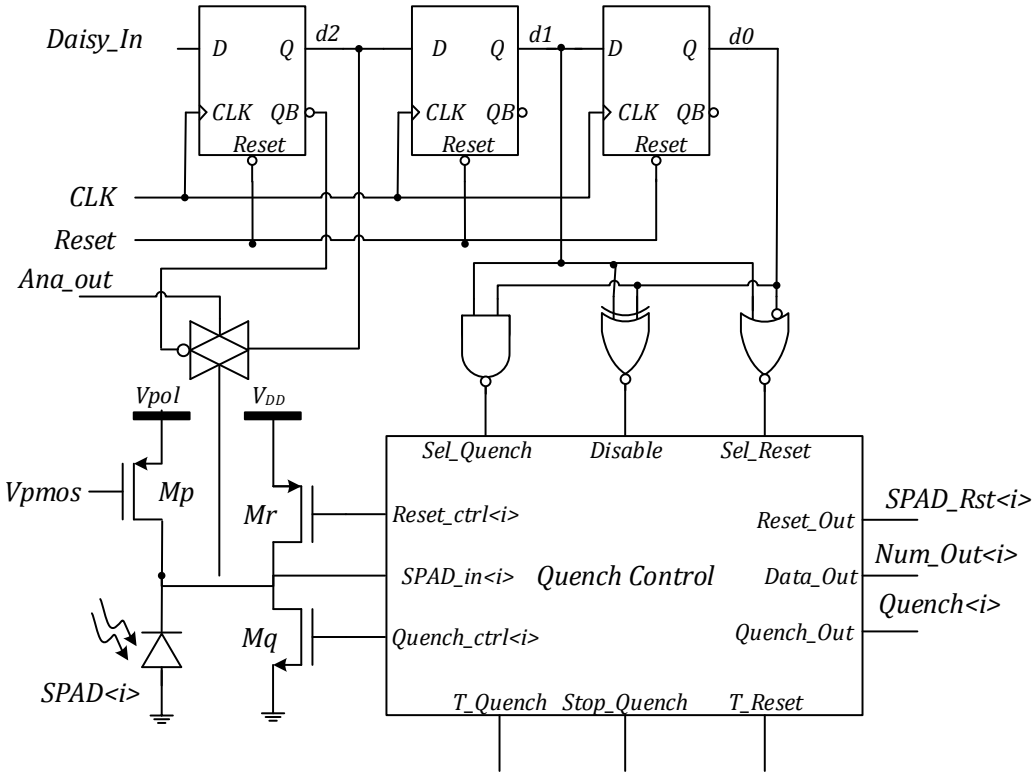


Figure 46 - Description of one pixel of the SPAD array: the SPAD is controlled by a versatile quenching circuit configurable via the daisy chain.

Table 2 - Configurable quenching modes of the proposed pixel.

Mode	Q_1	Q_0	Description
Off	0	0	The SPAD bias is kept below its breakdown voltage to prevent unwanted accidental avalanches and crosstalk between the pixels
Passive	0	1	The avalanche is quenched passively via the quench resistor
Active Reset	1	0	A passive quenching stops the avalanche followed by an active reset to restore the SPAD original biasing voltage
Auto	1	1	A mixed active quenching is used to stop the avalanche and reset the SPAD

2.2.2.3 Versatile Quenching Control System

The Block diagram of the Quench control system is illustrated in Figure 47. When an avalanche is triggered *SPAD_in<ij>* changes from high to low. As a result *data_out<ij>* changes to high indicating that a photo or thermo electron was detected by the SPAD. A buffer is used as a level shifter between the analog circuitry operating in 3.3 V and the digital block operating in 1.8V. Two other buffers are used to perform the reverse translation and allow the *Rest_ctrl* and *Quench _ctrl* to command the two transistors M_Q and M_R operating in 3.3V. The timer is the basic part of the block; in the absence of an avalanche its output is always set to high, however when an avalanche is triggered causing the input to become low, the timer's output goes from high to low after a certain time delay set via the analog input *T_Quench*. It stays low for a time interval chosen via the second analog input *T_Reset* which fixes the low signal width. To set the quench mode, the quench control block is configured through the 3 inputs *Sel_quench*, *Sel_reset* and *Disable*. These signals are driven by the Q_0 and Q_1 bit of the daisy chain via the glue logic as shown in the Figure 47. The operation of each mode is summarized as follows:

Off Mode (Sel_Quench=1, Disable = 1, Sel_Reset=0): *Quench_ctrl<ij>* is always high as M_Q is always "on" keeping the SPAD below its breakdown voltage.

Passive Mode (*Sel_Quench* = 1, *Disable* = 0, *Sel_Reset* = 1): *Quench_ctrl* is low and *Reset_ctrl* is high. The outputs are independent of the timer's output and both M_Q and M_R are always "off". The pixel is equivalent to a SPAD with a quench resistor and when an avalanche takes place, it is stopped by passive quenching only.

Active Reset Mode: (*Sel_Quench*=1, *Disable* = 0, *Sel_Reset*=0): When an avalanche is triggered, a passive quenching is set in action and the timer's input changes to low. *Disable* is low while *Sel_Quench* is high thus *Quench_ctrl* is set to low and M_Q is "off". The Active

Quenching is deactivated and avalanches are stopped passively via the quench resistor. After a time delay set by the analog input *T_quench*, the timer's output changes to low, it remains low for a time set via the input *T_reset*. As a result *Reset_ctrl* which was initially high changes to low and starts the reset phase by making M_R conducting until the timer's output returns to high.

Auto Mode (*Sel_Quench* = 0, *Disable* = 1, *Sel_Reset* = 0) : Both *Sel_Quench* and *Sel_Reset* are set to low and the timer's output is thereby directly connected to the block's outputs. When an avalanche is triggered, the block input's level changes to low, *Stop_Quench* =1 and the timer's output stays high for a time set but *T_Quench*. *Quench_ctrl* and *Reset_ctrl* are both high, consequently M_Q is conducting, $M_R$is blocked and the quenching phase is activated. After the timer's delay, its output goes to low, *Quench_ctrlReset_ctrl* therefore changes to low blocking M_Q and making M_R conducting, the quenching phase is terminated and the reset phase starts until the timer's output goes back to its original level.

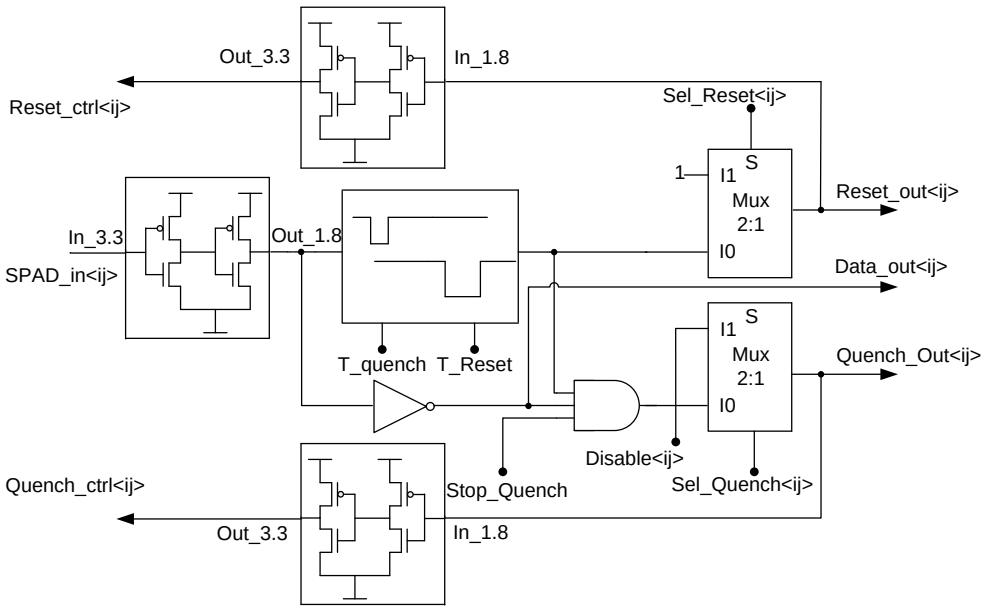


Figure 47 - Detail of the versatile quenching control.

2.3 Characterization Results

To fully characterize a SPAD performance several parameters needs to be measured: The noise is measured as the rate of faulty detection in the dark or dark count rates (DCR) which also includes the afterpulsing probability, the sensitivity measured in terms of the photodetection probability (PDP), and the timing jitter. These parameters depends of several factors: excess bias, temperature, active area, dead time length , as well as wavelength (timing jitter and PDP), the characterization tests of the SPAD should take into account the effect of the variation of these factors on the measured parameters.

2.3.1 Breakdown voltage

The first step of the testing process was the determination of the breakdown voltage (V_{br}) of each structure in order to ensure that the structure is functional and in which case the limit of the Geiger mode. To determine the breakdown voltage we plotted the I-V characteristics while increasing the SPAD reverse-bias. The results are presented in Figure 48, they show that the 6 Structures show an avalanche at different breakdown voltages as summarized in Table 3.

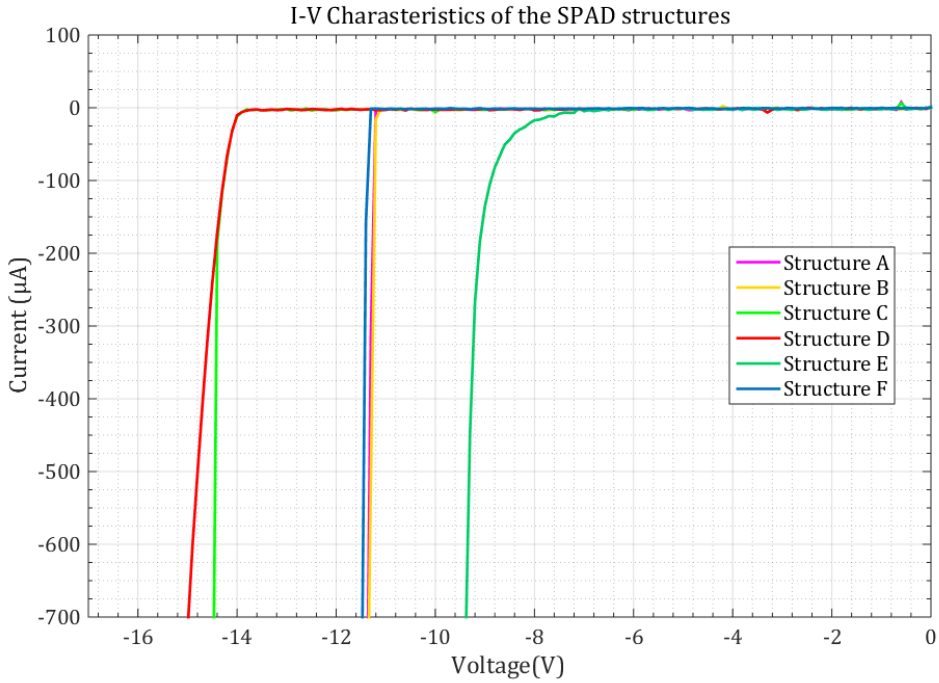


Table 3 - Breakdown voltages of the demonstrated SPAD structures

Structure	Breakdown Voltage
A	-11.2 V
B	-11.2 V
C	-14 V
D	-14 V
E	-7.5
F	-11.3 V

2.3.2 Electroluminescence light emission test

Based on the I-V measurements the 6 SPADs showed an avalanche when polarized above their breakdown voltage but to guarantee that the SPADs are properly functional we needed to determine the quality of these avalanches especially in term of uniformity over the all the active area which can be done by means of the electroluminescence light emission test. Electroluminescence is a the result of material emitting light in response to the passage of an electric current, this phenomenon is the result of spontaneous emission when an excited electron in the state (E_2) spontaneously decay into a lower energy state (E_1) releasing the difference of energy between the two states in the form of a photon Figure 49.

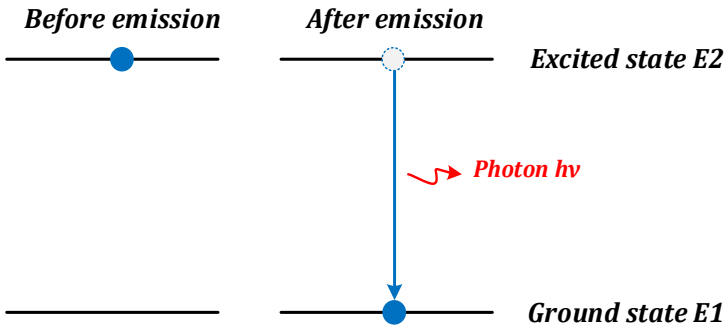


Figure 49 - Spontaneous emission phenomenon

To perform the electroluminescence light emission test, an avalanche is triggered in each SPAD resulting in photons being emitted by the hot carriers crossing the p-n junction high field region. Assuming a uniform electric field and as a result a uniform avalanche triggering probability, the electroluminescence light emission test should show a uniformly bright active area. The electroluminescence light emission test was performed on the 6 SPAD structures and the results are shown in Figure 50. As we can see, structures A, B and C (Figure 50- a, b, c) the avalanche showed a strong premature edge breakdown phenomenon. On the other hand, structure D (Figure 50 - d) showed a uniform avalanche but all the SPADs seemed to be shortcut together as was the case for the structure E (Figure 50- e). Finally the structure F (Figure 50 - f) showed a uniform avalanche process all over the surface of the active region and appeared to be functioning correctly. Based on these results, we were able to determine that only the SPAD structure E can be used as a SPAD and the characterization tests were carried on for this structure only.

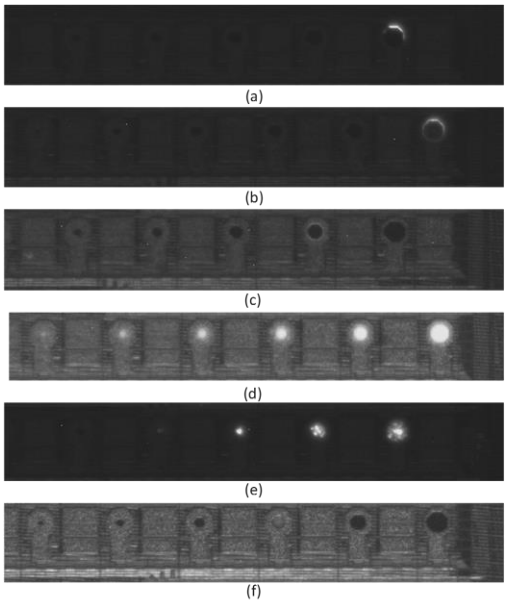


Figure 50 - Electroluminescence light emission test results for SPAD structures A (a), B (b), C (c), D (d), E (e) and F (f)

2.3.3 Dark count rate

Dark Count Rate (DCR) represents the measure in the dark of spurious avalanche events that are not resulting from a photon interaction. The DCR is triggered by several generation mechanisms: thermal generation, trap assisted generation and band to band tunneling; it is strongly dependent on the SPAD design as well as the used CMOS technology process. Primary DCR is random and statistically uncorrelated but SPADs display also secondary avalanches statistically correlated to primary avalanches known as afterpulsing. Afterpulsing is due to trapping and de trapping of charges flowing through the multiplication region during the primary avalanches; it is dependent of the current amplitude, the associated electronics as well as temperature and can be reduced by increasing the SPAD hold-time to allow the releasing of the trapped charges before activating the SPAD again. Figure 51 shows the afterpulsing probability measurement results using the inter-arrival time histogram method [95] for the 20 μm SPAD with an excess bias of 300mV. Figure 52 shows the afterpulsing probability of the 7 SPAD devices as a function of their active area diameter; the measured probability was around ~0.21% at 300mV for the 7 devices, indicating that the afterpulsing probability in not related to the surface of the device. Furthermore the inter-arrival time histograms obtained from the measurements by calculating time durations between 2 consecutive avalanches events showed that the undesirable afterpulsing can be eliminated by increasing the dead time to 1 μs. The total DCR was measured for the 7 devices at different excess bias values and with a temperature ranging from -25°C to 60°C. The DCR variation against temperature for an excess bias of 100 mV and 400 mV is represented in Figure 53 which shows that the measured DCR increases with the SPAD active area, at 15°C it is lower than 5 KHz for SPADs with an active

region diameter below 30 µm and lower than 10 KHz for the 30 µm and 40 µm SPADs. Raising the excess bias from 100 mV to 400 mV increases the DCR value with a mean factor of 2.2 over the whole range of considered temperatures for all the SPAD diameter variations, which proves that the impact of the excess bias on the DCR does not change with the value of the active area. To better illustrate the effect of SPAD diameter on the measured DCR, the ratio DCR/Surface variation against temperature for the 7 devices at an excess bias of 400 mV is plotted in Figure 54. The SPADs showed the same DCR/Area value with the exception of the 30 µm SPAD that generated more noise/µm^2 most likely due to higher defects concentration compared to the rest of the SPADs. The results also show a linear variation below 20°C with the DCR doubling every 10°C which suggests that it is mostly dominated by the tunneling mechanism. However, when the temperature is above 20°C, the DCR increases exponentially and the thermal generation becomes the dominating mechanism.

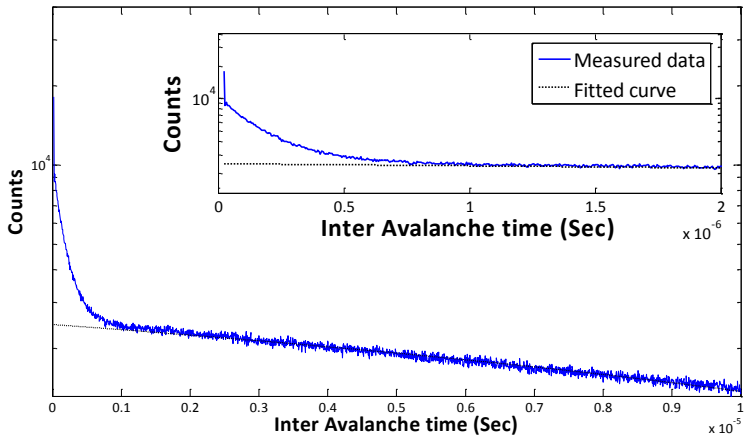


Figure 51 - Afterpulsing measurement results for a dead time of 30 ns and an excess bias of 300 mV

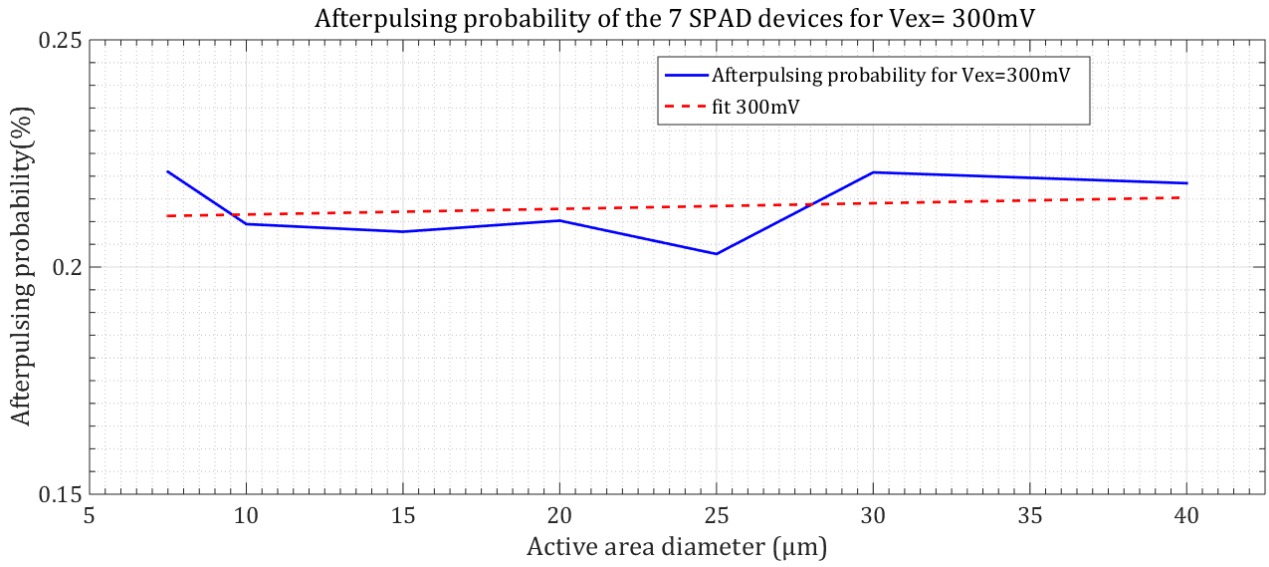
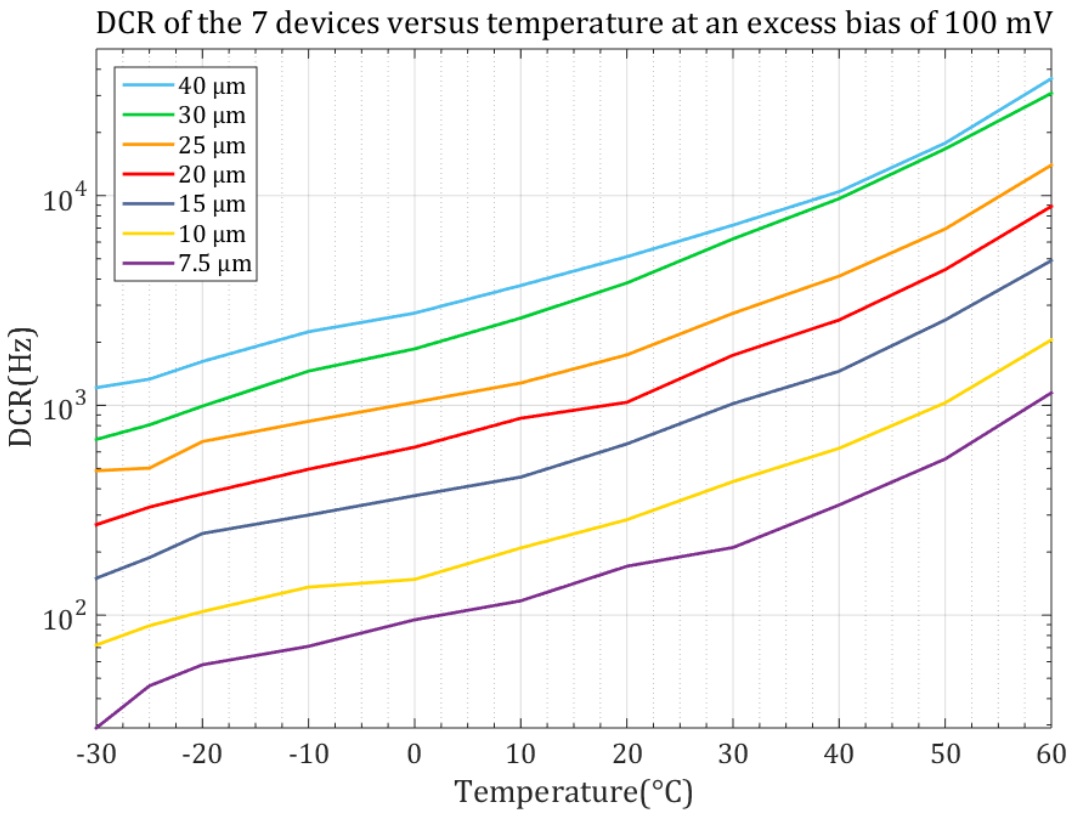


Figure 52 - Afterpulsing probability measurement results for the 7 devices with an excess bias of 300 mV



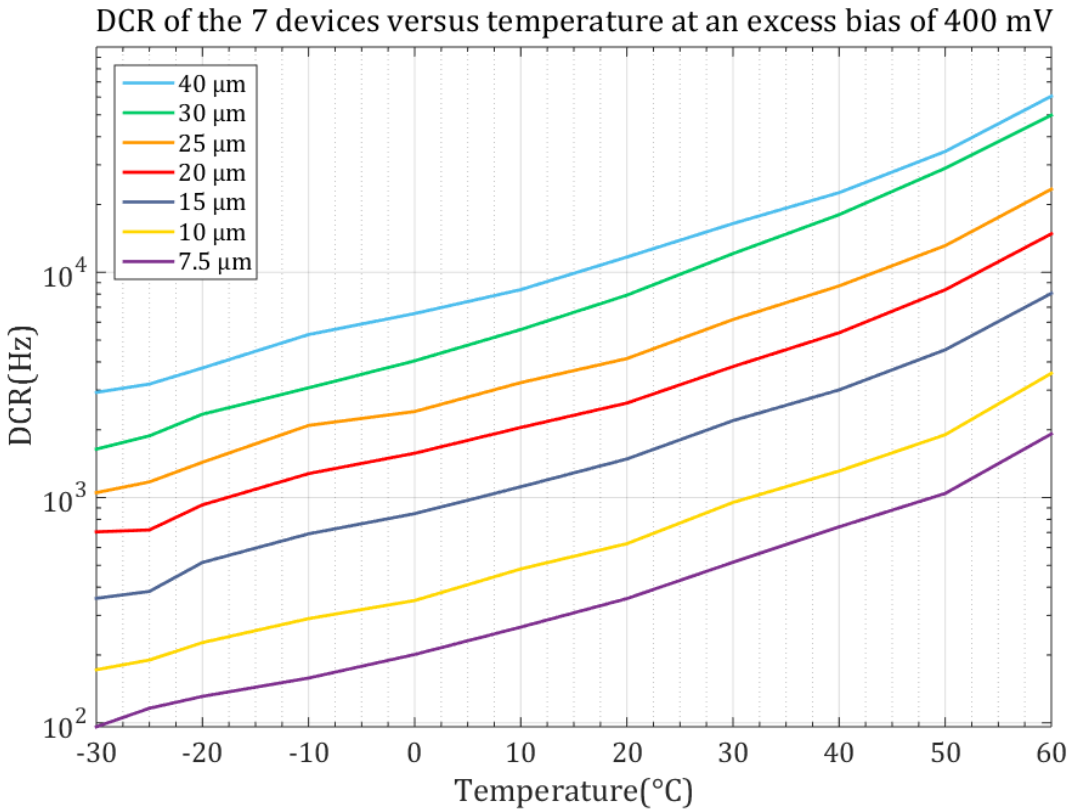


Figure 53 - DCR of the 7 devices versus temperature at an excess bias of 100 mV and 400 mV

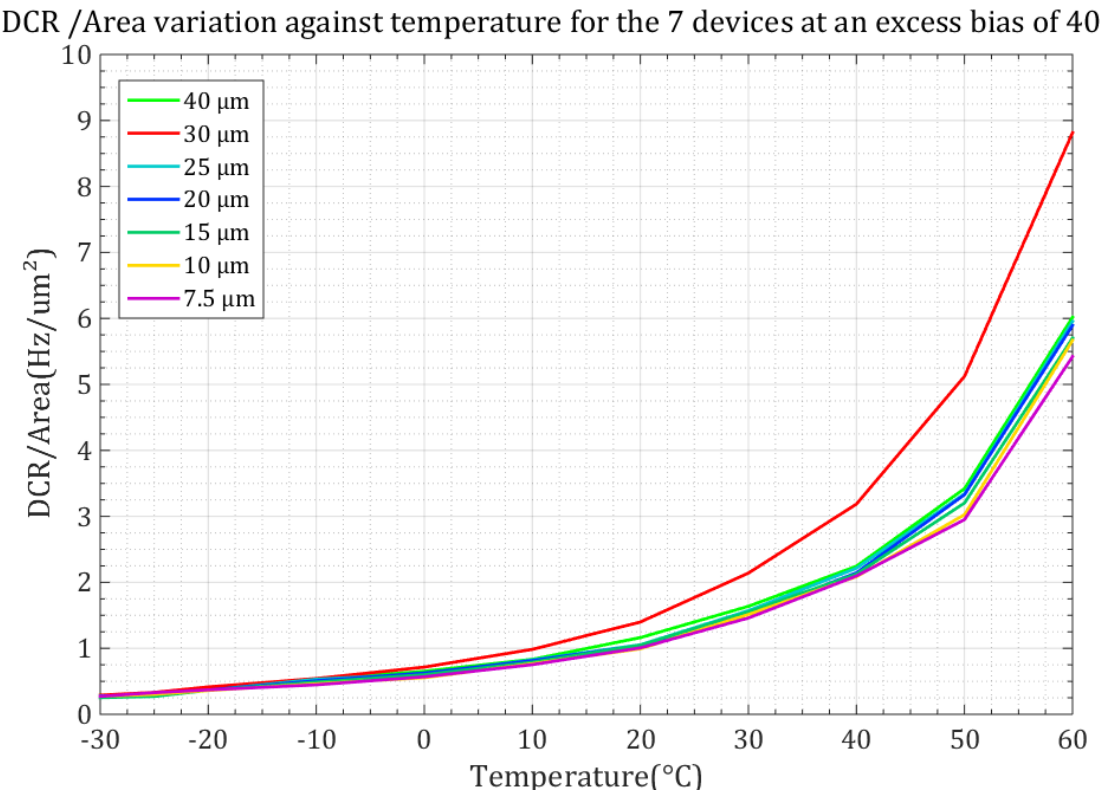


Figure 54 - DCR /Area variation against temperature for the 7 devices at an excess bias of 400 mV

2.3.4 Photodetection probability

The photodetection probability of the 7 working SPADs were measured at room temperature, over a wavelength range between 350 nm and 1000 nm, for various excess biases between 100 mV and 400 mV and a dead time of 30 ns. The results are shown in Figure 55-Figure 61. Figure 62 shows the PDP of the 7 devices for an excess bias of 300mV. The measurements showed that the PDP peaks around 430 nm for the 7 devices. The maximum value for a given active area increases with the excess bias and the maximum value for a given excess bias value increases with the active area although the amount of PDP enhancement tends to decrease and becomes very small for diameters higher than 20 µm (Figure 62). Finally, the measured PDP peak varies from 11% for the 7.5 µm SPAD up to 19% for the 40 µm SPAD.

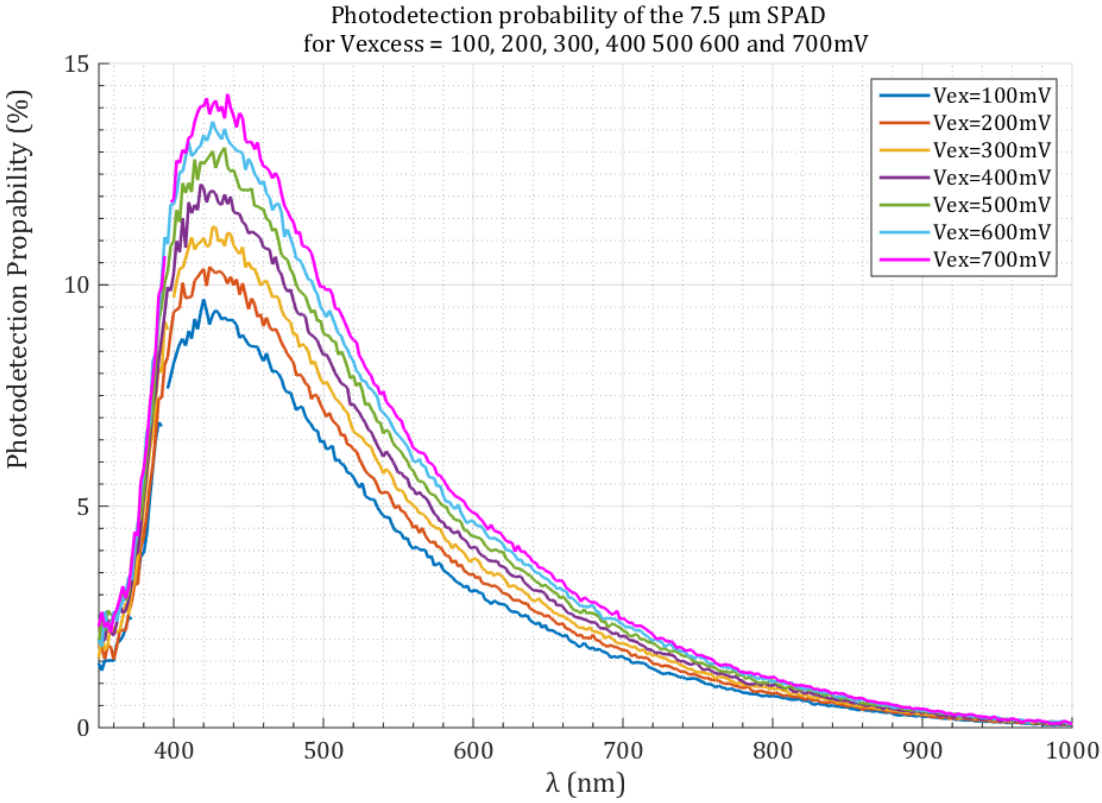


Figure 55 - Photodetection probability of the 7.5 µm SPAD for various excess bias

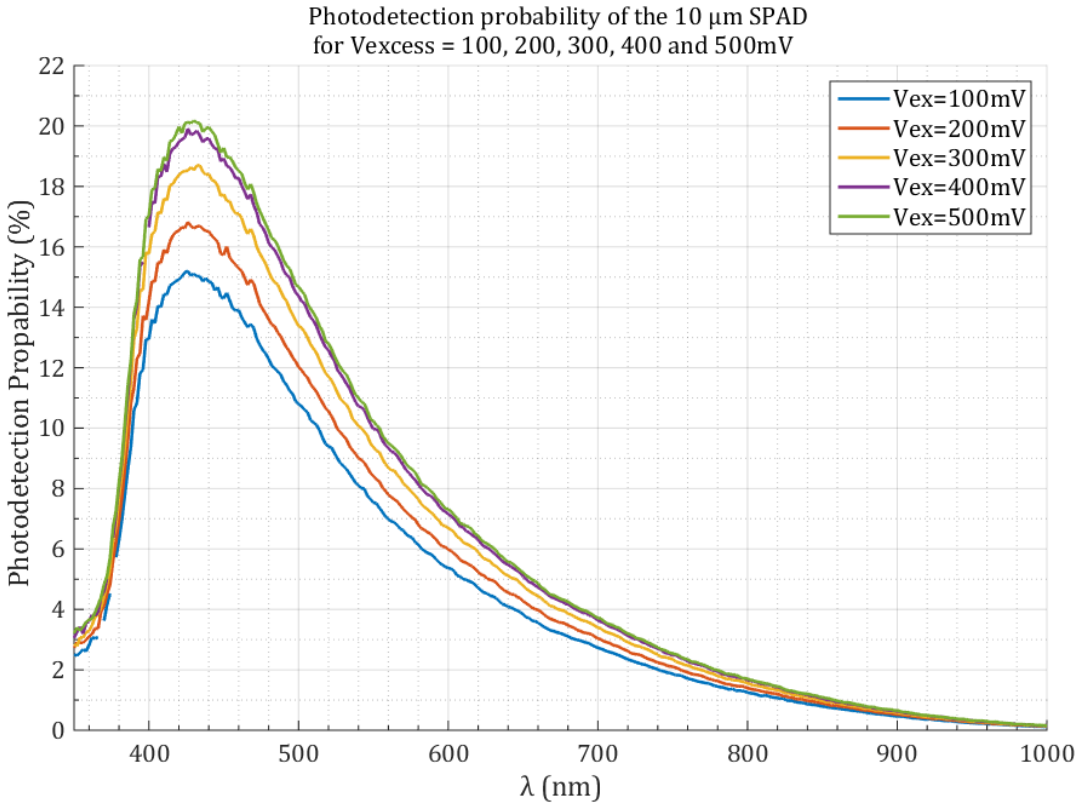


Figure 56 - Photodetection probability of the 10 µm SPAD for various excess bias

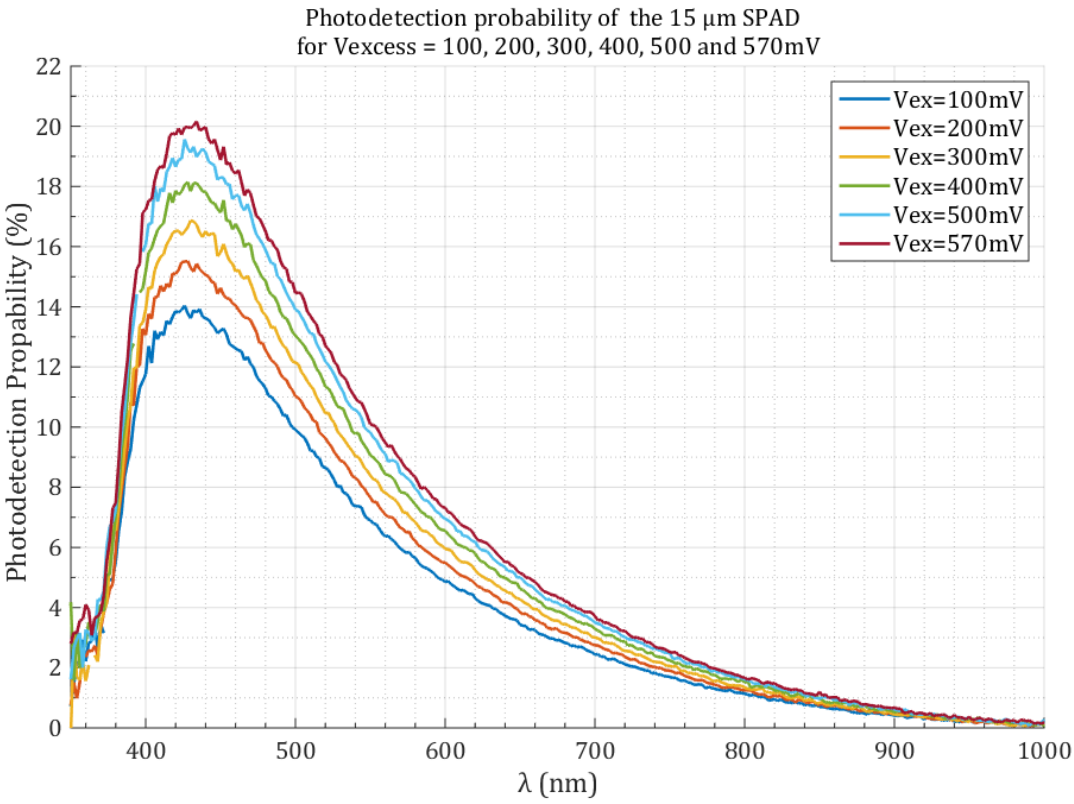


Figure 57 - Photodetection probability of the15 µm SPAD for various excess bias

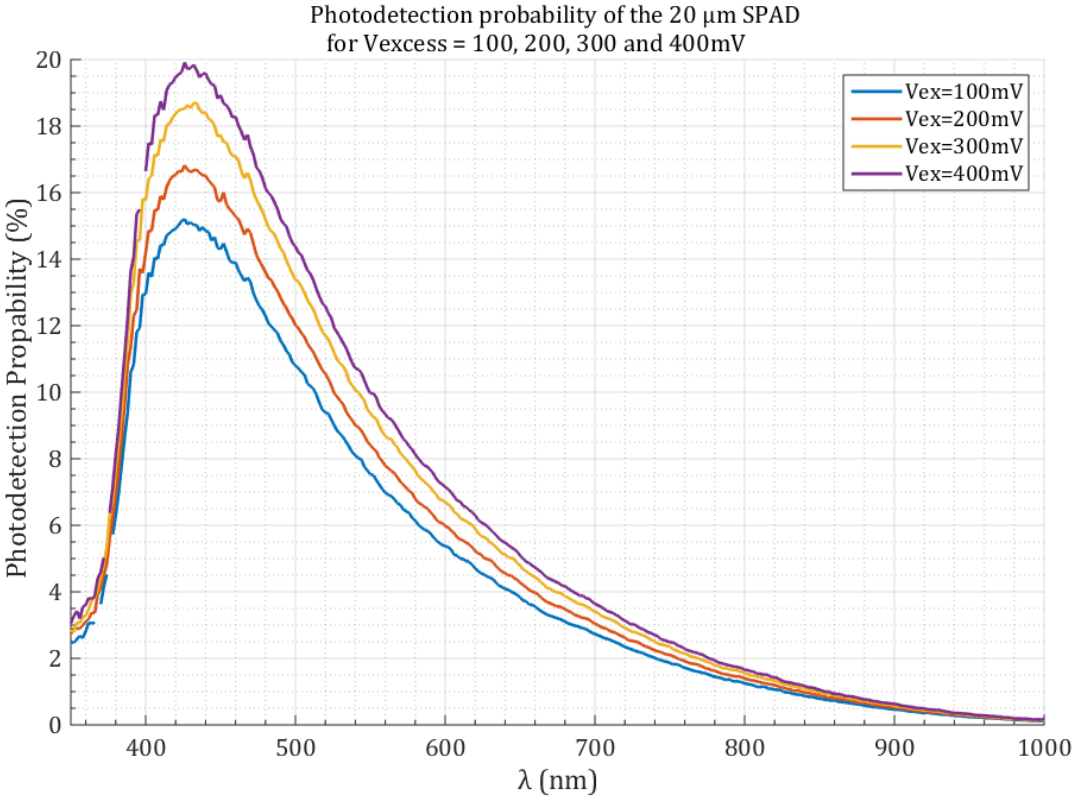


Figure 58 - Photodetection probability of the 20 µm SPAD for various excess bias

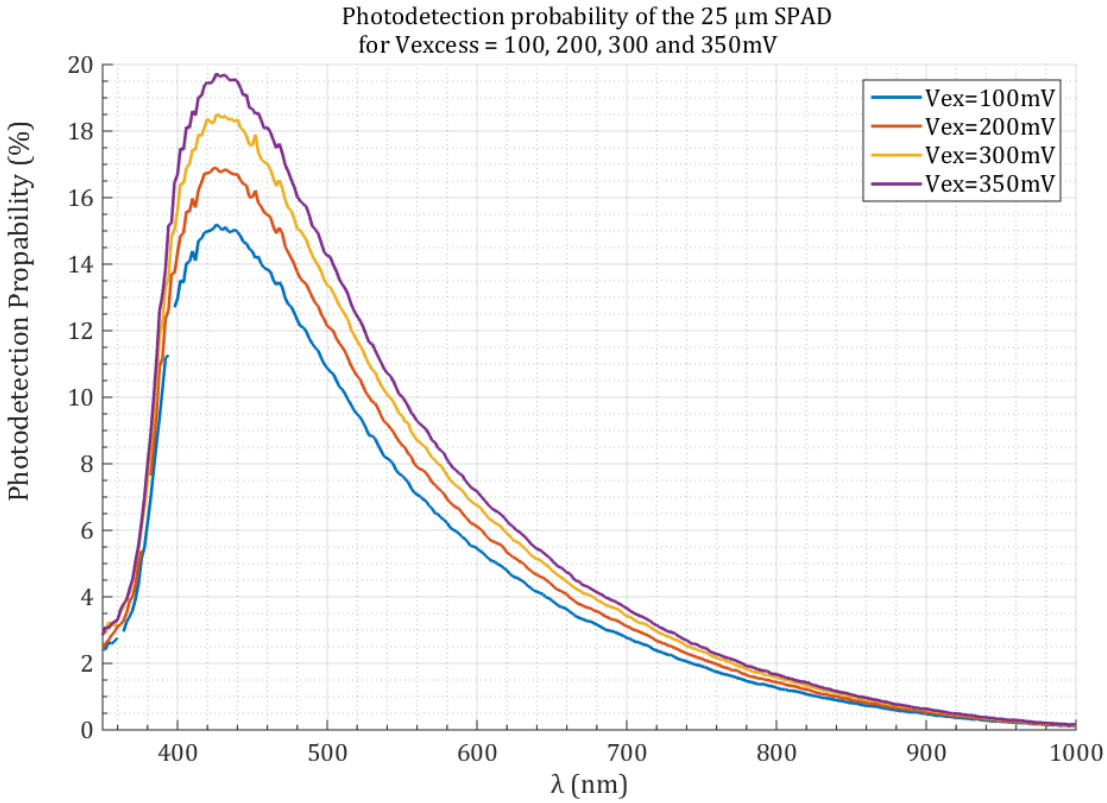


Figure 59 - Photodetection probability of the 25 µm SPAD for various excess bias

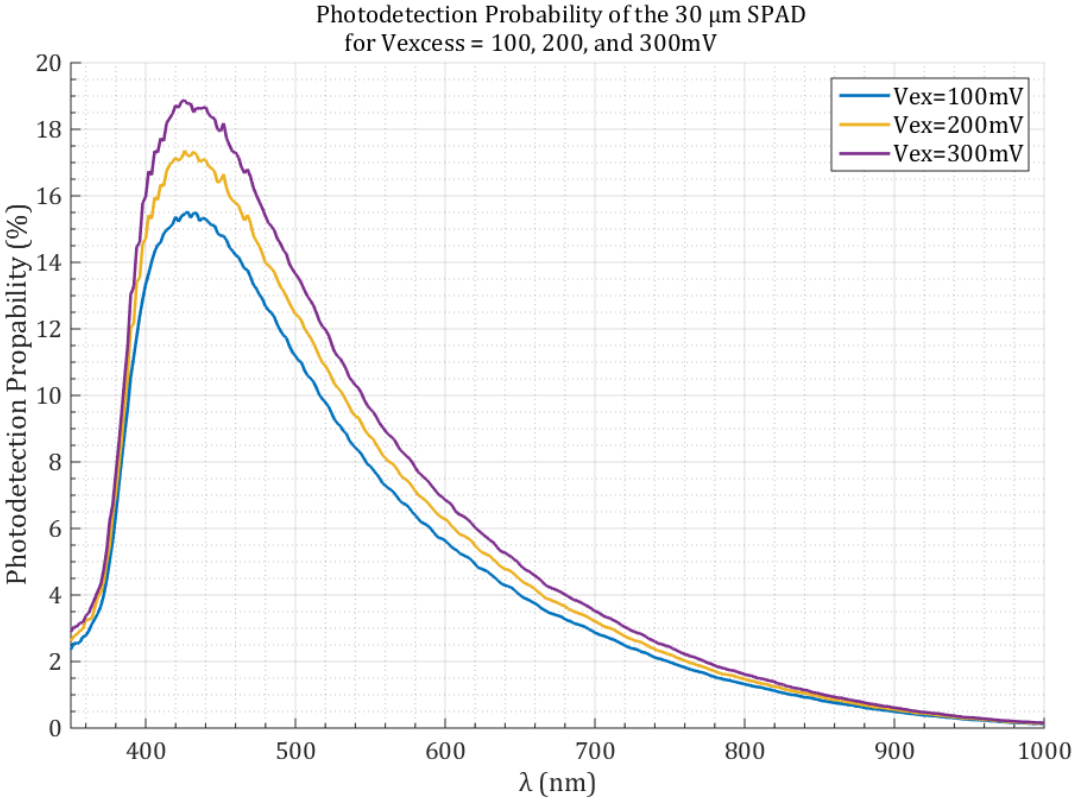


Figure 60 - Photodetection probability of the 30 µm SPAD for various excess bias

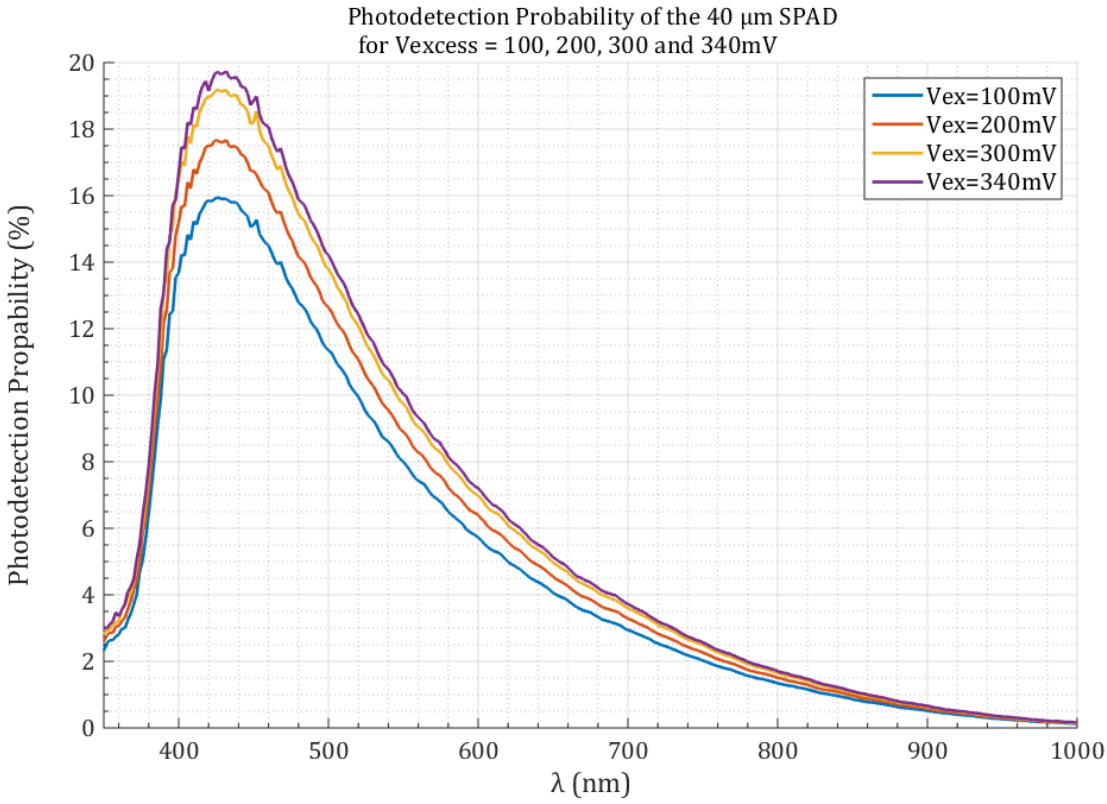


Figure 61 - Photodetection probability of the 40 µm SPAD for various excess bias

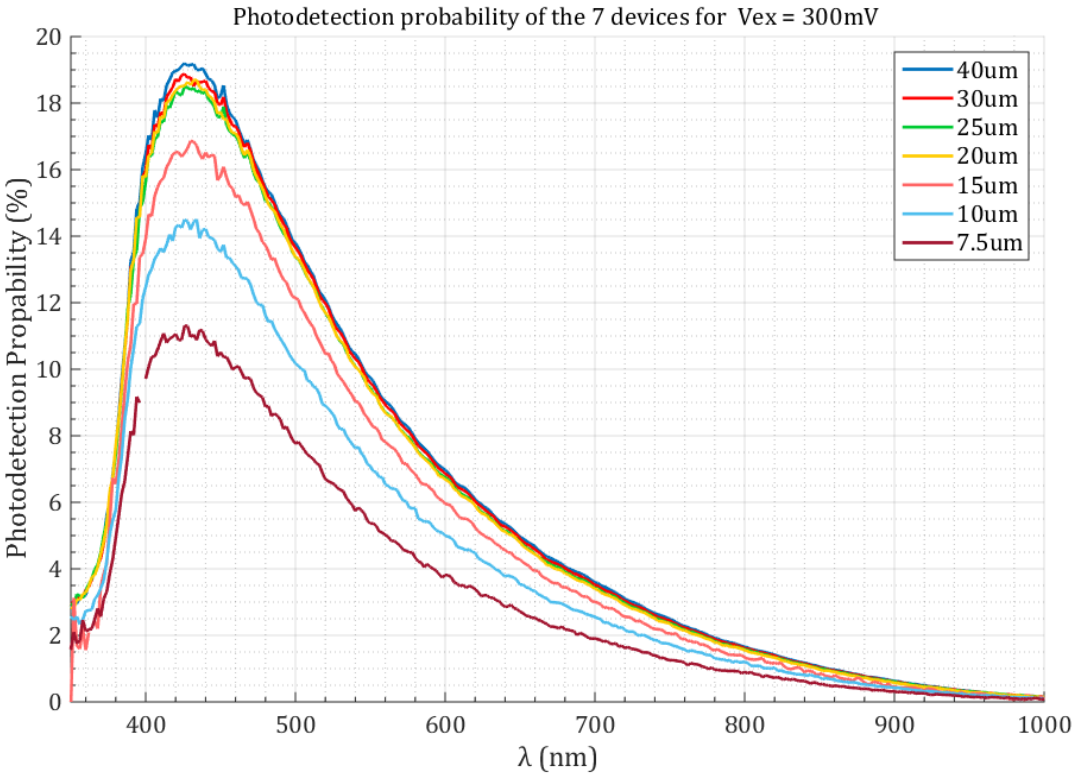


Figure 62 - Photodetection probability of the 7 SPADs for an excess bias of 300 mV

2.3.5 Timing resolution (Jitter)

The timing jitter values of the 7 devices were characterized using a TCSPC setup with two 20 MHz picoseconds laser diode [96] The timing response of the 7 SPADs are shown in Figure 63 for λ=450 nm and λ=808 nm at an excess bias of 300 mV. The pulse width of the light emission was measured with the streak camera described in [97] to be 67 ps FWHM and 99 ps FWHM for the 450 and the 808 nm laser diode respectively.
The distributions showed a sharp Gaussian peak and an exponential tail, the sharp Gaussian peak is due to photons absorbed in the multiplication area while the tail is due to minority charge carriers diffusion and photons absorbed below the multiplication region [75]. In general, the timing response FWHM for a giving excess bias increases with the active area diameter due to lateral and vertical dependence of timing uncertainty on photon absorption location thus better timing resolutions are obtained with the smaller SPADs for the 2 laser source wavelength values. Furthermore the 450 nm laser yields better timing resolutions due to the fact that photons with higher wavelength are absorbed near the SPAD surface, thus near the multiplication region. This is better illustrated in Figure 64 which shows the raw measured jitter (Jitter Measured) values (solid line) of the 7 SPADs for the 2 wavelength values and for an excess bias of 300 mV. The effect of the optical pulse width have been decorrelated to the measured jitter to obtain a value closer (Jitter Real) (dotted line) to the real device performances according to the equation

$$Jitter_{real} = \sqrt{(Jitter_{measured})^2 - (Pulse\ width)^2} \quad (12)$$

At 300 mV the best measured timing resolution is equal to 34 ps FWHM achieved using the SPAD with an active area of 7.5 µm for λ=450 nm. The maximum jitter for λ=450 nm is obtained using the 40 µm SPAD and it is equal to 300 ps. for λ=808 nm, the minimum measured jitter is equal to 67 ps and the maximum jitter is equal to 350 ps, they were measured using the 40 µm SPAD. Figure 65 shows the timing jitter variation of the 7.5 µm to 25 µm SPADs as a function of the excess bias for λ = 405 nm at 300 mV. The maximum jitter decreases as the excess bias increases due to the increase of the electric field strength in the multiplication region which reduces the time needed for a photon to generate a reverse current high enough to be detected by the SPAD's associated electronics. For λ=450 nm (Figure 65), the measured timing jitter of 7.5 µm device is around 100 ps at 200 mV and decreases to 65 ps at 500 mV. Similarly the measured jitter for the 25 µm is 166 ps at a 200 mV excess bias and decreases to ~130 ps at 500 mV.

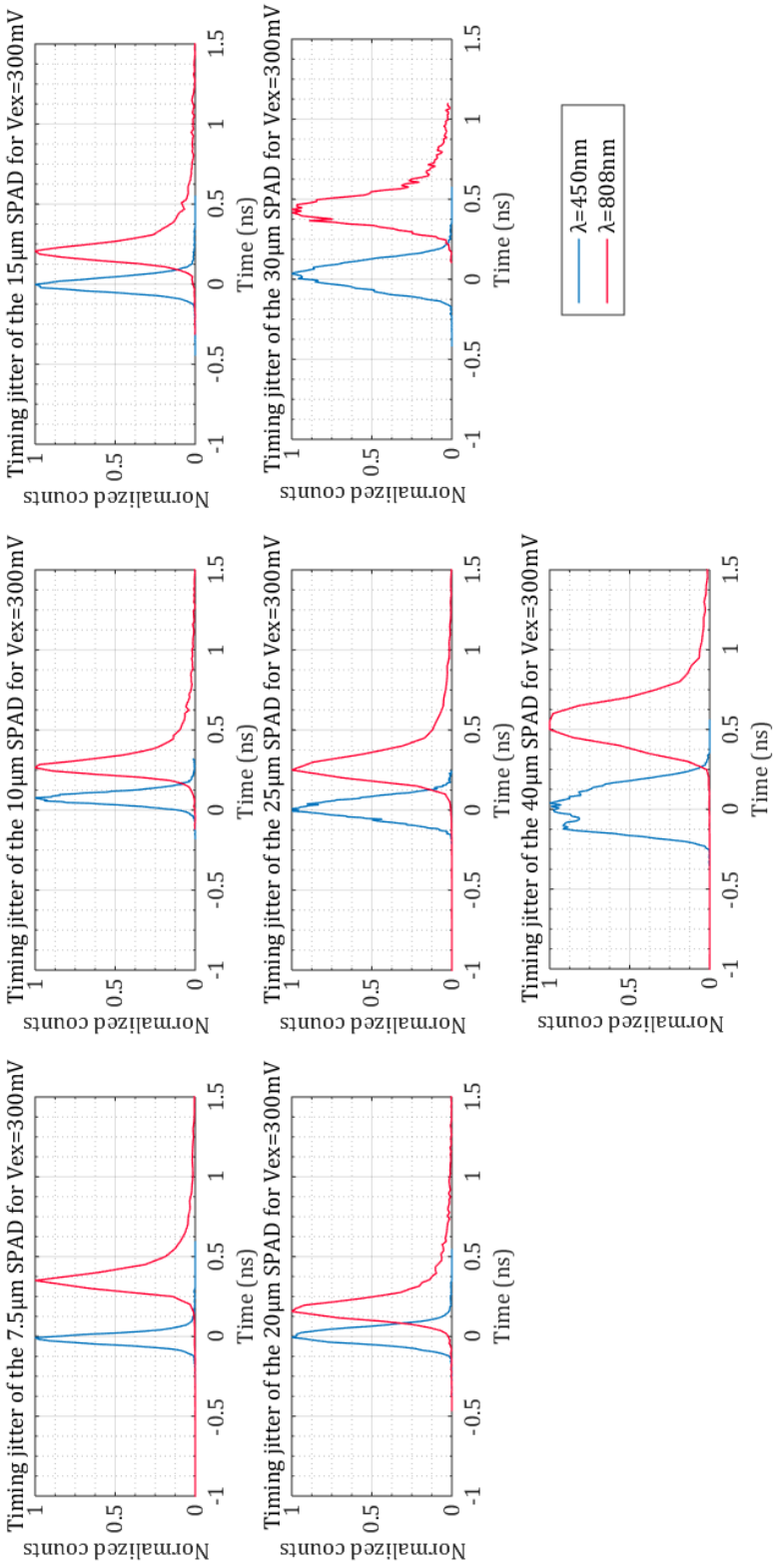


Figure 63 - Normalized timing response measurement results using a 450 nm and an 808 nm laser with an excess bias of 300mV for the 7 SPAD diameters

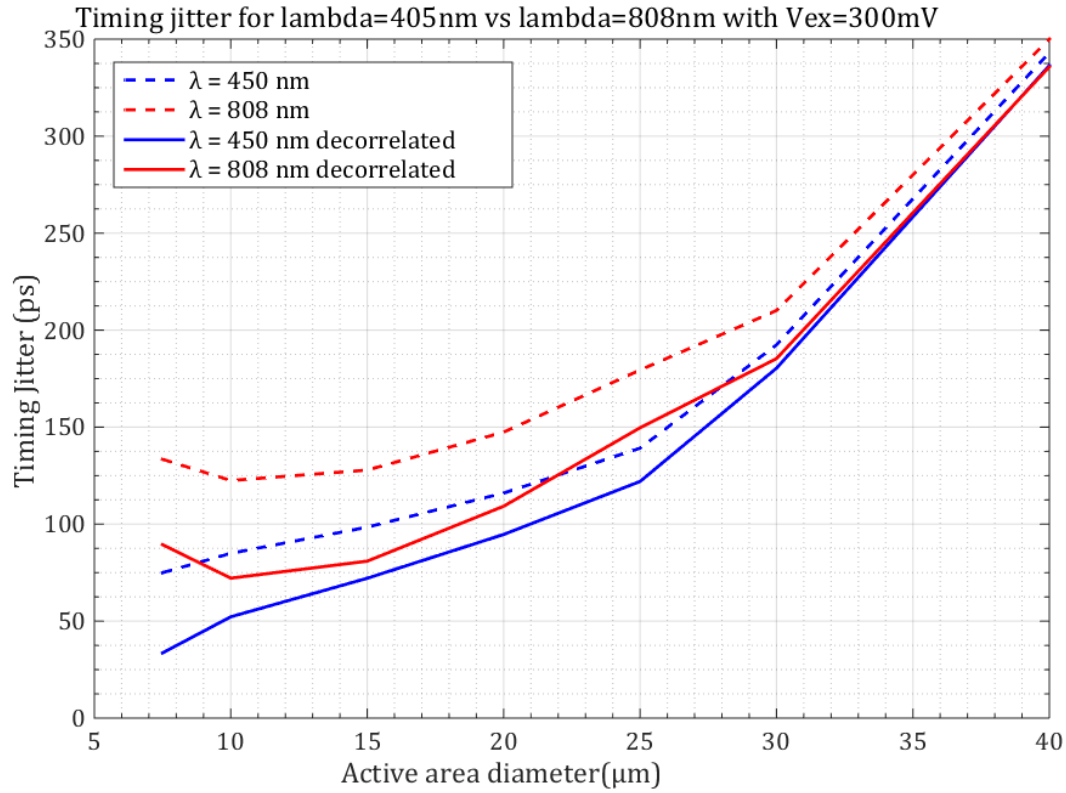


Figure 64 - Timing jitter results for λ = 450 nm and λ = 808 nm at an excess bias of 300 mV

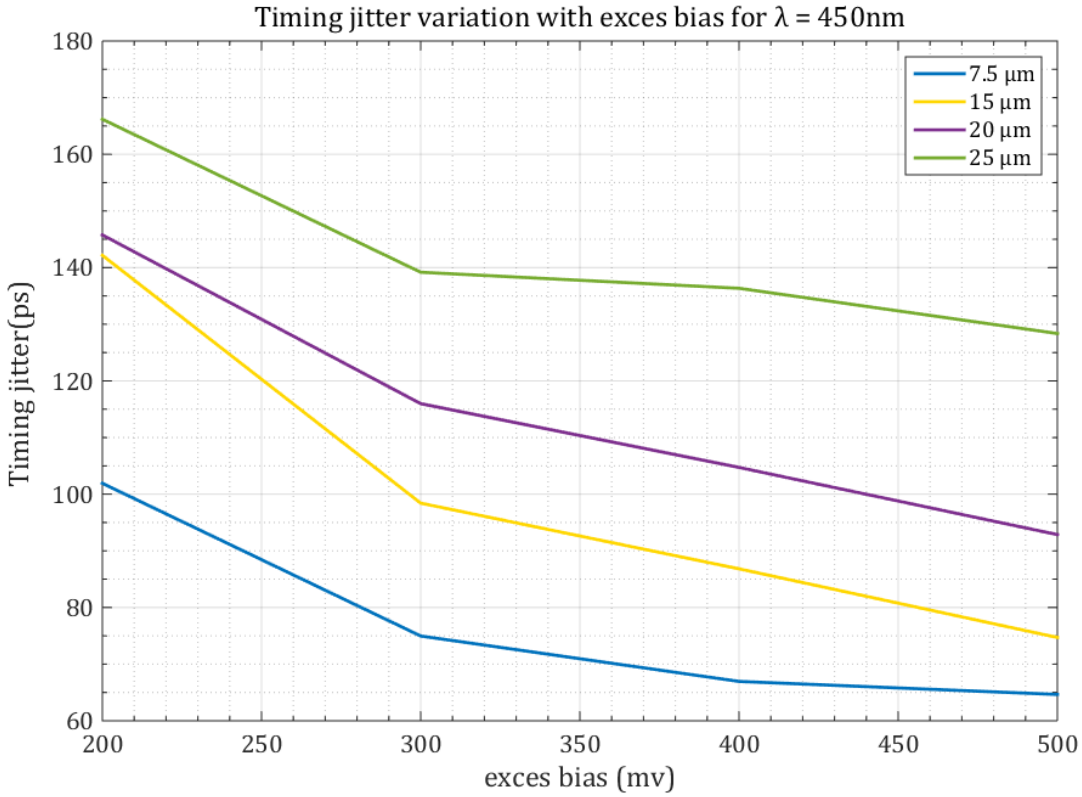


Figure 65 - Timing jitter variation of the 7.5µm to 25 µm SPADs as a function of the excess bias for λ = 405 nm

2.3.6 Conclusion

Despite a limited excess bias value due to the detection electronic, the SPAD structure compares relatively well with several reported SPADs fabricated in standard CMOS technology process (c.f. Table 4). The measurements showed a DCR below 5 KHz at 15°C with a highly suspected strong band to band tunneling and field-enhanced generation effects due to the relatively low breakdown voltage (11.4 V). The afterpulsing probability of the devices is around 0.2% for a hold time of 30 ns. The PDP measurements showed a maximum peak value of 20% around 430 nm and the timing resolution was less than 100 ps FWHM for λ =450 nm at an excess bias of 500 mV. SPADs characterization is known to be a very delicate process as their performances is sensitive to many variables (excess bias, temperature, size, etc.) and the importance of the active area value was quite obvious as it affected all the SPAD parameters except for the afterpulsing probability. The DCR was proportional to the devices area, the jitter was improved when the diameter is reduced whereas the PDP decreased for SPAD diameter below 20 µm. In our case, the tradeoff between all the SPAD performance parameters yields the conclusion that the best compromise was obtained with SPADs that have an active area diameter between 15 µm and 25 µm as they showed a reasonable DCR (less than 5 KHz), a good PDP probability (~18%) and a timing jitter lower than 100 ps.

Table 4 - Performances figures of several SPADs fabricated in standard CMOS technologies

	Performance figures			
	DCR	***Afterpulsing***	***PDP***	***Jitter***
This work	**<10 KHz**	**0.2%**	**20%**	**100 ps**
[59]	1 KHz			
[91]	100 KHz		5.5%	
[98]	<43 Hz		40%	77 ps - 120 ps
[99]	13 KHz		17.4%	
[100]	1 MHz			
[101]	<100 KHz			55ps
[102]	600 Hz	2.6%	45%	
[103]	100 Hz	2.1%	30%	170 ps

Chapter 3
The Time to Digital Converter

Time to digital converters (TDCs) are highly precise devices used for time measurements in many applications such as nuclear physics, time of flight measurement (TOF), Positron emission tomography (PET), RF data transmission, digital to analog converters and many more. They operate by converting a time interval defined by two pulses, "*START*" and "*STOP*", into a digital representation.

Time to digital converters are divided into digital and analog architectures, the traditional conversion approach employs an analog circuit to convert a time interval into a voltage which is then translated into a digital code using an analog to digital converter. However, the digital approach have gained more popularity due to the many advantages of digital compared to analog circuits and in particular owing to the facts that unlike analog devices, digital devices benefit from the scaling trend resulting in a more compact, efficient and flexible implementation and that the resulting data is more robust to noise and can be stored easily without any loss of information.

Numerous TDC implementations were presented in the literature in the aim of achieving better performances in terms of Dynamic range, resolution, linearity and power consumption. In this chapter, the main operation parameters of a TDC and the inherent error sources affecting these parameters are listed; the most relevant TDC architectures are then presented and discussed concisely with an explanation of their operation concepts and their main advantages and limitations. Next we present the devised hybrid TDC architecture that combines the concepts of digital counter-based TDCs digital, DLL-based TDCs and TACs to overcome their inherent limitations thus achieving a high time resolution and a wide dynamic range with reasonable power consumption.

3.1 TDC Operation Metrics

A time to digital converter operation is based on converting a time interval defined by the arrival of two consecutive signals (*"Start"* and *"Stop"*) into a digital representation. This quantization leads to an input-output characteristic such that represented in Figure 66 where the input time interval is plotted on the x-axis and the digital representation output is plotted on the y-axis. The conversion principal is very similar to that of an ADC but instead of quantizing a continuous voltage or current level, TDCs quantize continuous time intervals. TDCs share similar characteristics with ADCs: Least Significant Bit (*LSB*) or Timing resolution, dynamic range, quantization error, offset, gain error, non-linearity errors, conversion time, dead time and power dissipation, and these parameters should be taking into account when designing a TDC.

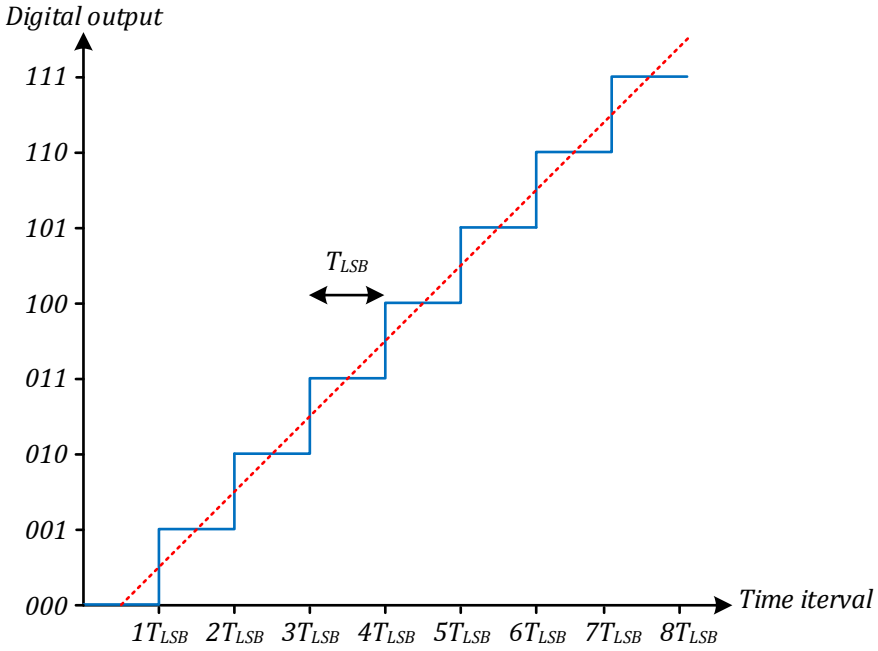


Figure 66 - ideal input-output characteristic of a TDC

3.1.1 Time Resolution

Like all converters, TDC have a least significant bit output value which represents the smallest time that it can discriminates T_{LSB} (Figure 66) also known as the time resolution. Time resolution depends of many factors: the TDC architecture, the used technology and the noise levels.

3.1.2 Dynamic Range

The dynamic range of a TDC is defined as the maximum time interval that can be measured by the TDC. The dynamic range (DR) is given as:

$$\Delta R = 2^N . T_{LSB} \tag{13}$$

where T_{LSB} is the time resolution of the TDC and N is the number of output bits.

3.1.3 Quantization Error

In a TDC, the quantization of continuous time intervals into a binary code results in a divergence between the ideal TDC transfer curve and the measured transfer curve due to mismatch and noise in the TDC electronics. The resulting error is known as the quantization error ε. Unlike ADCs where the quantization error is usually symmetrical around zero ($^{-T_{LSB}}/_2 < \varepsilon < {}^{T_{LSB}}/_2$) , the quantization error of a TDC is not mean free ($0 < \varepsilon < T_{LSB}$).

3.1.4 Offset

In an ideal TDC the first step occurs at the position T_{LSB}, the offset error occurs when the first step is deviated from its ideal value (T_{LSB}) in the time axis (Figure 67) in which case the offset error is defined as:

$$E_{offset} = \frac{T_1 - T_{LSB}}{T_{LSB}} . \tag{14}$$

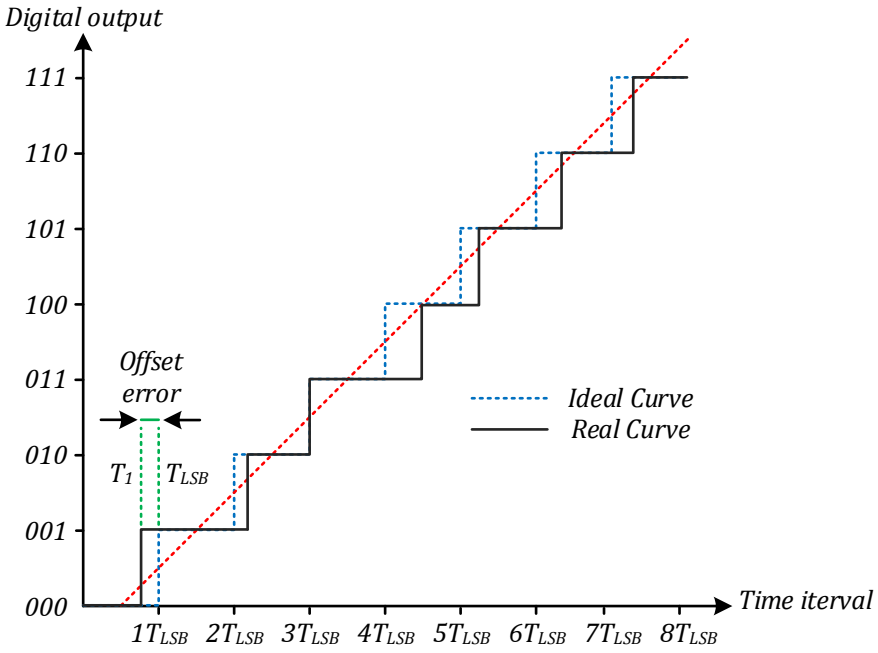


Figure 67 - Input–output characteristic with offset error

3.1.5 Gain Error

The gain of a TDC (G_{TDC}) is defined as the slope of the input-output characteristics or in mathematical terms:

$$G_{TDC} = {\Delta B}/{\Delta T} \tag{15}$$

Where ΔB is the digital output variation for a time interval of ΔT. In an ideal TDC, the gain is defined as:

$$G_{TDC} = 1/T_{LSB} \tag{16}$$

This means that an increment of one T_{LSB} will lead to an increase by one *LSB*. The gain error is equal to the deviation of the last step position from its ideal value in terms of *LSB* after removing the offset.

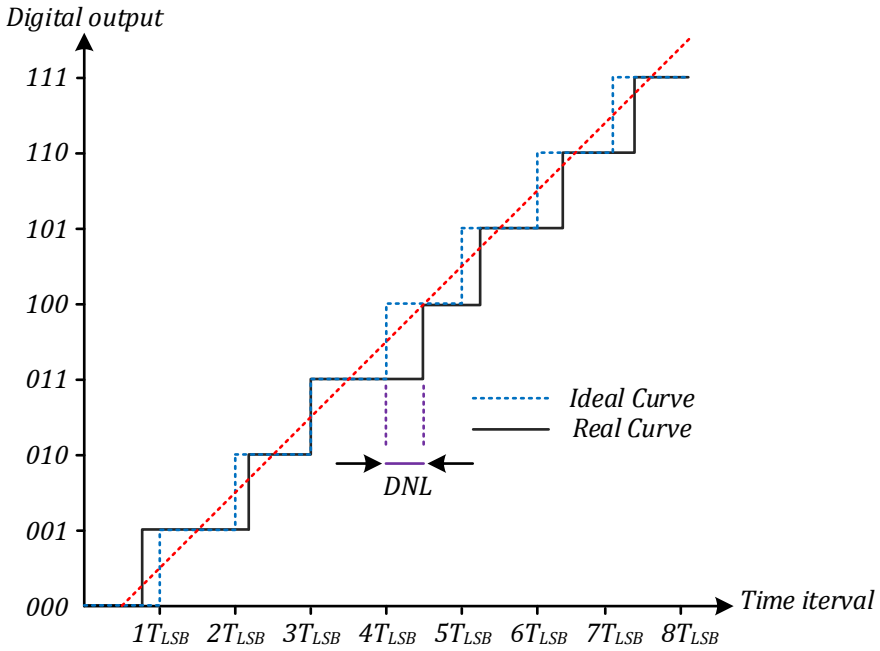


Figure 68 - Input–output characteristic of a TDC with gain error

3.1.6 Non-linearity errors

Non linearity errors are deviations of TDC characteristics from its ideal form that lead into nonlinear distortion. Unlike gain and offset errors, non-linearity errors cannot be simply compensated by adding or multiplying a term to the input time interval. Non-linearity in a TDC is divided into differential non linearity (DNL) and integral non linearity (INL). DNL expressed in LSB represents the deviation of the output bin size from its ideal value of one T_{LSB} normalized to one T_{LSB} (Figure 69):

$$DNL_i = \frac{T_i - T_{LSB}}{T_{LSB}} \tag{17}$$

Where DNL_i is the *ith* value of the DNL, T_i is the width of the *ith* step in the input-output characteristics. The INL expressed in LSB is defined as the defined as the departure of the step position from its ideal value normalized to one T_{LSB} it is defined as the deviation of each step from its ideal value, it is given as:

$$INL_i = \sum_{j=0}^{i} DNL_j \tag{18}$$

As suggest by their names, the DNL gives a microscopic view on the non-linearity while the INL is a macroscopic description of the bending of a converter characteristics.

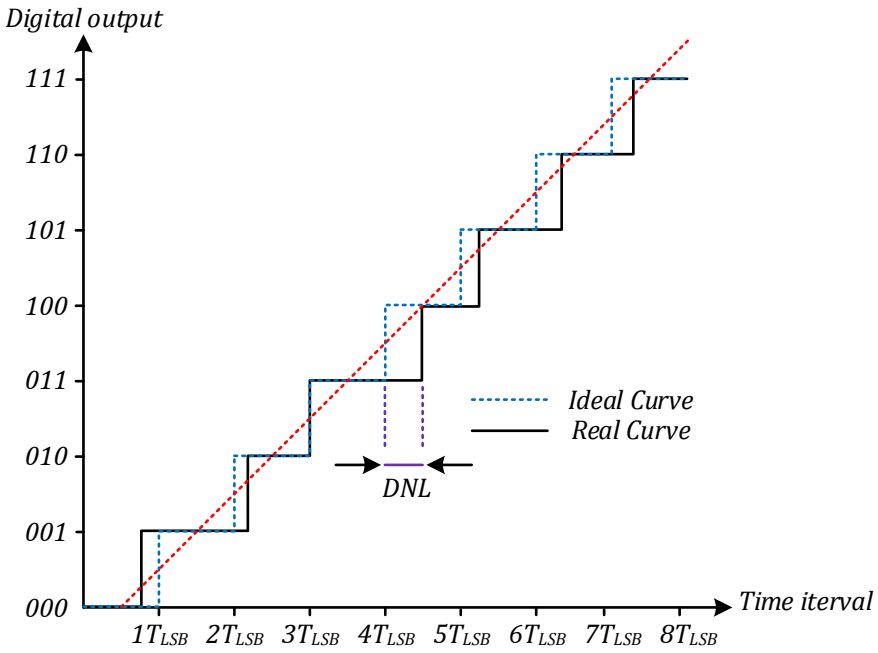


Figure 69 - Input–output characteristic of a TDC with DNL error

3.1.7 Single Shot Precision

Noise and mismatch in a TDC leads to variations in the measured values thus causing a fixed time interval measurements to yield different values. The single shot precision (*SSP*) measures the reproducibility of a TDC measurement in the presence of noise, it represents the standard deviations of a fixed time interval conversion results. The *SSP* is not a constant value but rather a function of time interval *SSP(T),* this can be understood given that longer time measurements lead to more elements contributing to the overall timing uncertainty although this is not exactly true in the case of looped topology as we will see later.

3.1.8 Conversion Time

The conversion time represents the time between the arrival of the "*Start*" signal and the availability of the subsequent measurement results. If the "*Stop*" signal is the reference the delay is called latency.

3.1.9 Dead Time

Most TDC require some time after the end of a measurement before a new one can be started. This is referred to as the dead time given that the TDC cannot be used. Dead time is a very important parameter as it defines the highest measurement repetition rate achievable by the TDC.

3.1.10 Power Dissipation

Power dissipation includes static power dissipation and dynamic power dissipation. The static power dissipation is defined as:

$$P_{static} = V_{dd}.I_{static} \quad (19)$$

Where I_{static} is the total static current consumption and V_{dd} is the power supply voltage. On the other hand, the dynamic power dissipation is determined by the switched capacitor, the clock frequency and the power supply voltage. It is given as:

$$P_{dynamic} = \alpha.C.V_{dd}.f \quad (20)$$

3.2 Analog TDC Architectures

3.2.1 Time to Analog Converter

The traditional time to digital conversion scheme (Figure 70) consists of a time to amplitude converter (TAC) followed by a conventional analog to digital converter (ADC). Two "*Start*" and "*Stop*" signals are used to generate a pulse which transformed by the analog integrator into a voltage. The voltage is fed to an ADC that transforms it into a digital code. The TAC allows for very good resolutions despite a limited dynamic range and the operation principle of this TDC is quite simple but the performance of the overall TDC (the TAC followed by the ADC) are usually limited by the used ADC especially in terms of resolution, dynamic range, dead time, occupation area and power consumption. Furthermore the TAC which is based on current integration in a capacitor suffers from weak linearity and tends to be difficult to calibrate and stabilize as it is sensitive to noise, process and temperature variations.

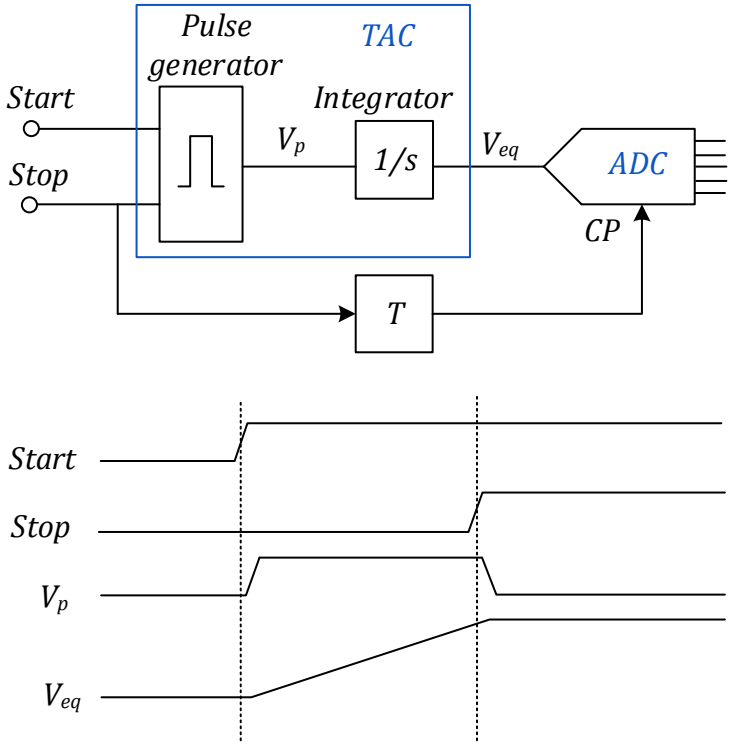


Figure 70 - Block and signal diagram of a time to analog converter

3.2.2 Dual Slope TDC

One problem of the traditional analog TDC approach is its sensibility to noise and temperature variation which makes it hard to stabilize thus requiring precise calibration to determine the exact current and capacitor values, a tedious task in the case of an integrated TDC. This can be avoided using a dual slope TDC.

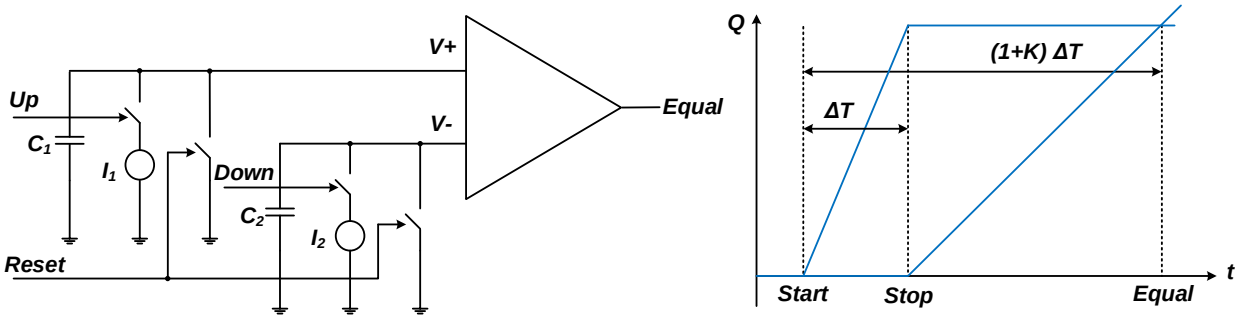


Figure 71 - basic time stretcher realization and the corresponding timing diagram

Dual Slope TDC is based on a technique known as time stretching where the time to be measured is stretched by a factor K. Like the traditional analog approach a pulse is generated by the *"Start"* and *"Stop"* signals and converted into a voltage by a first integrator. Upon the arrival of the *"Stop"* signal a second integrator with a lower integration factor k starts integrating until a comparator signals that its output is equal to the first one. One way of realizing a time stretcher is represented in Figure 71 and the operation principal of such circuit is very similar to a Wilkinson ADC. At first the 2 capacitors are discharged by the reset devices then the first capacitor *C1* is charged by the current source *I1* for the time interval *ΔT* determined by the 2 signals *"Start"* and *"Stop"*. Upon the arrival of the *"Stop"* signal *C1* becomes floating and *C2* is charged by the second current source *I2* until the comparator connected to the 2 capacitors signals that the two voltages;

$$V_+ = \frac{I_1}{C_1}.\Delta\mathrm{T} \text{ And } V_- = \frac{I_2}{C_2}.\Delta\mathrm{T}_2 \qquad (21)$$

where *ΔT2* is the time elapsed since the arrival of the "Stop" signal, are equal. This happens at:

$$\Delta T \ + \Delta\mathrm{T}_2 = 1 + \frac{C_2}{C_1}\frac{I_1}{I_2} = (1 + \mathrm{M.N}) \qquad (22)$$

where $M=\frac{C_2}{C_1}$ and $N = \frac{I_1}{I_2}$.

3.3 Basic TDC architectures

Traditional time converters were based on analog techniques but with the VLSI technology progress digital solutions became more popular, this is due to the many advantages of digital circuits compared to analog circuits. Indeed digital circuits offer more robustness against noise and process variations, they are also easier to implement and the resulting data can be easily stored without loss of information. But mostly, unlike analog devices, digital devices have benefited greatly from the technology scaling trend which led to area reduction, lower power consumption and faster devices due to the reduction in gate delays. This is particularly advantageous to functionalities related to the time domain which is the case with time to digital converters. In this section the main fully digital TDC architectures are presented with an explanation of their operation principle, their advantages And Possible Limitations.

3.3.1 Counter Based TDC

The basic technique to measure a time interval is using counter to determine the number of periods elapsed between the *"Start"* signal and the *"Stop"* signal arrivals as shown in Figure 72. This approach is quite simple and allows measurements over a wide dynamic range limited only by the used counter but the asynchronous nature of the "Start" and "Stop" signals lead to a measurement error $\varepsilon_T < 2.Tclk$. Furthermore the achieved resolution is limited by the used clock frequency and good time resolutions require the use of very high frequency rates (a 1 ns time resolution requires the use of a 1 GHz clock frequency) as a result this approach is not suitable for precise time measurements.

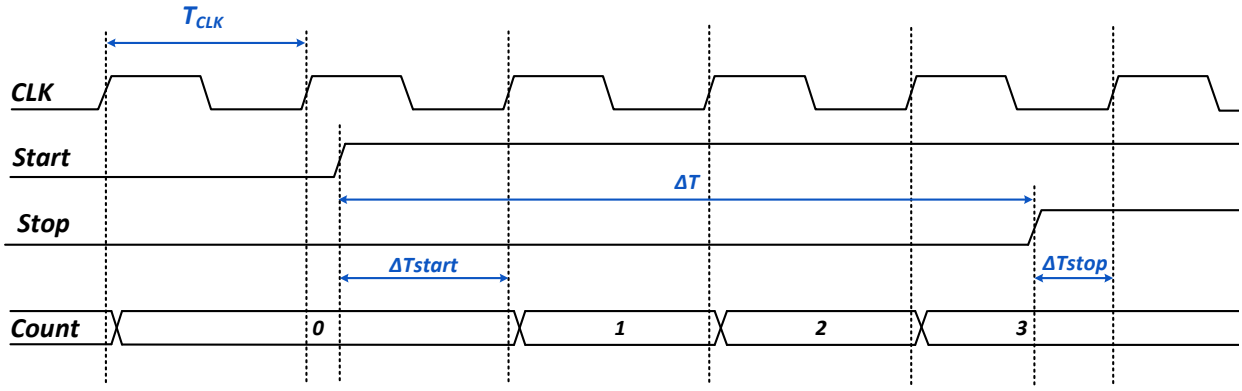


Figure 72 - operation principle of a counter based TDC [104]

3.3.2 Delay Line Based TDC

The use of very high clock frequency that limits counter based TDC can be avoided by taking the gate delay as a measurement unit which can allow time measurements with less than 100 ps time resolutions. A delay line based TDC is shown in Figure 73, the "Start" signal is propagated through a delay line where a buffer is used as a delay element, the output of

the buffers are connected to Flip-Flops and the intermediate taps of the delay line are sampled on the rising edge of the "*Stop*" signal. The obtained thermometer code and in particular the position of the high to low transition indicates how far the "*Start*" signal have propagated during the time interval between the arrivals of the "*Start*" and "*Stop*" signals. The resolution obtained using this technique is only limited by the used technology; it is equal to the buffer gate delay which is equal to 2 times the inverter delay. The resolution can be theoretically increased by choosing an inverter as a delay element but such approach is not practical due to rise and fall times mismatch as well as a more complex circuitry to evaluate the thermometer code.

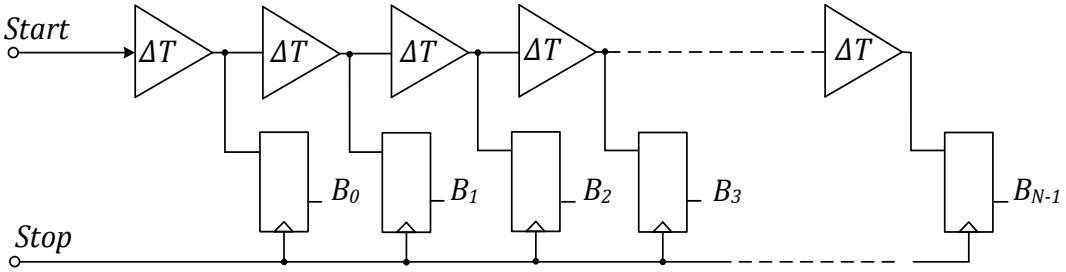


Figure 73 - Implementation of a Delay line based TDC using a buffer as a delay element

In general this technic is simple, efficient and fast but it also suffers from several limitations as it is sensitive to process, voltage and temperature variations (PVT variations) leading to measurement errors. These errors could also arise from skews in the buffer line driven by the "*Stop*" signal. Another limitations is the limited dynamic range since the measured time interval cannot exceed $N\times \Delta T$ where N is the number of buffers in the delay line. A reasonable dynamic range requires a big number of delay elements leading to a large area and an increase in power consumption. The limited dynamic range can be further expanded using a looped structure (Figure 74) where the delay line is bent into a loop such as the start signal can go through it several times, a counter is added to keep track of the number of times the delay line has been passed by the "Start" signal. This approach requires careful layout design to limit nonlinearities due to loop asymmetry, it's also crucial to avoid any *LSB* error in the counter output which would translate into an error of $N\times \Delta T$ leading to an enormous time measurement error.

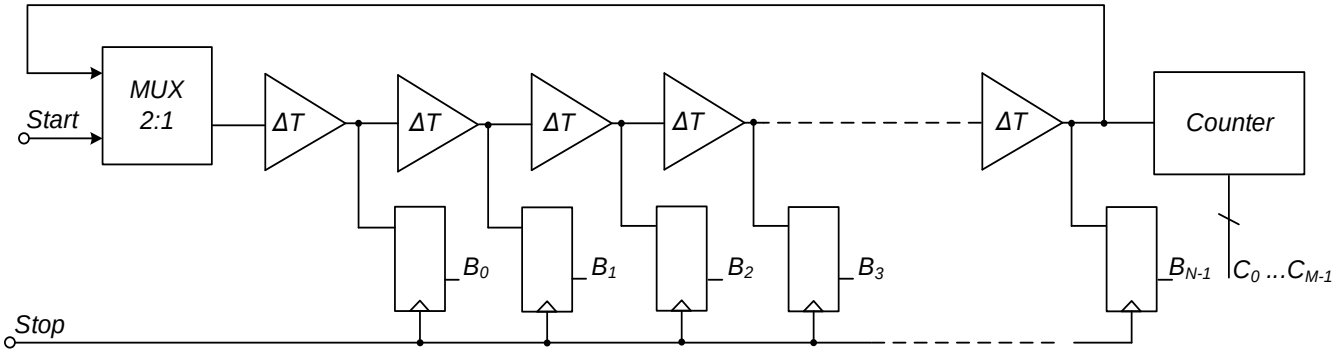


Figure 74 - schematic representation of looped TDC architecture

3.3.3 Delay Locked Loop Based TDC

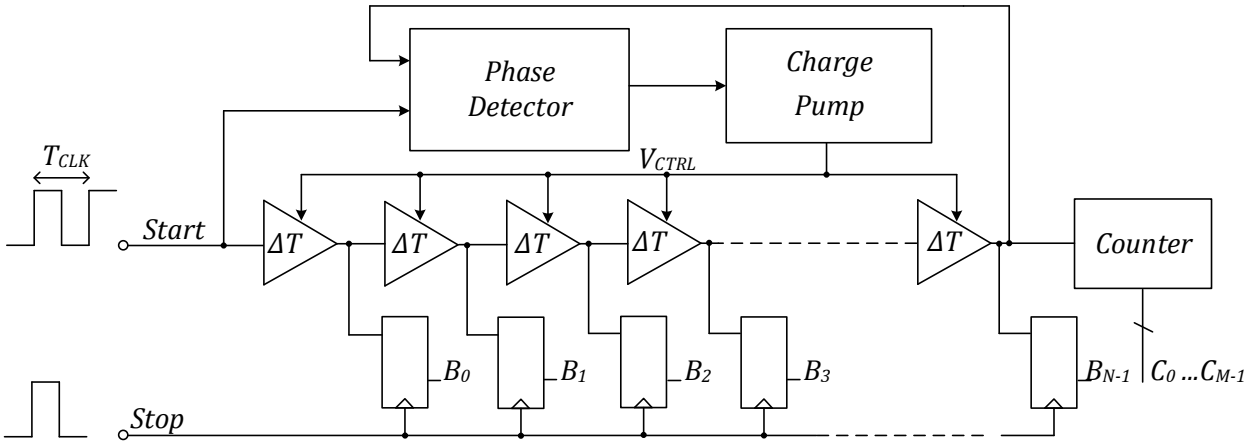


Figure 75 - Implementation of a DLL based TDC

Digital TDC presented so far are based on delay elements which make them vulnerable to process variations, temperature and supply voltage (PVT variations). As a result absolute time measurement cannot be done without a calibration phase to determine the exact gate delay of the used delay elements; this can be avoided using a delay locked loop (DLL) based TDC architectures. A DLL based TDC implementation is presented in Figure 75, a periodic "Start" signal is injected into a delay line where basic inverters are replaced by current starved inverters (Figure 76). The signal at the end of the delay line is compared to the periodic "*Start*" signal using a phase frequency detector (PFD) which provides a signal proportional to the delay between the 2 signals. This signal is used to control the delay of the tunable buffers by means of a charge pump that control the gate voltage of the current starved inverters. The measurable delay along the line is equal to the reference clock period but a wider dynamic range is achievable using a counter synchronous to the reference clock providing the "*Start*" signal. Since the two signals are provided by the same reference the resulting thermometer code represent the time measurement result in reference to the clock period.

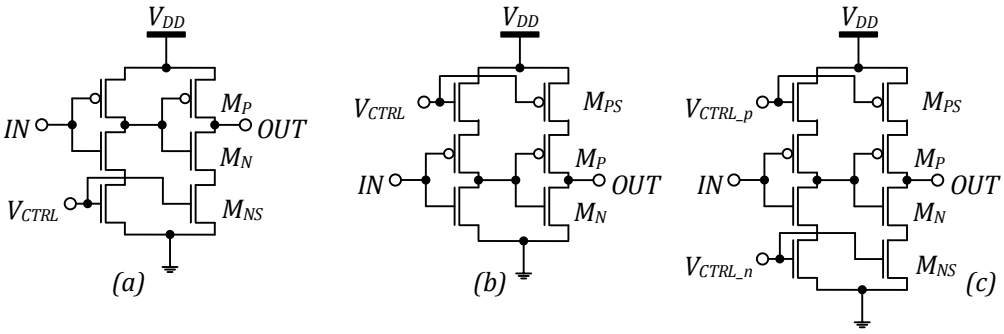


Figure 76 - The 3 types of current starved inverters used as tunable delay elements in the DLL based TDC: N-side (a), P-side (b), and both (c)

Using this DLL based TDC approach; the sensitivity to environmental conditions is eliminated resulting in high resolution time measurement on a wide dynamic range. However the DLL is susceptible to other problems such as jitter, noise and delay cells mismatch but the impact of these factors can be limited using a stable reference clock and careful layout and power distribution scheme.

3.4 Advanced TDC Architectures for Sub-Gate Resolutions

In Basic TDC architectures, the time interval is measured as a multiple of a gate delay which is limited by the inverter delay in the used technology. As a result the achievable resolution is limited only by the used technology and higher resolutions require the use of faster technologies. However it is possible to escape the limitations of the used technology by adapting more complex TDC architectures thus achieving delays lower than the gate delay or sub-gate delays.

3.4.1 Parallel Delay Elements Based TDC

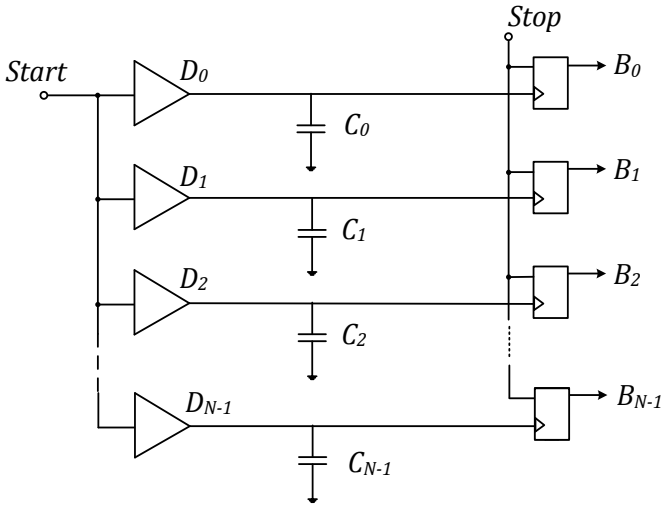


Figure 77 - Implementation of a parallel delay elements based TDC

A basic implementation of parallel delay elements based TDC is represented in Figure 77. In this architecture, the "*Start*" signal is connected to parallel lines of delays elements with a Δt step gradually increasing delay equal to the TDC resolution. Upon the arrival of the "Stop" rising edge, the parallel delay elements are sampled through a buffer line. The delay of each line is fixed by scaling the load capacitance which can be done with a very fine step; as a result the delays can be tuned quasi-continuously resulting in a sub-gate resolution. The conversion principle is quite straight forward and the measurements results are immediately available after the arrival of the "*Stop*" signal rising edge, but this conversion technique suffers from several inconveniences in particular sensitivity to local variations and an unsuitability to high dynamic range as it leads to a high occupation area. Furthermore, this architecture requires a very careful design to minimize skew problems in the "*Start*" and the "*Stop*" nets.

3.4.2 Pulse Shrinking TDC

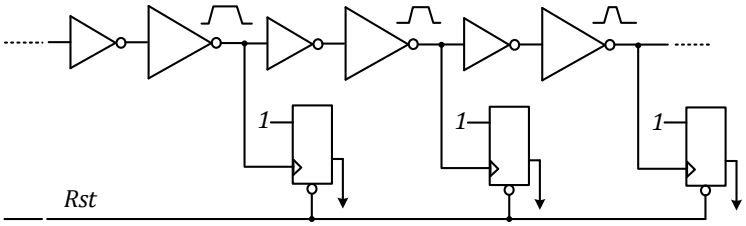


Figure 78 - Cut-out of a pulse shrinking TDC line showing the pulse width reduction as the pulse propagates

In the pulse shrinking TDC a pulse with a width defined by the arrivals of the two signals "*Start*" and "*Stop*" is propagated through a delay line with an intentionally asymmetric rise and fall times in each stage, as the pulse propagates through the delay elements its width shrinks thus becoming smaller and smaller until it disappears (Figure 78). A line of flip flops with a constant one connected to their data inputs had their clock inputs connected to the delay line taps, as the pulses propagates the flip flops outputs switches to 1 however when the pulse vanishes the remaining flip flops stop switching and the position where the data outputs goes from 1 to 0 determine the time interval measurement. The T_{LSB} is defined by the reduction factor of the original pulse after each delay stage which is the result of the asymmetrical rise and fall times thus allowing a sub gate delay resolution (Figure 79).

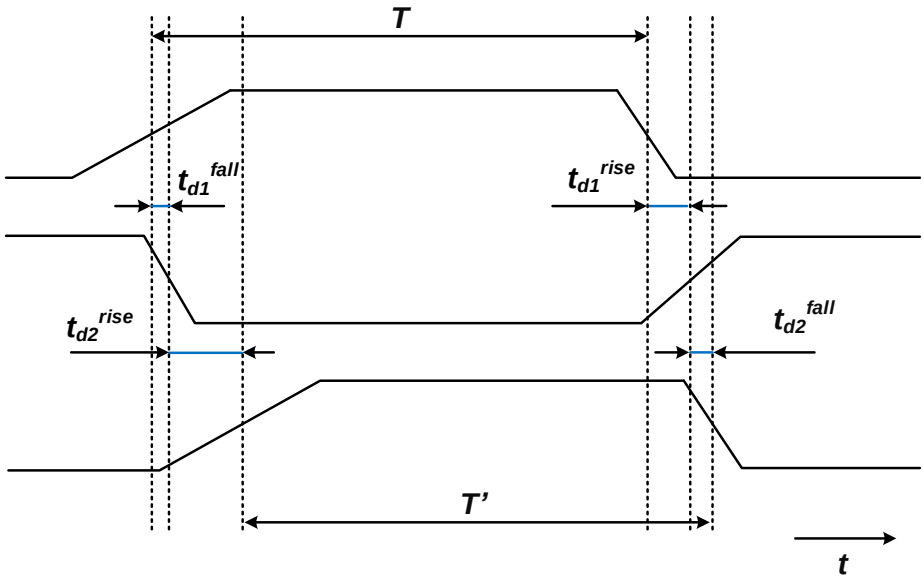


Figure 79 - Operating principle of the pulse-shrinking TDC. Input signal (top), intermediate signal (middle), and output signal (bottom) of the pulse shrinking element [104]

This architecture offers sub-gate resolution without the need for complex circuitry but like most delay elements based TDC it is sensitive to PVT variations resulting in measurements nonlinearities furthermore the dynamic range is determined by the number of stages as well as the time resolutions. As a result there is a trade-off to be made between the timing resolution and the maximum measurable time interval in order to limit the nonlinearities as well as the area and power consumption. A looped structure can also be used to further reduce these effects although it only results in a moderate reduction as the number of the shrinking elements in the loop must be sufficient so that the maximum complete pulse fit inside the loop.

3.4.3 Vernier Delay Line

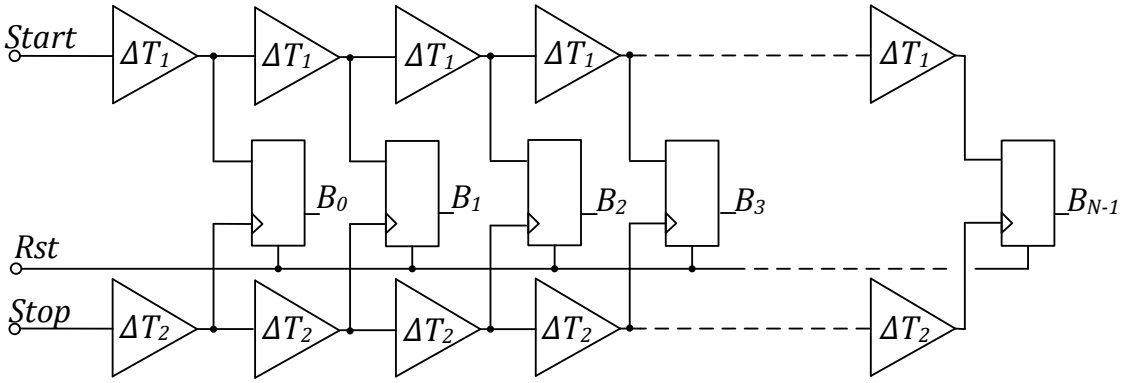


Figure 80 - Implementation of a Vernier TDC with a delay line for the start signal (top), a delay line for the stop signal (bottom) and an early late arbiter line (middle)

The Vernier TDC (Figure 80) is a delay line based TDC, it consists of two delay lines, one of the start signal and one for the stop signals. The delay of the elements in the start delay line (ΔT_1) is slightly higher than the delay of the elements in the stop delay line (ΔT_2). When the "Start" signal occurs it propagates along the slower delay line then when the "Stop" signal ensues it propagates along the faster delay line in a chase for the "Start" signal catching up by $T_{LSB} = \Delta T_1 - \Delta T_2$ at each stage. An line of early late arbiters usually consisting of a line of flip flops detects the position "n" where the 2 signals become in phase and the time interval between the "Start" and "Stop" signals is determined as $n \times (\Delta T_1 - \Delta T_2)$. This technique offers a very good resolution equal to $T_{LSB} = \Delta T_1 - \Delta T_2$ but the total dynamic range is very limited as it is equal to $N \times (\Delta T_1 - \Delta T_2)$ where N is the number of stages in each line thus requiring the use of long delay lines to achieve reasonable dynamic ranges which leads to a high power and area consumption. Another drawback of the Vernier TDC is its high sensitivity to delay matching and PVT variations as it is the case in other delay lines based schemes. However, this is even more critical in the Vernier TDC due to the high resolution and the additional variability of the second delay line. Lower area consumption and a bigger dynamic range can be achieved by substituting the linear structure by a loop structure.

In a looped structure (Figure 81), the 2 delay lines are both configured in a ring like structure through a multiplexer while 2 counters are used to determine the number of times each signal went through its loop. This topology is quite effective in increasing the dynamic range while limiting the occupation area but the loop structure is also the source of many drawbacks a careful design of the layout is thus required to limit problems due to asymmetry and non-linearity which is already challenging even in the case of a single looped delay line. In general, a looped structure results in a degradation of the linearity compared to a linear structure. Another consequence of the closed loop structure is the integration of all the timing errors that may occurs during the measurement time which leads to the degradation of the system accuracy especially for higher time intervals. Finally, to limit the effects of PVT variations and possible coupling between the two delay lines, a double DLL – control loop can be used for calibration, but in high repetition rate measurements this can only be done between measurements and will result in a long dead time. Another approach is to use DLL controlled dummy lines to derive control information needed to stabilize the delays. Both approaches result in an increase of power and area consumption and it's safe to say that despite being an elegant approach, the Vernier TDC is quite hard to implement.

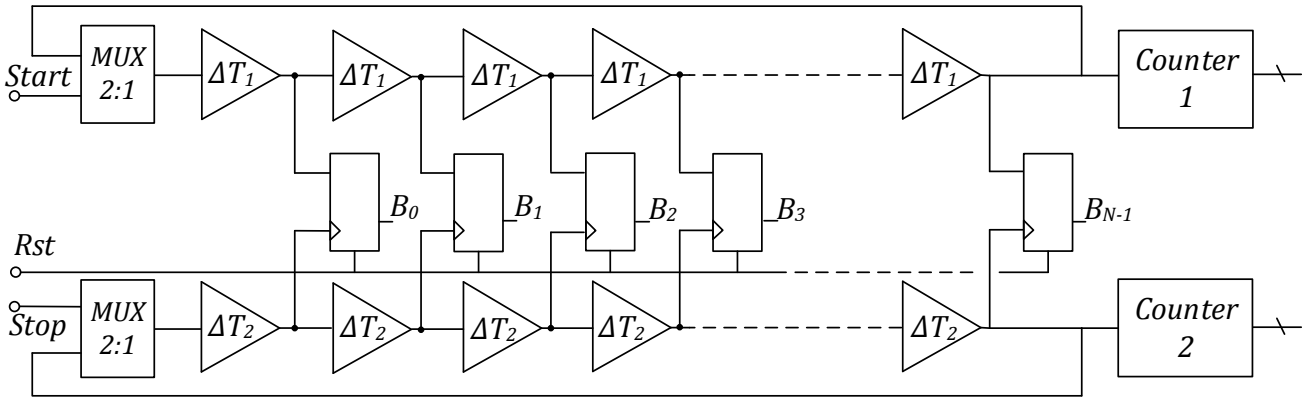


Figure 81 - Implementation of a looped Vernier TDC

3.4.4 DLL Arrays

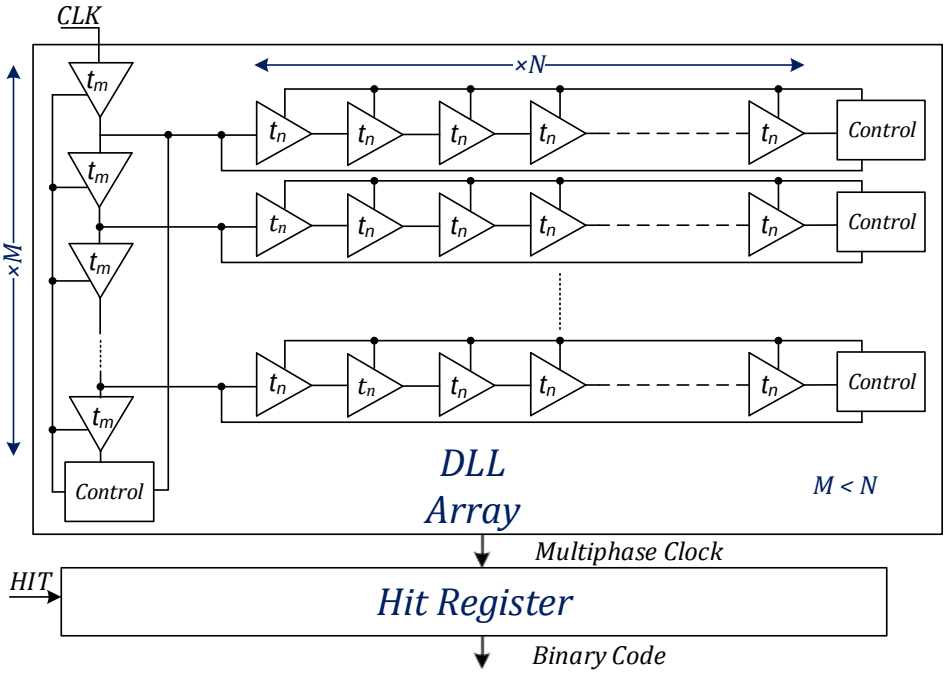


Figure 82 - Block diagram of a time-to-digital converter based on DLL arrays

A very effective scheme to achieve sub-gate resolution is time interpolation using DLL arrays. A bloc diagram of a TDC based on this method is illustrated in Figure 82, a reference clock is propagated through a DLL of M stages with a delay t_m driving an array of smaller DLLs with N stages and an identical time delay t_n . This TDC offers a time resolution equal to the difference between the two time delays

$$\Delta T = t_m - t_n = T_{CLK}\left(\frac{1}{M} - \frac{1}{N}\right) = \frac{T_{CLK}}{N.F} \quad (23)$$

Where F is the number of DLLs in the array [105]. By choosing t_m slightly larger than tn a very good time resolution in the order of few picoseconds can be achieved. However, a minor drawback of this scheme is the fact that the result is not obtained in a pure binary unit of $1/2^N$ but rather in a unit of $1/N.F$, the output is therefore a pseudo-binary code which imposes an additional processing step to convert it into binary form. Like other delay lines based TDCs, this method is prone to jitter and offset which limits its accuracy. Furthermore, a careful layout is crucial to minimize mismatch and asymmetry; this is especially important due to the use of several DLLs. The use of several DLLs also results in large static power dissipation consequently low power design should be strongly considered to limit power consumption. An example of interpolation by phase shifting in an array of DLLs is shown in Figure 83. In this example, M=28, N=35 and F=4. With tm = 5×ΔT and tn = 4×ΔT, the resulting time resolution is equal to ΔT.

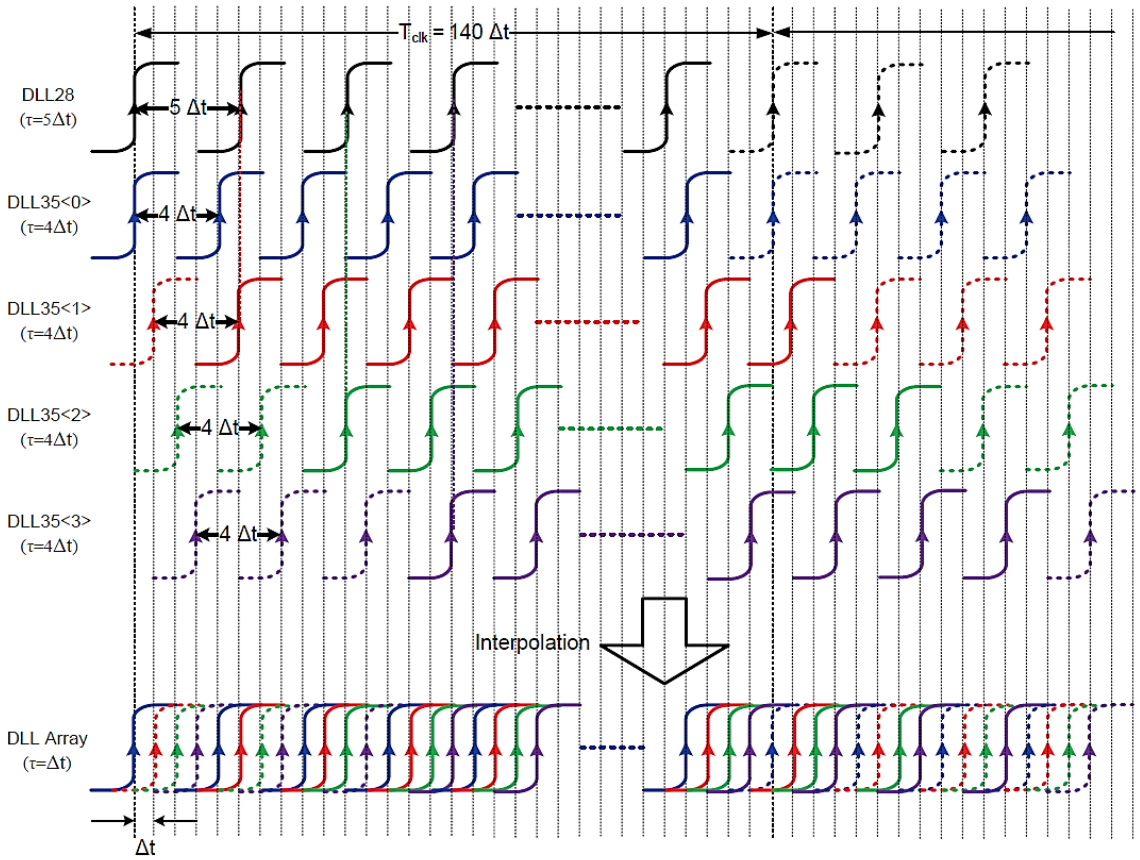


Figure 83 - Interpolation by phase shifting in an array of DLLs. [105]

3.4.5 Gated Ring Oscillator TDC

Like other oscillator-based TDC, the Gated ring oscillator TDC (GRO TDC) uses the number of delay element transitions occurring during the measured interval to measure its length. However unlike traditional oscillator based TDCs these transitions can only occur during the measurement period as the oscillator is only enabled during the measurement interval and disabled otherwise thus preserving the ring oscillator state between measurements. The architecture and operation principle of the GRO TDC is depicted in Figure 84. Initially the gating signal “Enable” is set to low as a result the Gated ring oscillator (GRO) is disabled and the counters are reset. Upon the beginning of the measurement period, the gating signal “Enable” is set to “1” thus activating the GRO and its outputs are used to drive the counters that compute the oscillator phase transitions within the gated operation period. The total sum of all counters is obtained via the binary adder and the result is proportional to the measured time interval.

The gating functionality in the GRO-TDC is particularly advantageous as it allows the first-order shaping of the quantization error and the mismatch errors by passing them to the next measurement where they can be deduced from the result [106]. Consequently, the GRO-TDC is capable of achieving high resolution and linearity without the need of calibration although it still requires continuous calibration to mitigate elementary delay mismatch due to process and temperature variations. The fact that the resolution of the GRO TDC is independent from inverters mismatch makes this architecture very suitable for sub-ps TDC with an achievable time resolution as high as 100 fs [106]. Furthermore GRO-TDCs are mainly realized by digital blocks thus making them capable of achieving low static power dissipation and more suitable for technology scaling. The gating operation principle that makes it possible to achieve very high time resolution leads also to many drawbacks. For example the floating output nodes during the off-state make the GRO-TDC particularly vulnerable to electronic noise (such as switching noise and cross-talk) as well as leakage current and charge sharing which degrade the system's linearity. Other possible problems includes premature reset of the counters [105] gating skew error, power supply coupling non-oscillation, metastability, and electronic noise thus requiring a careful design and layout to mitigate these problems.

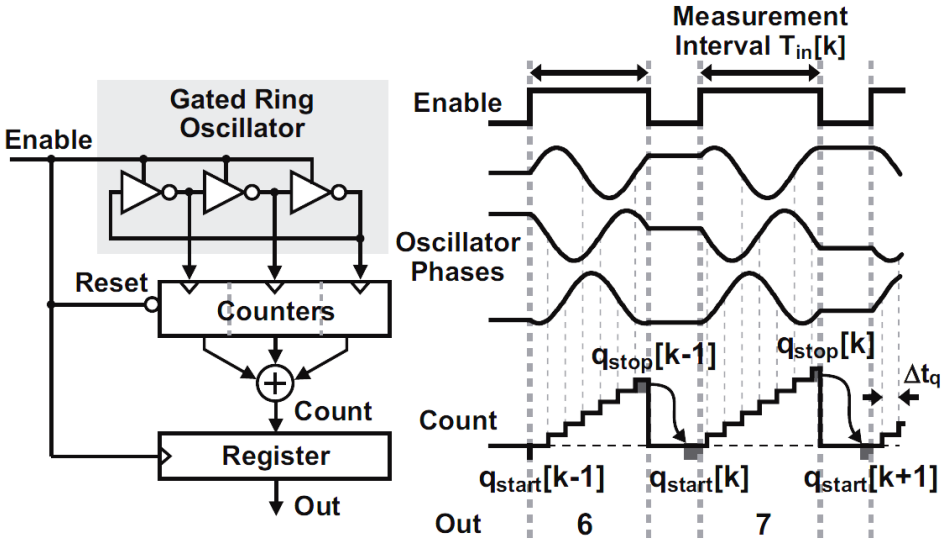


Figure 84 - Concept of the gated ring oscillator TDC [107]

3.4.6 Local Passive Interpolation TDC

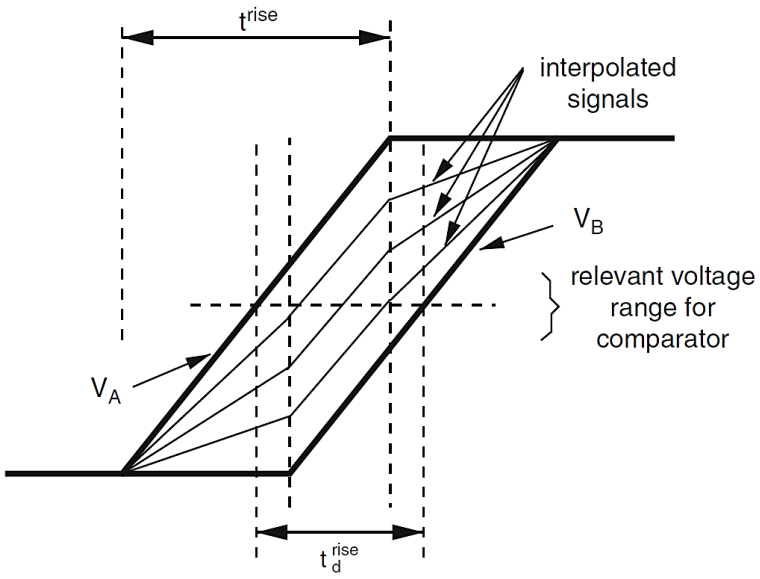


Figure 85 - Principle of local passive interpolation TDC [104]

In the local passive interpolation TDC (LPI-TDC) scheme the coarse resolution of the conventional delay line TDC is increased by subdividing the delayed signals using resisting dividers between delay cells resulting in an interpolation signal giving by :

$$V_{int,i} = V_B + a_i \times (V_A - V_B) \quad (24)$$

with $0 < a_i = i/IF < 1$ and $i = 1...IF-1$ (Figure 85). V_A and V_B are the two delayed signals with an inverter delay skew, a_i is the interpolation coefficient and IF is the interpolation factor (*IF=4* in the example). A basic implementation of a *LPI-TDC* with *IF* = 4 is shown in Figure 86 , the interpolation principle requires that the delay is small enough compared to the rise time. To fulfill this condition, a differential delay-line is used to propagate both the "*Start*" and the inverted "*Start*" signals. Each delay elements includes just two parallel inverters and after each stage an inverted and a non-inverted copy of the "*Start*" signal is available and the signals of any two adjacent stages have a skew of only one inverter delay. The interpolation is done by resistive voltage dividers connected between the delay cells. Interpolates can be based on ohmic resistors as well as transistors and diodes can be used but resistors are more suitable due to better linearity and symmetry as well as the absence of a threshold voltage. Upon the arrival of the "*Stop*" signal, the state of the delay line is sampled is sampled by differential comparators [104] connected to the pairs of complementary signals (R_i and F_i) to obtain sub-gate resolution results. The proposed *LPI-TDC* allows time measurement with sub-gate resolution, no latency and low dead-time and without the need for long delay lines. The use of interpolators are based on resistive voltage dividers make this scheme robust

against global variations, however resistance values must be chosen wisely to achieve reasonable power consumption and RC delays with acceptable linearity .

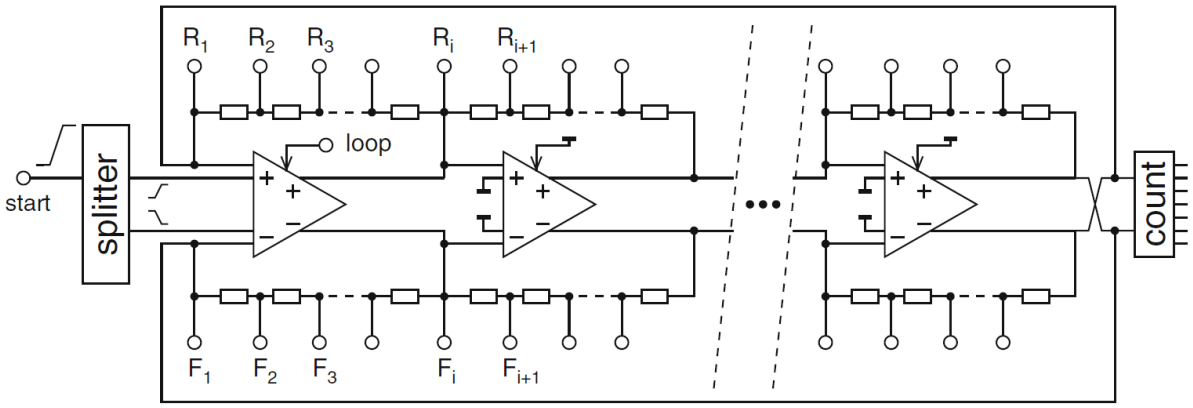


Figure 86 - Basic implementation of a local passive interpolation TDC with an interpolation factor IF =4 [104]

3.5 The Hybrid Time to Digital Converter

The work achieved during this thesis aimed toward designing a prototype of a Streak Camera based on a TCSPC system. TCSPC systems require the use of a TDC or a TAC for precise time measurement however a Streak camera based on a TCSPC system (SC-TSCPC) will require the use of a TDC array that englobes several TDCs, as a result in addition to the traditional requirements of high resolution, a large dynamic rate and a high repetition rate the designed TDC that represents the elementary block of the time measurement unit in the SC-TCSPC should be easily scalable with reasonable power consumption and area occupation. In order to meet these requirements we devised a hierarchical TDC scheme that employs several TDC architectures in order to combine their advantages and overcome their limitations. Indeed, Conventional TACs offer a very good resolution with low power consumption but they represent a limited dynamic range. Alternatively digital DLL-based TDCs offer a reasonable dynamic range but it comes with the price of a poor time resolution often determined by the delay of an inverter cell in the used technology. Array-DLLs architectures have been proposed to improve resolution but they require high power consumption and a large chip area. The proposed hybrid TDC combines the two concepts of digital DLL-based TDCs and TACs to overcome their limitations thus achieving a high time resolution and a wide dynamic range with reasonable power consumption.

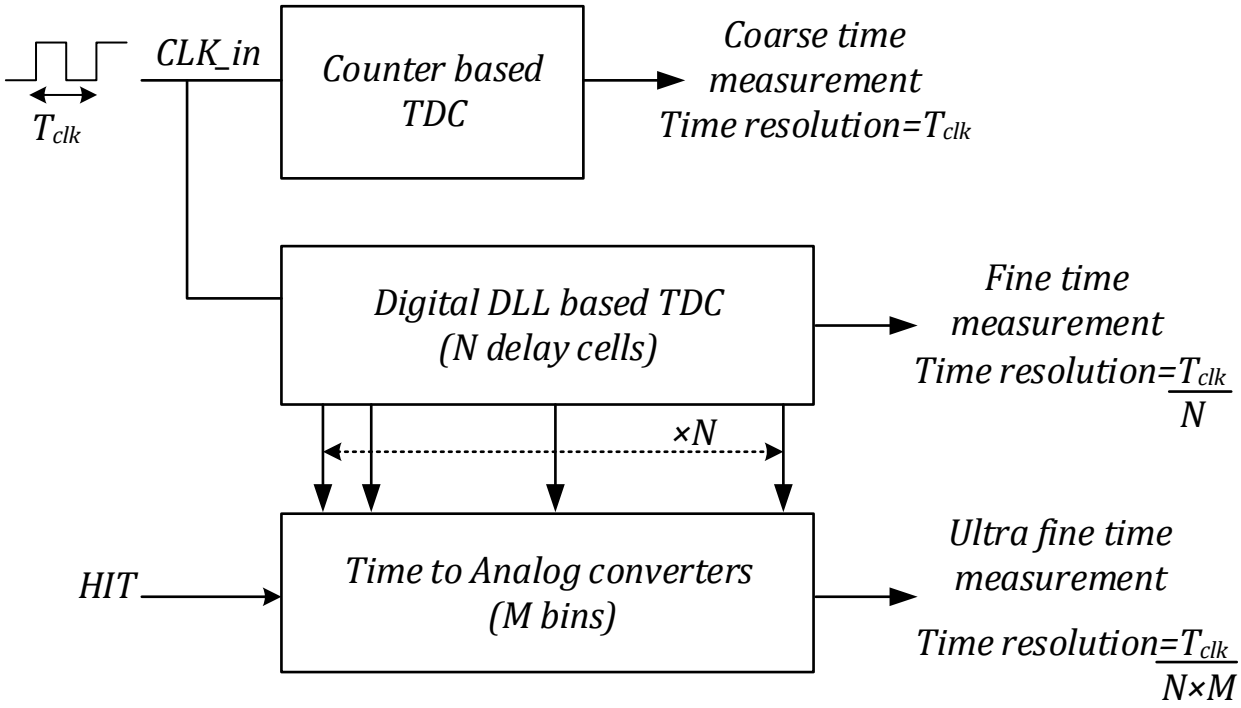


Figure 87 - The 3 stages of the proposed hybrid TDC

The hybrid TDC can be represented as a 3 stage time converter (Figure 87) where the first stage is a counter-based TDC that performs coarse time measurements with a resolution equal to its clock input's period (T_{clk}) and offers a high dynamic range. The second stage is a DLL-based TDC that delivers a fine time measurement with a higher resolution of $T_f = \frac{\text{Tclk}}{\text{N}}$ where N is the number of cells in the DLL's voltage controlled delay line (VCDL). The last stage consists of a TAC where the N DLL's taps are used to charge a line of capacitors so that when a "*HIT*" signal occurs their voltage is sampled. The edge slope is quantized into M equal bins and the stored voltage of the sampling cell where a '1' to '0' transition occurred is used to get the Ultra-fine time measurement resulting in a total time resolution equal to $T_{uf} = \frac{\text{Tclk}}{\text{M}\times\text{N}}$.

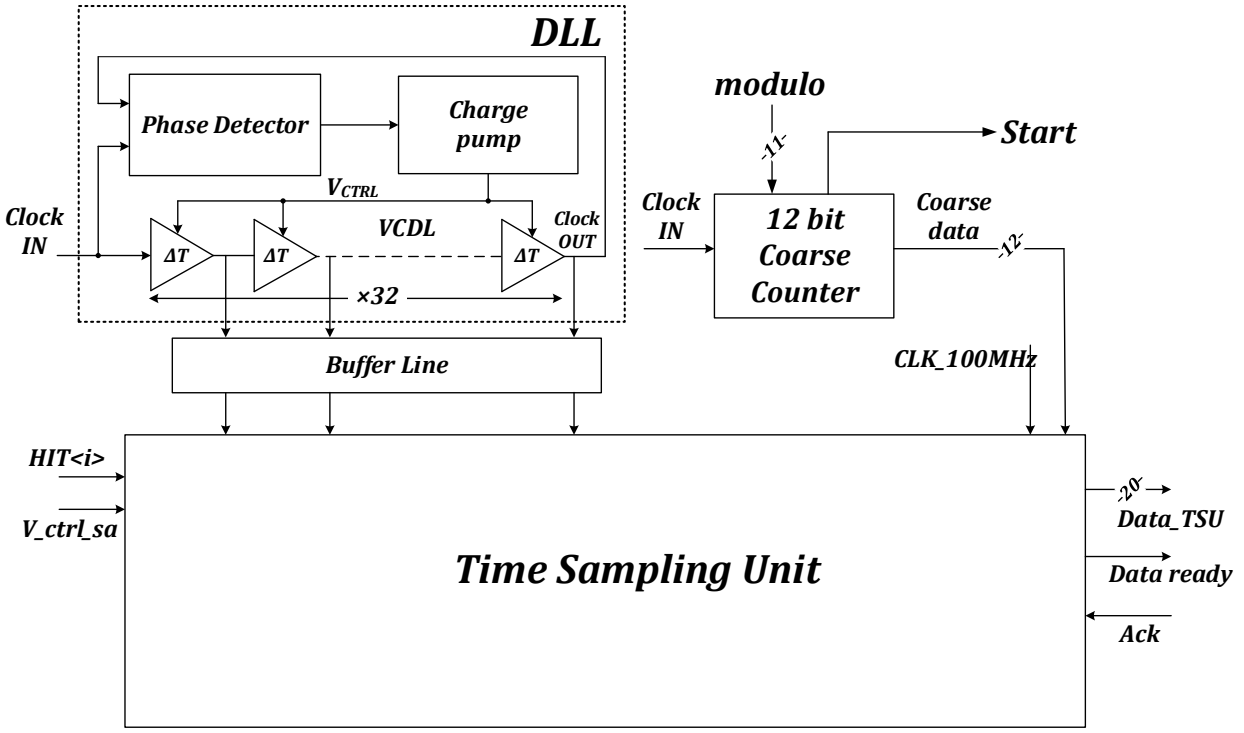


Figure 88 - Block diagram of the hybrid TDC

The block diagram of the hybrid TDC is represented in Figure 88, it consists of 3 main parts, a digital 12-bit counter, a DLL and a Time sampling unit (TSU). A 400 MHz clock input is fed into the coarse counter to deliver a coarse time measurement with a time resolution equal to the clock period (2.5 ns), a DLL with a 32 stage VCDL and a time sampling unit that reads the 32 VCDL to deliver a T_f ≈80 ps resolution fine time measurement and inspects the stage where the high to low transition occurred to obtain an ultra-fine time measurement with a T_{uf} ≈ 10 ps time resolution. This chapter is divided into three sections. In the first section the hybrid TDC architecture is thoroughly represented and discussed, in the second one the experimental results are presented and finally a summary is given at the end of this chapter.

3.5.1 Delay Locked Loop

Traditional DLL consists of three main blocks (Figure 89): a voltage controlled delay line (VCDL), a phase detector and a loop filter. The loop filter typically consists of a charge pump connected to a loop capacitor, the charge pump is considered as a source current delivering a current proportional to error signal which is integrated by the loop capacitor thus converting the phase difference into a voltage signal controlling the delay of the VCDL. A clock input (*CLK_in*) is propagated through the VCDL, the phase detector compares the output of the VCDL (*CLK_out*) to its input and delivers an error signal proportional to the phase difference between the two signals. The produced error signal is then filtered and used as a voltage control signal to gradually reduce the phase difference between "*CLK_in*" and "CLK_out" until it is equal to 2π and the delay of the VCDL is equal to the input clock's period in which case the DLL is said to be in a "locked" state.

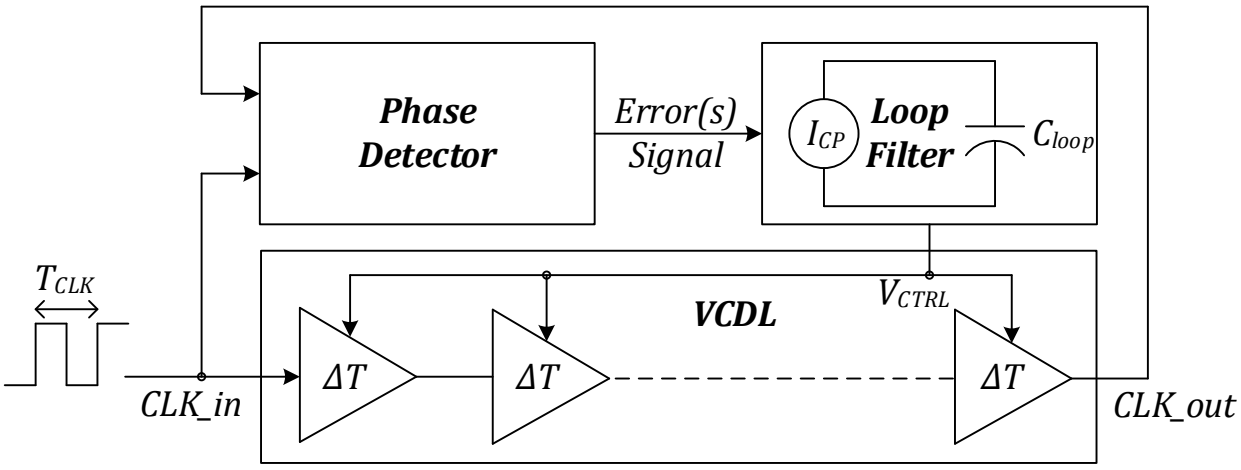


Figure 89 - Block diagram of a traditional DLL

The DLL's operation can be analyzed with a continuous time approximation, a valid approximation as long as the DLL's lower bandwidth limit is at least 10 times smaller than the operating frequency. Figure 90 shows a simplified block diagram of the DLL in the Laplace domain where T_{REF} is the period of the input reference clock, K_{PD} is the phase detector gain (second/rad), I_{CP} is the charge pump current (A), F(s) is the loop filter transfer function and K_{VCDL} is the VCDL gain proportional to the number of delay cells (rad/V). The Loop filter usually consists of a single capacitor. Thus, the transfer function F(s) can be written as

$$F(s) = \frac{1}{s.C_{Loop}} \quad (25)$$

As a result, the closed-loop behavior of a DLL can be characterized in the steady locked state by a first-order transfer function given by:

$$T(s) = \frac{D_o(s)}{D_i(s)} = \frac{1}{1 + s.w_n} \quad (26)$$

Where w_n represents the close loop band-width defined as:

$$w_n = \frac{I_{cp}.K_{PD}.K_{VCDL}}{2\pi.C_{Loop}} w_{REF} \quad (27)$$

With w_{REF} is the frequency of the input reference clock. Based on these equations a conventional DLL is a single-pole system and therefore is unconditionally stable. Furthermore, as long as the term $\frac{I_{cp}.K_{PD}.K_{VCDL}}{2\pi.C_{Loop}}$ is kept constant, the loop bandwidth ω_N can track the operation frequency ω_{REF}. Such tracking can be advantageous because as the operation frequency ω_{REF} increases, the loop bandwidth ω_N also increases which results in a faster acquisition speed. Although a narrower loop bandwidth can help attenuate more high-frequency input noise.

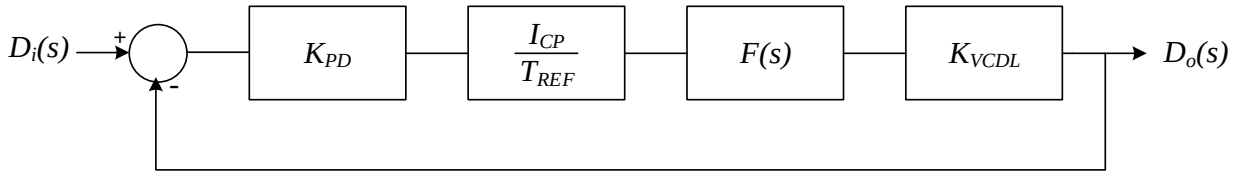


Figure 90 - Simplified DLL model in the Laplace domain

3.5.1.1 Voltage-controlled Delay Line

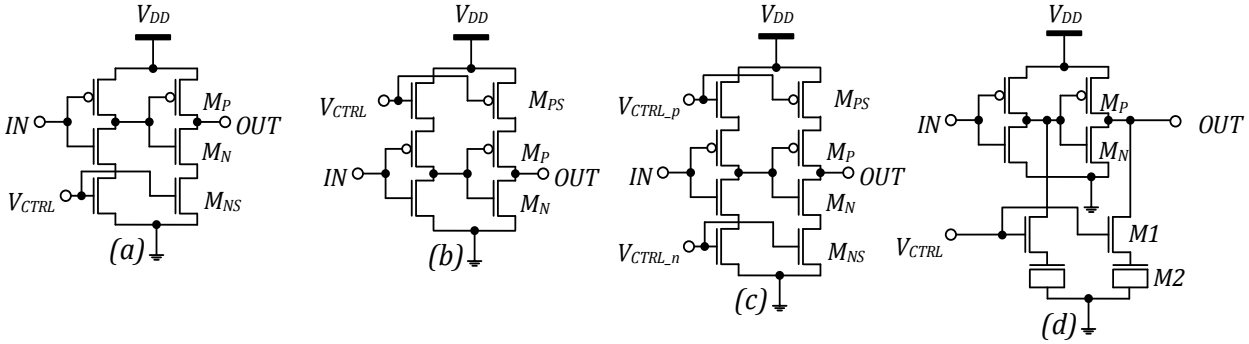


Figure 91 - Single ended voltage controlled delay elements: (a) n-voltage controlled current starved cascaded inverters, (b) p- voltage controlled current starved cascaded inverters, (c) np- voltage controlled current starved cascaded inverters, (d) Shunt-capacitor voltage controlled delay element

The Voltage-controlled Delay Line (VCDL) is the central part of the DLL; it consists of several delay elements with similar elementary delays connected in series to form a delay line with a larger delay. VCDL can be realized with fully digital elements based on invertors or buffers, such systems are easy to design and require low power consumption but they suffer from high jitter levels and are only suitable for coarse measurements. Analog VCDL on

the other hand are more suitable for precise measurements as they allow continuous fine delay regulation, they are usually divided into capacitive load delay elements, current starved delay elements, and differential delay elements. A capacitive load delay element is shown in Figure 91- d; it consists of 2 inverter stages with a variable capacitive load. In this configuration M2 acts as a capacitor and M1 is used to control the charging/discharging current via V_{ctrl} thus controlling the propagation delay. This structure suffers from a limited regulation range as well as a big occupation area due to the use of the capacitor. Another approach to construct analog VCDLs is the current starved delay elements (Figure 91-a,b,c) which consists of 2 starved inverter stages with one or more additional transistors added to control the DC operating point and as a result the current charging /discharging the load capacitance. Current starved delay elements represent a better single ended alternative as they are simple to implement and allow a wider range of delay regulation with low power consumption. However they suffer from some disadvantages in particular a low immunity to noise, as well as supply and temperature variations. A better immunity to supply and temperature fluctuations can be obtained using differential delay elements but this comes at the price of static power consumption, an increased complexity, a bigger occupation area and a longer propagation delay.

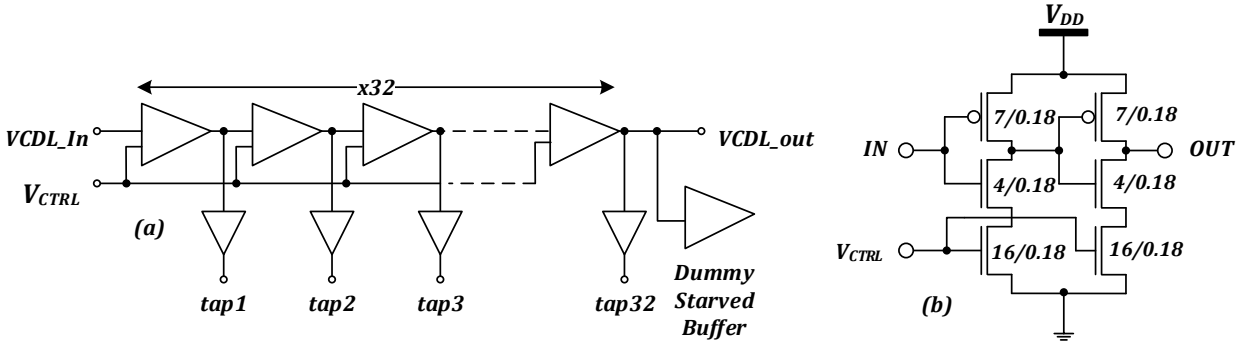


Figure 92 - The voltage Controlled delay line structure (a) and the elementary voltage controlled delay stage (b)

The voltage controlled delay line structure used in the presented DLL is shown along with the elementary voltage controlled delay stage in Figure 92. In order to ensure the fastest propagation delay possible for our design, we opted for the N-side current starved delay element which consists of a cascaded inverter pair with an additional NMOS transistor inserted in each of their pull-down path and controlled by a global control voltage to adjust their propagation delay. The fact that this structure requires low power consumption with reasonable occupation area is an additional advantage. Giving that during any transition of the input, one of the two inverters of the delay element will be discharging through a controlled transistor, and the other will be charging normally through a PMOS transistor, the total delay of the double starved inverter is therefore equal to the sum of a normal inverter delay and a controlled inverter delay resulting in a total propagation delay equal to:

$$t_p = \frac{1}{2}(t_{p-HL} + t_{p-LH}) = \frac{C_L}{2}(\frac{1}{K_p V_{DD}} + \frac{V_{DD}}{K_n {V_{CTRL}}^2}) \quad (28)$$

Where t_{p-HL} and t_{p-LH}represents respectively the propagation delays for high to low and low to high transitions, C_L is the load capacitance, and V_{CTRL} is the global control voltage [108]. If the control transistors are large enough and the control voltages are set to maximum the overall delay of the cell can be approximated to that of a simple cascaded inverter pair but in reality it is slightly higher due to the additional resistance and capacitance added by the starved transistor. The dimensions of the transistors were chosen such that the rising and the falling delays are similar; in theory this can be achieved by widening W_P compared to $W_n$with a factor K defined as:

$$K = \frac{W_p}{W_n} \approx \frac{\mu_n}{\mu_p} \quad (29)$$

to compensate the slower holes mobility. A wider PMOS improves t_{p-LH} by increasing the charging current but it also degrades t_{p-HL} due to an increased intrinsic capacitance. Consequently, symmetrical rising and falling delays are achieved at the price of an overall propagation delay greater than the minimum achievable value. Furthermore, based on [109] the optimum propagation delay in this scenario is obtained by choosing W_P and W_n such as:

$$\frac{W_p}{W_n} \approx \sqrt{\frac{\mu_n}{\mu_p}} \quad (30)$$

The results obtained via simulations were indeed consistent with this theory as the optimal propagation delay was obtained for W_p = 7 µm and W_n = 4 µm with $L_p = l_n$ = 0.18 µm. In order to ensure that the starved NMOS (M_{ns}) operated as a linear current source, we chose a larger value for its width (W_{ns}) moreover simulation results showed that the elementary delay decreases when the width of the starved NMOS transistor increases however it tends to stabilizes for $W_{ns} \geq$ 16 µm (Figure 93). Based on these data, the dimensions of the starved NMOS transistor were fixed as W_{ns} = 16 µm and l_{ns} = 0.18 µm. The variation of the rising transition and falling transition delays for an elementary stage delay loaded and driven by an identical buffers as a function of W_{ns} are shown in Figure 93. Figure 94 shows the variation of the rising transition and falling transition elementary delays as a function of the NMOS Control voltage.

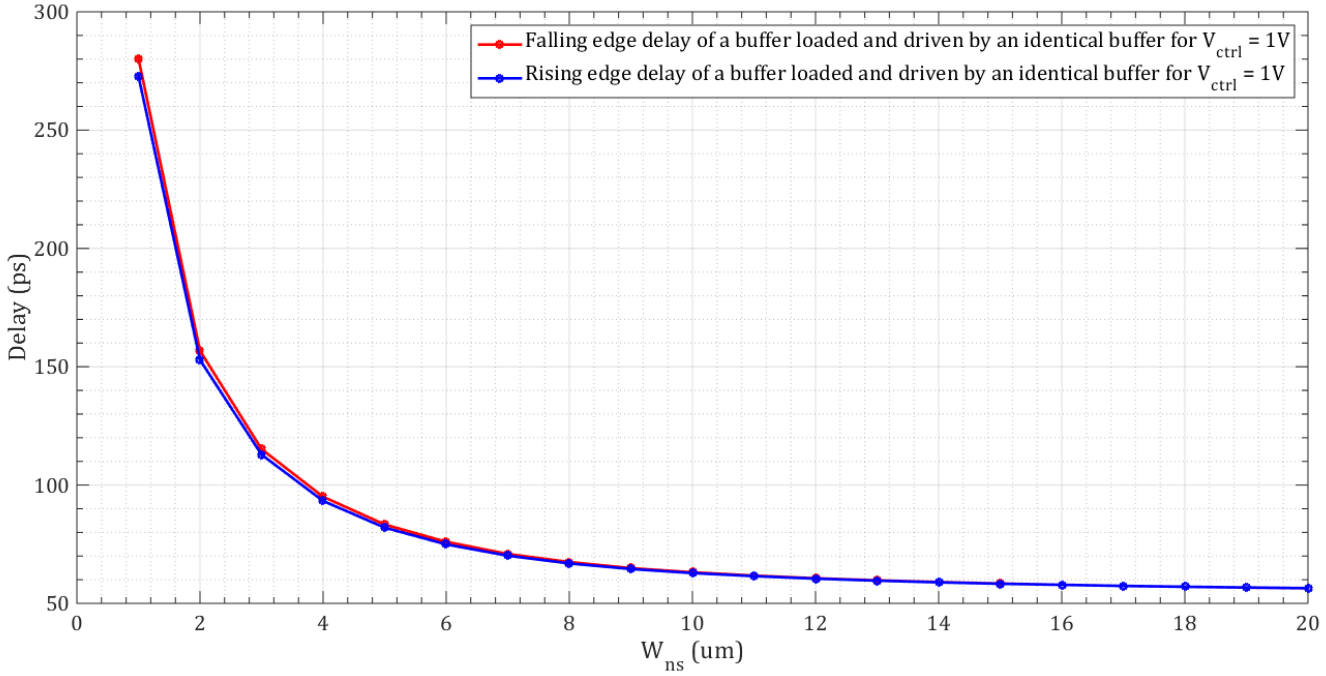


Figure 93 - Falling and rising delays of a buffer loaded and driven by an identical buffer as a function of W_{ns} for V_{ctrl} = 1 V

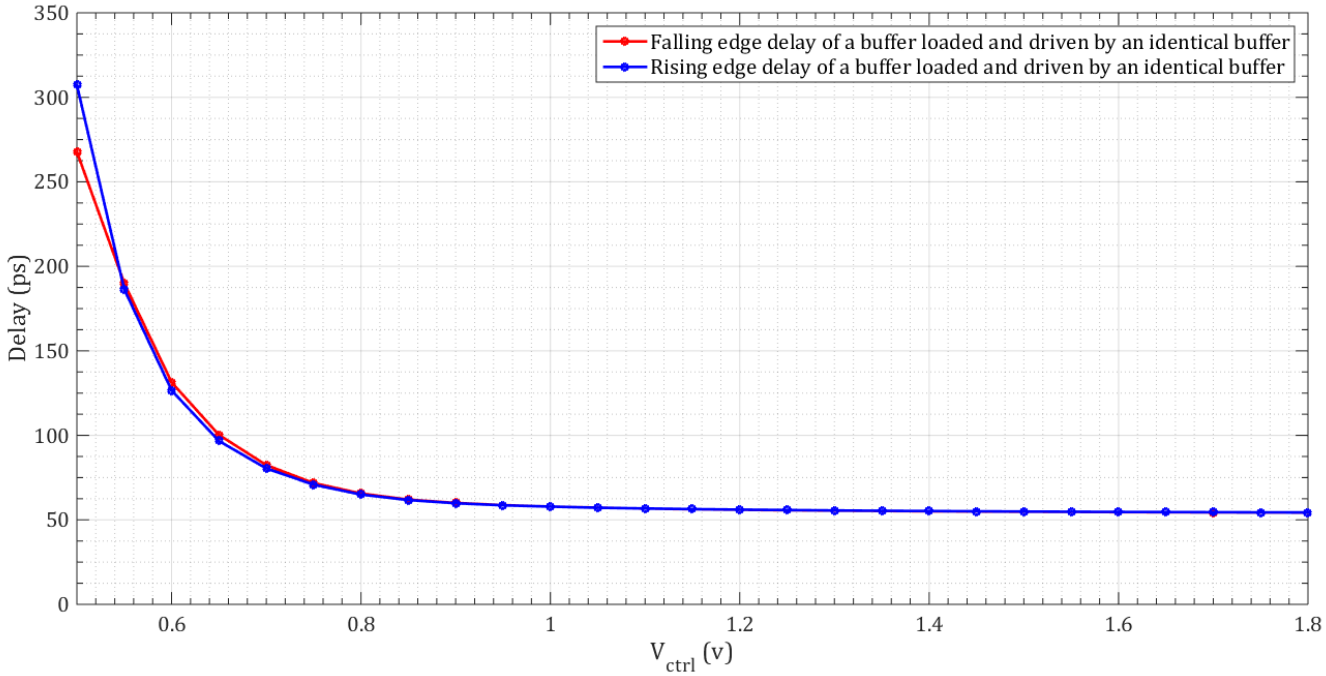


Figure 94 - Falling and rising delays of a buffer loaded and driven by an identical buffer as a function of V_{ctrl}

Deterministic Jitter evaluation:

The VCDL generates two types of jitter in the DLL: random noise jitter and deterministic jitter also known as skew. Random noise jitter is caused by many factors such as thermal noise, power supply noise, substrate noise, crosstalk and process variations, it can be minimized by using appropriate design techniques. Deterministic jitter results from device mismatch which is due to the stochastic nature of physical implementation processes. This mismatch is inherent in a real circuit and cannot be eliminated or reduced during the design process, it will cause the delay of the elementary stages in the VCDL to deviate from the ideal value for a certain tuning voltage which will result in jitter. However, although device mismatch is caused by a stochastic process, its ensuing jitter is deterministic, because once the chip has been processed; the mismatch properties are more or less permanent and the timing errors caused by this mismatch are systematic. In order to evaluate the statistics of deterministic jitter due to delay mismatch which will set the lower bound of the achievable jitter, Monte-Carlo simulations were performed. Figure 95 shows the simulation results for a stage voltage control of 715 mV which correspond to a global delay of 2.5 ns. For 1000 runs, the results show a standard deviation less than 1 ps (~372 fs) and a peak to peak variation of 2.55 ps. furthermore; the taps propagated through the standard buffers show a standard deviation of less than 1 ps (~738 fs) and a peak to peak variation of 5.1 ps.

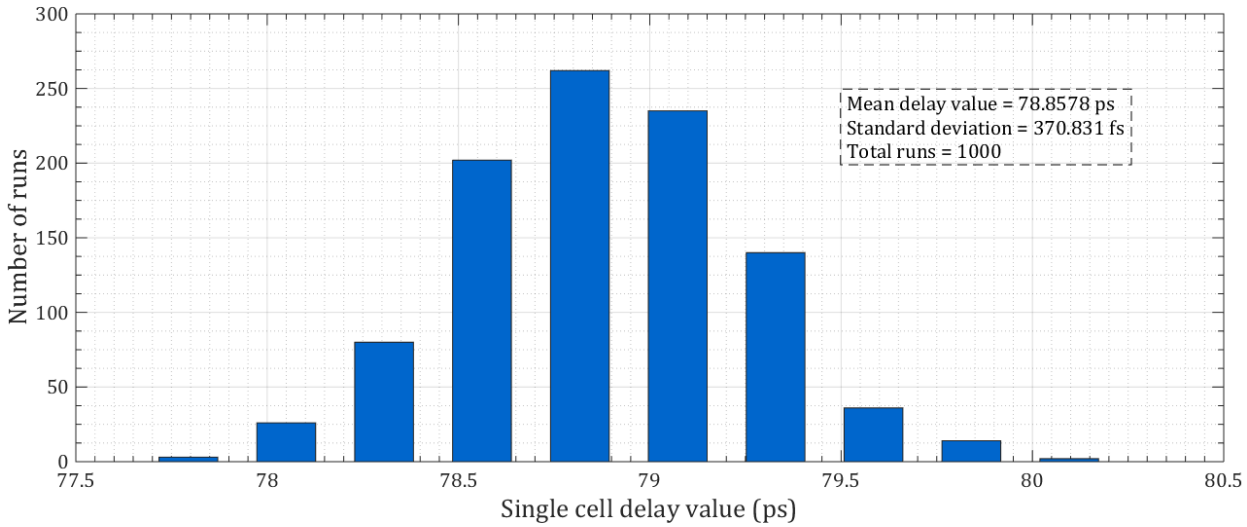


Figure 95 - Distribution of the single cell delay values for a Voltage control of 715 mV based on device mismatch Monte-Carlo simulation

3.5.1.2 Phase Detector

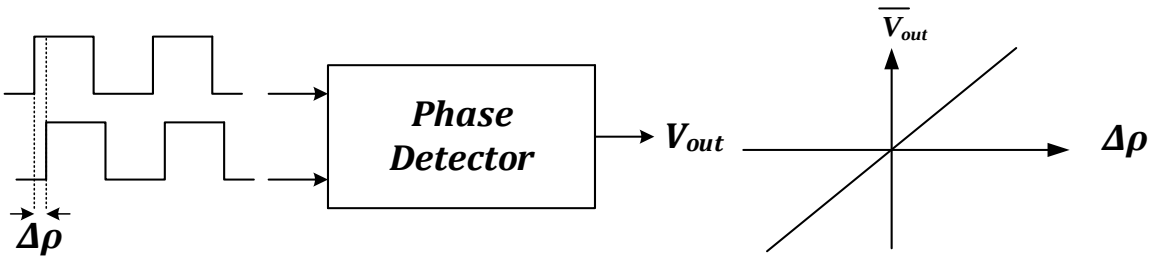


Figure 96 - characteristic of an ideal phase detector

An ideal Phase detector (PD) produces an output signal whose dc value is linearly proportional to the difference between the phases of 2 inputs:

$$\overline{V_{out}} = K_{PD}.\Delta\rho \tag{31}$$

Where $\Delta\theta$the input is phase difference and K_{PD} is the gain of the phase detector. Ideally the relation between $\overline{V_{out}}$ and $\Delta\rho$ is linear crossing the origin for $\Delta\rho = 0$ (Figure 96) and K_{PD} $(rad.\mathrm{V}/\theta)$ is defined as the slope of the line. In the DLL, the phase detector detects the phase difference between the VCDL input and its output, the output can be a signal whose width is proportional to the detected error or in some case one or more signal indicating if the VCDL's output is early or late compared to the VCDL's input. A phase detector can be as simple as an exclusive OR (XOR) gate as shown in Figure 97; the width of the output pulses varies with the phase difference between the two inputs resulting in a dc level proportional to the phase error. This phase detector operates around a static phase difference of $\pm\pi/2$ and the sign of the static phase shift is set by the sign of the gain slope. Furthermore the phase detector gain is independent of the inputs' amplitude and constant over the phase range. Unfortunately, the XOR phase detector suffers from several limitations in particular a short detection range and a duty cycle sensitivity causing it to lock with a phase error if the duty cycle is not 50%. The duty cycle sensitivity can be eliminated using a JK-Flip Flop phase detector [110] shown in Figure 98. The addition of the two flip flops to the XOR gate makes the phase detector only sensitive to the rising edges of the input signals, rather than their duty cycles. Because this phase detector uses only the leading edge of the input signals, the linear region of the phase detector is expanded to ±π (Figure 99-b). Unfortunately; it also it also leads to potential false locking to input reference harmonics of the input reference clock as they can display the same DC output (Figure 99-c). False locking to reference harmonics can be avoided by using a device that detects both frequency and phase errors.

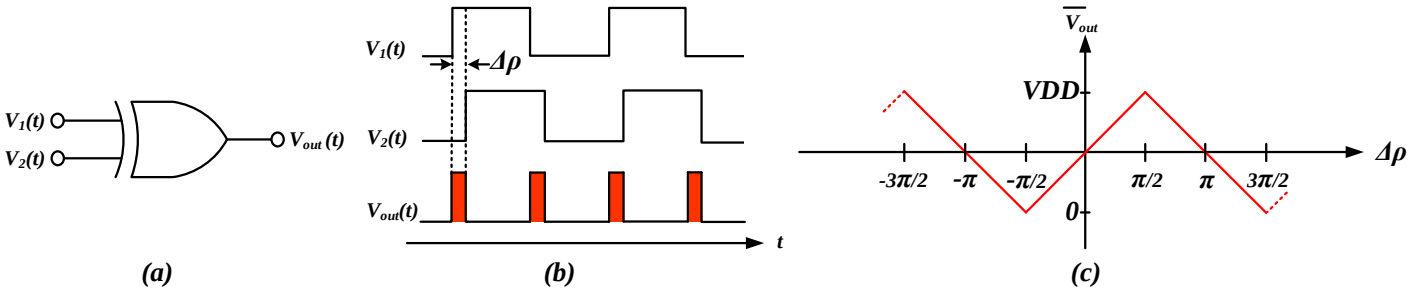


Figure 97 - (a) XOR gate operating as a phase detector, (b) Inputs and outputs waveforms and (c) Ideal transfer characteristics

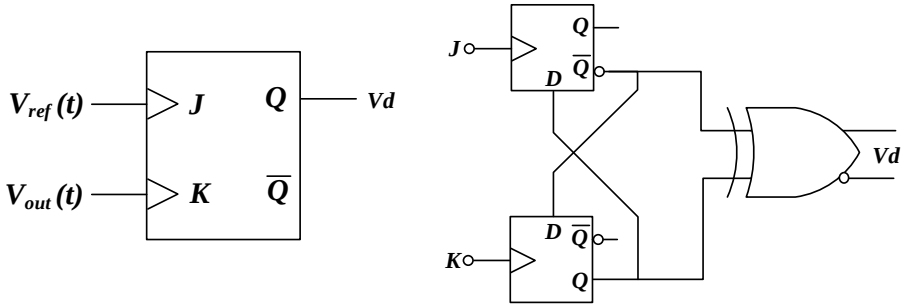


Figure 98 - JK Flip-Flop phase detector

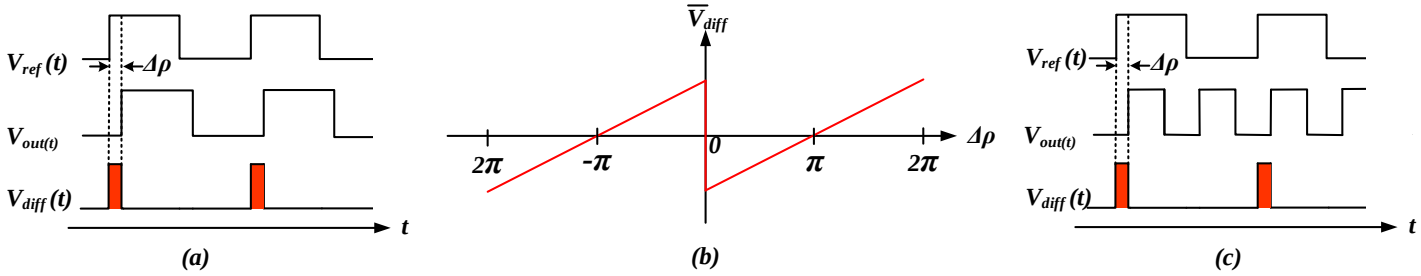


Figure 99 - (a) Inputs and outputs waveforms of a JK Flip-Flop phase detector, (b) Ideal transfer characteristics and (c) example of false locking to a harmonic of the reference clock $V_{ref}(t)$

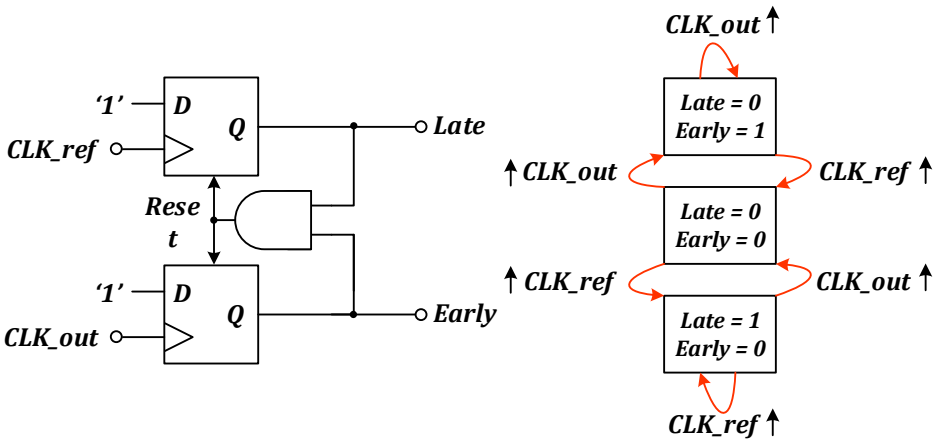


Figure 100 - Conventional phase-frequency detector (left) and state diagram (right)

Such device is called phase-frequency detector (PFD), it produces an output pulse whose width is proportional to the phase error between the reference input signal and the controlled input signal. The PFD deliver two terminals signals, early and late. It is implemented as an edge-trigger sequential. Figure 100 shows the state transition diagram summarizing the PFD operation with a conventional PFD implementation which consists of two D Flip Flops and an AND gate inserted in the reset path. D Flip Flop (FF) used here are edge triggered with their input connected to logical "one", the inputs are connected to the clock inputs. The ideal output characteristics the conventional PFD are illustrated in Figure 101-a, in addition to immunity to false locking and duty cycle issues, this approach also offers a wide acquisition of range ($\pm 2\pi$) and a faster locking speed but they suffer from dead zone and blind zone problems.

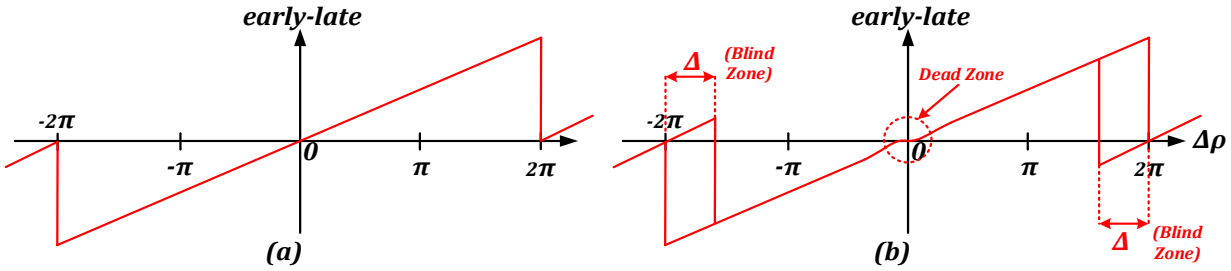


Figure 101 - Characteristics of an ideal (a) and real (b) PFD

The Dead zone problem occurs when the phase error is small, the short output pulses produced by PFD are too short and thus cannot be effectively propagated to switch charge pump. This leads to a "dead zone" which translates into low loop gain and increased jitter [111]. One way of eliminating the dead zone is by inserting a delay element in the reset path thus forcing a minimum *UP* and *DOWN* pulse length (Figure 102) at the price of a lower PFD frequency operation. The blind-zone problem arises when the phase difference of the two input clocks approaches ±2π resulting in missing clock edges and output polarity reversal. The reset phase is triggered when both the *UP* and *DOWN* signals are simultaneously high as a result the PFD is insensitive to any transition of the input signals. An example of polarity reversal due the blind zone is illustrated in Figure 101-b, the rising edge of the reference signal occurs during the reset phase. As a result this transition is not detected by the PFD which causes the leading reference signal to be incorrectly perceived as late in in the following cycle (*DOWN* is wider than *UP*). The ideal linear phase detection range of the 3-state PFD is [-2π, +2π] but in reality, the full detection range can't be achieved due to the blind-zone problem [112] which limits the PFD detection range to [-2π+Δ, +2π-Δ] where Δ is blind-zone in phase domain. The only way of avoiding this problem is by taking special care during the circuit design to ensure that the input signal rising edge does not occurs during the reset phase by limiting the reset time of the PFD [113].

One of the PFD architectures that virtually eliminates dead-zone without limiting the operation speed is a design based on dynamic CMOS logic, in which the PFD's output signals are directly used to reset the PFD without any intermediate logic [114]. A Dynamic-Logic Phase Frequency Detector (Figure 103-a), based on this architecture was reported by Li and Ismail [115]with no dead zone and near zero phase error but it is still suffered from blind-zone. In [112], two architectures were suggested to improve this issue. The Delayed-Input-Pulse Dynamic PFD (DIP-PFD) (Figure 103-b), and the Delayed-input-edge dynamic PFD (DIE-PFD) (Figure 103-c), reduce the blind zone issue by using a delayed version of the input signal when the PFD operates in the blind-zone to ensure that its rising edge comes after the reset phase thus ensuring that the rising pulse edge is properly detected by the PFD. In [112] Cheng Zhang et al. proposed a PFD architecture that virtually eliminates the blind-zone. This architecture was modified in [116] where a dead-zone free PFD with near zero blind-zone was reported. The detection range of the PFD was enhanced in by reducing the pre-charging time of the internal nodes to further limit the blind zone which comes at the price of an increased area. Dynamic PFDs have become very popular in recent years due to their faster operating speed and reduced power consumption and occupation area. However, dynamic logic requires a minimum clock rate high enough so that the output and the intermediate logic states stored on gate capacitances are used before they are lost due leaking.

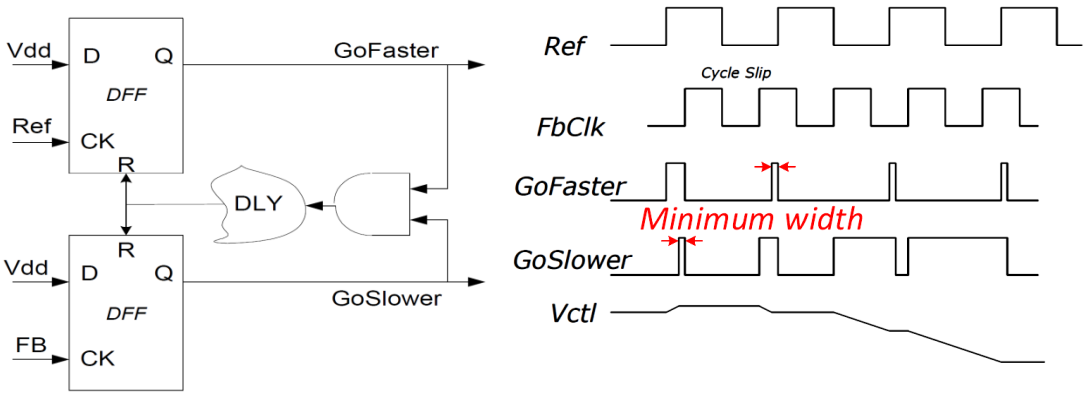


Figure 102 - Conventional PFD with an inserted delay to limit the minimum width of the PFD's outputs [117]

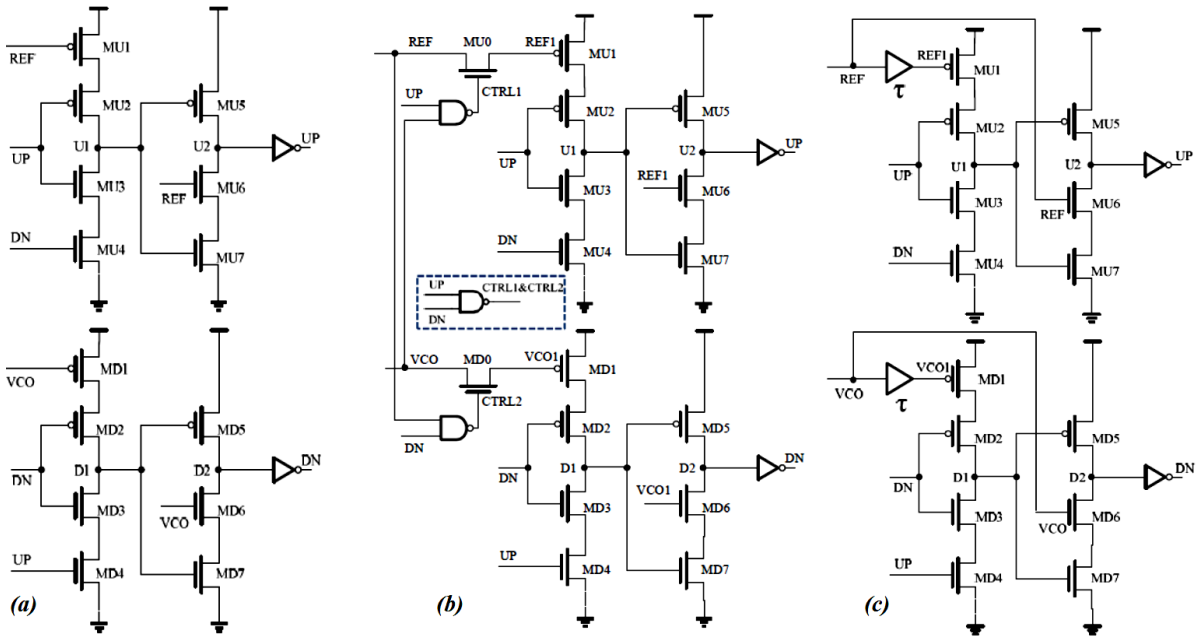


Figure 103 - Circuit diagram of Circuit diagram of Dynamic-Logic PFD by Li & Ismail (a), the DIE-PFD (b) and the DIP-PFD (c) [112]

The PFD block used in our DLL is based on the dynamic PFD device reported in [118], the architecture of the used device is illustrated in Figure 104, it features only 12 transistors and the inputs are driven by the *CLK IN* and *CLK OUT* signals thus allowing it to operate with less phase offset at high frequencies due to symmetry and shallow logic depth. The transistor aspect ratios are resumed in Table 5; non-minimum sized transistors were used in order to obtain higher parasitic capacitances on the intermediate switching nodes thus ensuring that the internal and output nodes of dynamic PFD remain at the correct logic state.

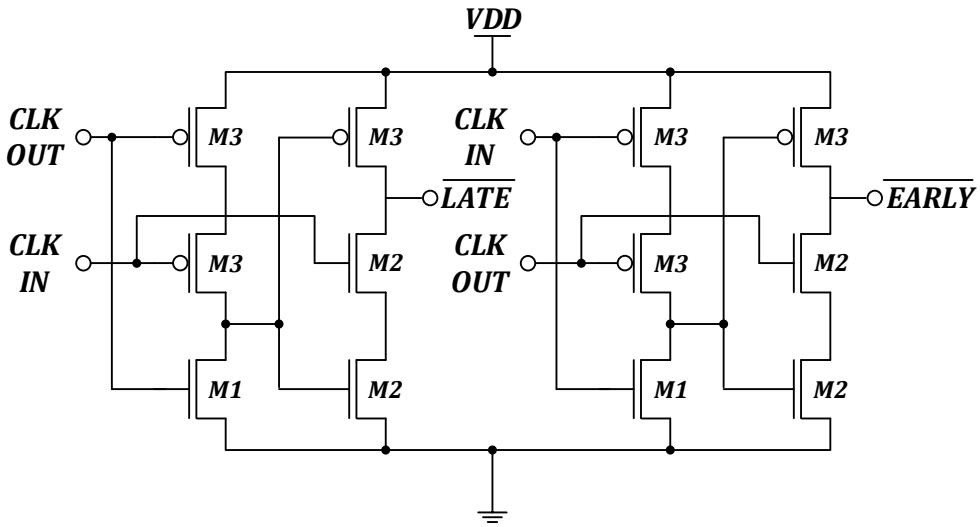


Figure 104 - dynamic logic open-loop PFD proposed in [118]

Table 5 - Phase detector transistor aspect ratios.

Transistor	Aspect Ratio
M1	2/0.18
M2	7/0.18
M3	2/0.18

3.5.1.3 Charge Pump and Loop Capacitor

In a DLL, the PFD detects the phase error between the VCDL input reference clock and its output and transfers information to the charge pump in the form of two voltage pulses. Based on these pulses the charge pump generates a charging or a discharging current and these currents are integrated through the loop filter in order to obtain a control voltage that adjusts the delay of the VCDL cells until the DLL reaches the lock state. Basically, the charge pump simply consists of two controlled switches, a current source, and a current sink, as shown in Figure 105. The two switches are controlled by the EARLY pulses and the LATE pulses, respectively. When the switch controlled by the LATE pulse is closed, the charging current is pumped into the loop capacitor and when the switch controlled by the EARLY pulse is closed the discharging current removes charges from the Loop capacitor into the ground, this charging and discharging will continue until the DLL is in the lock state, Once the lock is achieved, the voltage of the loop filter is kept constant by maintaining equal charging and discharging in the loop. In theory, the sinking and the sourcing currents must be equal so that the DLL will lock properly but in reality many factors contribute to inevitable current mismatches resulting in static phase offset as well as dynamic jitter in the DLL. Current mismatch is the result of the inherent current sources dissimilarities [119] as well as the pumping currents deviating from each other as the output voltage varies between VSS and VDD, due to the channel length modulation effect [120]. Figure 106 shows three typical charge pump topologies, the charge pump with switch in drain (Figure 106 – a) is the basic charge pump implementation, however the charge pump with switch in gate (Figure 106 – b) ensures the current mirrors are operating in the saturation region while the charge pump with switch in source (Figure 106 – c) offers a faster switching time than the gate switching without the need for high bias current. Improving the basic charge pump implementation is achieved by enlarging the output impedance of the current sources using long channel MOSFET and cascode structures such that the current variation is less sensitive to the output voltage.

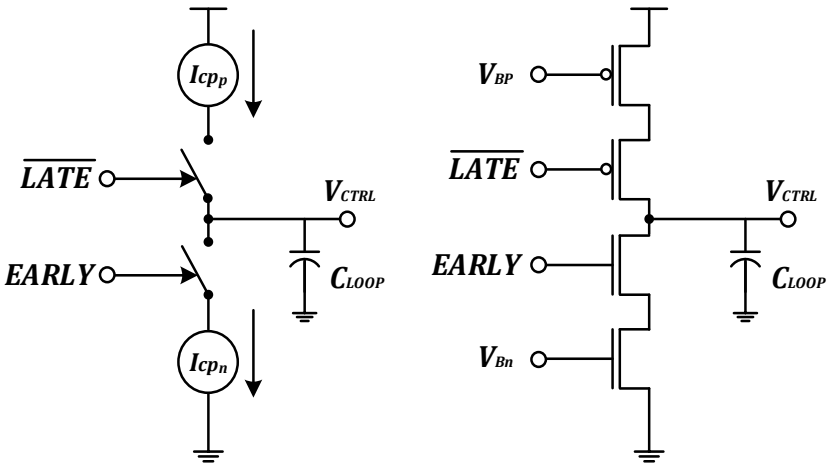


Figure 105 - Ideal representation of a charge pump (left) and basic implementation of a charge pump circuit

To avoid the inherent mismatch between PMOS and NMOS a charge pump topology using only NMOS switches was proposed in [121] but good performance required the use of a sufficiently large current [122]. A more conventional solution to the charge pump current mismatch problem rely on the use of analog feedback to match the charging and discharging currents [123] [120] , various employs operational amplifiers in a feedback loop to equate the NMOS and PMOS source currents via direct gate [124]or drain voltage adjustment using cascade transistors [125]. A low power rail to rail current conveyor is used for current matching in [126] but in general, these amplifier-based techniques result in bigger area and higher power consumption making them more costly and unsuitable for low supply voltage applications. In addition to current matching, special attention must be paid to limit dynamic errors caused by source switching and the presence of parasitic capacitance. These errors can be alleviated by improving the switching speed thus reducing parasitic effects such as charge coupling, glitches and noise contribution. The interference caused by Clock feedthrough and charge injection is proportional to the value WLC_{ox}, therefore a smaller sized MOS switch is recommended. Another source of dynamic errors is charge injection which can be handled by moving the switchers' positions to the source of the current source so that they do not connect to the output. This will also lead to speed improvement without the need to increase the biased current resulting in reduced power consumption.

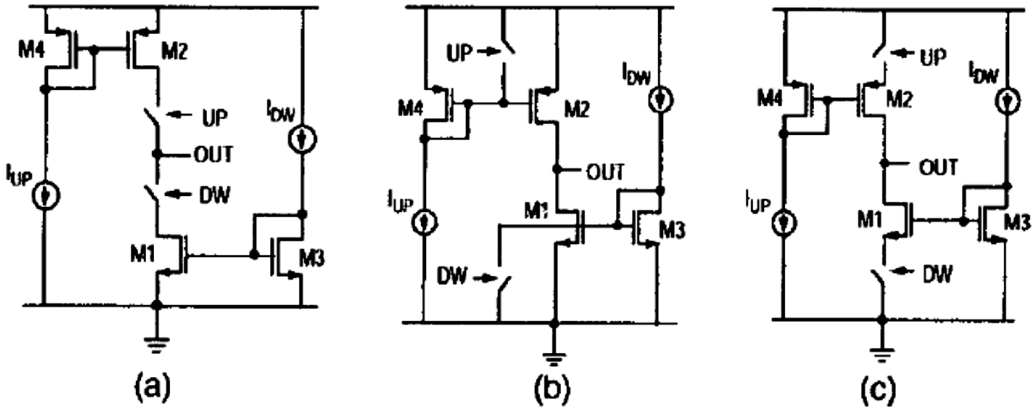


Figure 106 - Typical Charge pump topologies: Single-ended charge pumps: (a) switch in drain, (b) switch in gate, and (c) switch in source [122]

Single Ended Vs Differential Charge Pump Architectures

Single-ended charge pump circuits are an elegant approach that allows low power consumption, and smaller occupation area. The output current of the charge pump can be as high as 4.5mA [125] at lock in order to provide better spur performance by limiting the leakage current and to have high SNR. The use of tristate operation scheme limits the current consumption of the charge pump to a few hundred μA depending on the reference clock frequency and the delay of the PFD. Some typical single-ended charge pump topologies are shown in Figure 107. Figure 107(a) shows a charge pump utilizing a current steering switch. This structure provides high speed switching for a single-ended charge pump, since the switching time is improved by the current steering properties of the associated switching pair (M1-M3 and M2-M4). Another charge pump approach utilizing current steering with an active amplifier [122] [127] is shown in Figure 107 (b). This unity gain amplifier, buffers the voltage at the output node forcing the drain voltage of the current sources IDN and IUP to be the same when M1 and M2 are on or when they are off. This reduces the charge sharing effect, when the switch is turned on. This architecture ensures fast transient response through current steering, reduces the effect of any parasitic capacitance, at the expense of extra current. Finally, in Figure 107(c) the inherent mismatch of PMOS and NMOS transistors is avoided by using only NMOS switches [128]. Since the current does not flow in the current mirror, (M5 and M6), when the UP switch is turned off, the current mirrors still limit the performance unless large current is used [122].

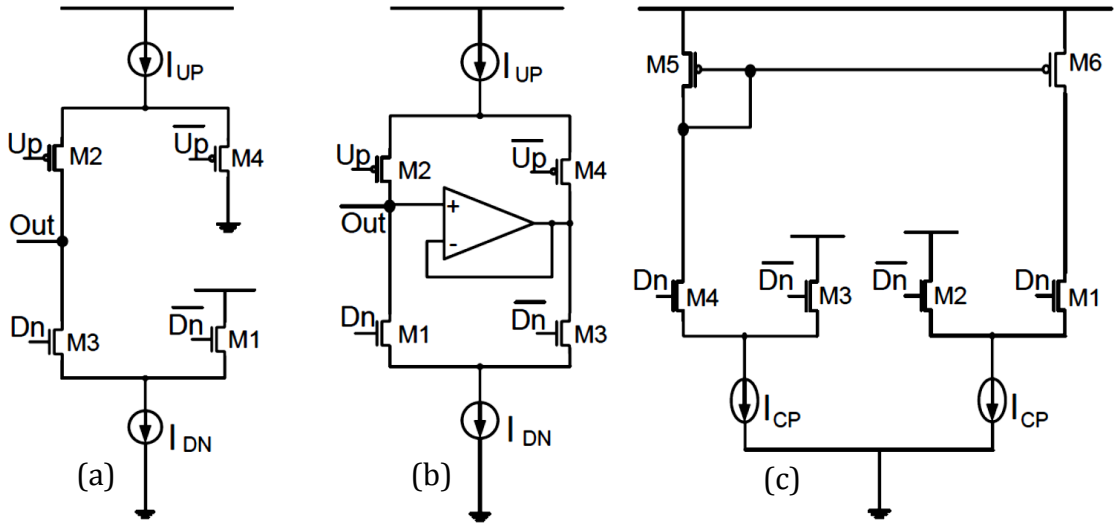


Figure 107 - Single-ended charge pump architectures: a) with current steering switch, b) with unity gain active amplifier and c) with NMOS switch only [125]

A fully differential charge pump (Figure 108) has several advantages over the conventional single ended charge pump. Firstly, the matching requirement between the two types of transistors is lessened as the switch mismatches between NMOS and PMOS transistors do not significantly affect the overall performance. Secondly, the differential charge pump has

only NMOS switching transistors thus the inverter delays for the Up (Early) and Down (Late) signals are fully balanced and consequently do not create any offset. Thirdly, the range of the output voltage compliance is doubled with this configuration compared to the single-ended architecture. Fourthly, the differential output stage is less sensitive to the leakage current, since the leakage current results in a common-mode offset at the dual output stages. Finally, the use of two on-chip loop filters fully differential charge pump provides better immunity to the supply, ground and substrate noise, with faster switching speeds and lower transient oscillations [125]. However, these advantages come at the price of higher power consumption and bigger area due to the use of two loop filters and common-mode feedback circuitry [122]. To alleviate dynamic errors induced by the parasitic capacitors of the switches and current source transistors, the charge-pump circuit proposed in [129] based on the structure in [130] is used here. The charge pump's architecture is shown in Figure 109. In this structure the switches M2 and M5 were putted up and down far from the output of the charge pump so that the switching charge injection will not influence the control voltage. Moreover, two clamping transistors were added to limit charge sharing, as suggested in [131]. The transistors M3 and M6 were chosen with long channels to reduce the channel length modulation coefficient [130], improve the matching characteristics of the devices and enhance the output impedance. The transistor aspect ratios are given in Table 6. The on-chip loop capacitor has been carried out with a MOS capacitor; its value is of approximately 100 pF and the charge pump can be connected to an outside capacitor in order to increase the loop capacitor value. The bias current of the charge pump (I_{CP}) has been supplied externally, the simulated sourcing and sinking current characteristics as a function of V_{CTRL} for I_{CP}=50 µA are shown in Figure 110 and a good current matching is observed in the practical operation range [0.3–1.3] V of the charge pump.

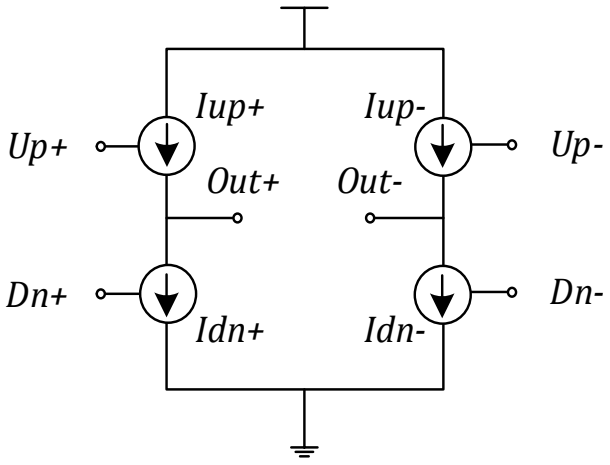


Figure 108 - Conceptual diagram of a differential charge pump.

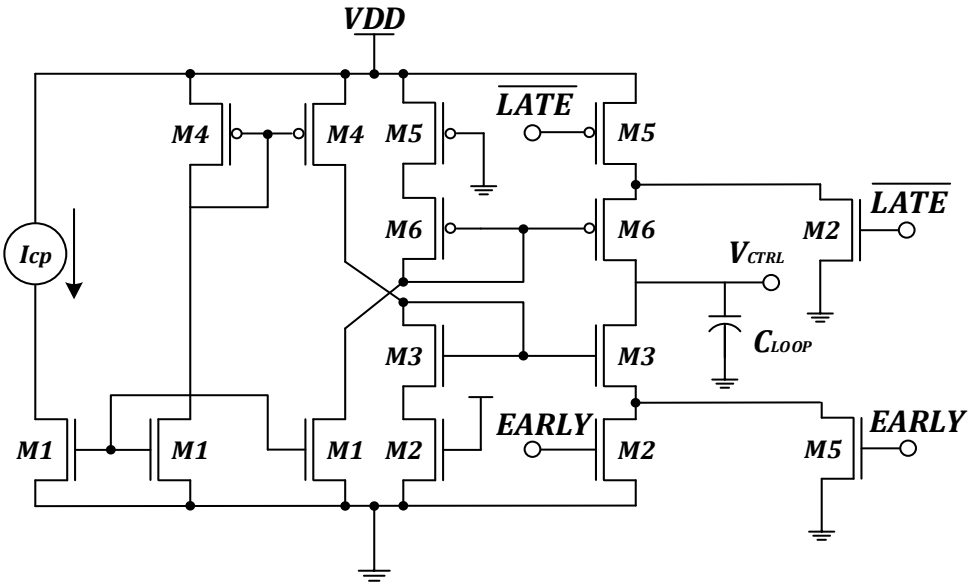


Figure 109 - The Enhanced charge pump circuit proposed in [129] based on [132]

Table 6 - Charge pump transistor aspect ratios

Transistor	Aspect Ratio
M1	7.5/2
M2	1.5/0.3
M3	10/2
M4	9/2
M5	4/0.3
M6	13/2

3.5.1.4 DLL Simulation Results

The DLL has been simulated using Cadence simulator. Figure 111 shows the initial phase difference of about 65 ps between the 2.5 ns-period input and output clocks. The charge pump bias current is set to 50 μA and the loop capacitor is 100 pF. The bottom figure shows a zoom on the clocks in locked state, 25 μs after the beginning of the lock acquisition. The static phase error is ~3 ps with CLK OUT leading CLK IN, it is mainly caused by the mismatch in the sourcing and sinking currents of the charge pump and the turn-on time of the phase detector [MZ.Kao08]. The evolution of the VCDL control voltages is shown in Figure 112, their final values are 0.7198 V which is conform to the results of the simulation in Figure 94.

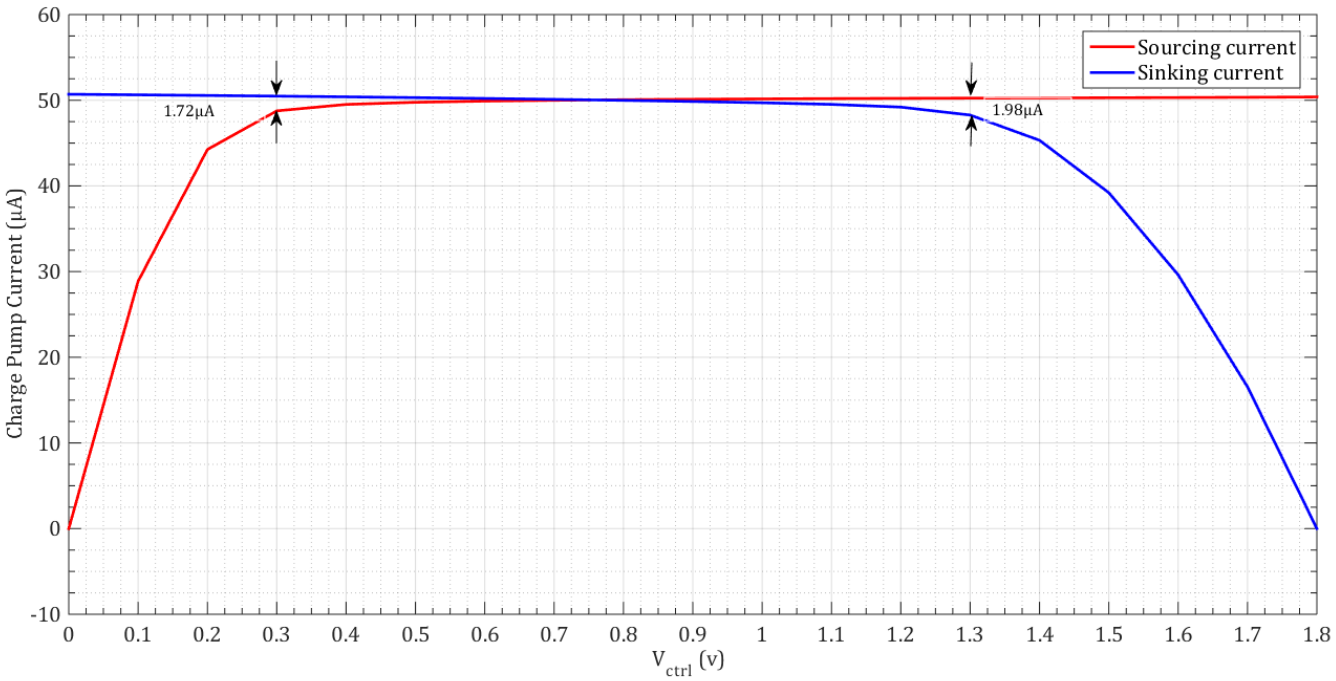


Figure 110 - Sourcing and sinking current characteristics as a function of VCTRL for ICP = 50 μA.

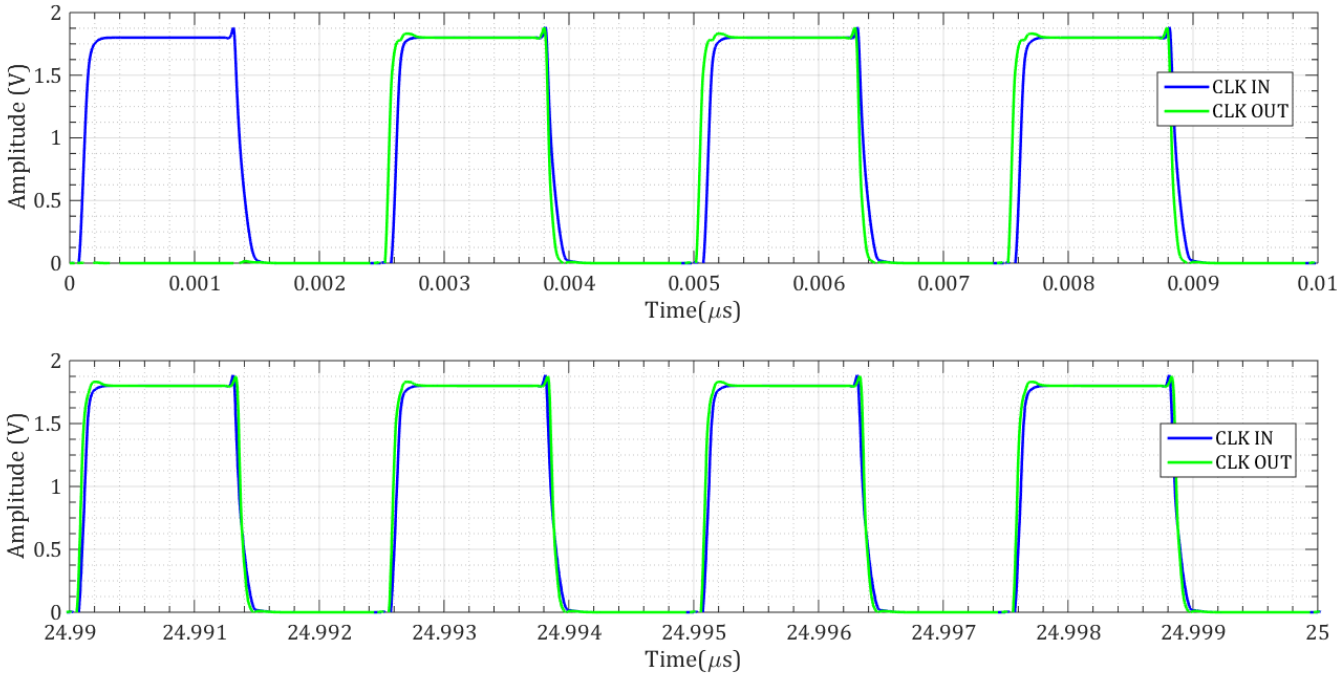


Figure 111 - Input and output clocks at the beginning of the simulation (top) and at the end of the locking procedure (bottom). T_{CLK} equals 2.5 ns and the initial phase difference is 65 ps. The static phase error when in lock is of -3 ps.

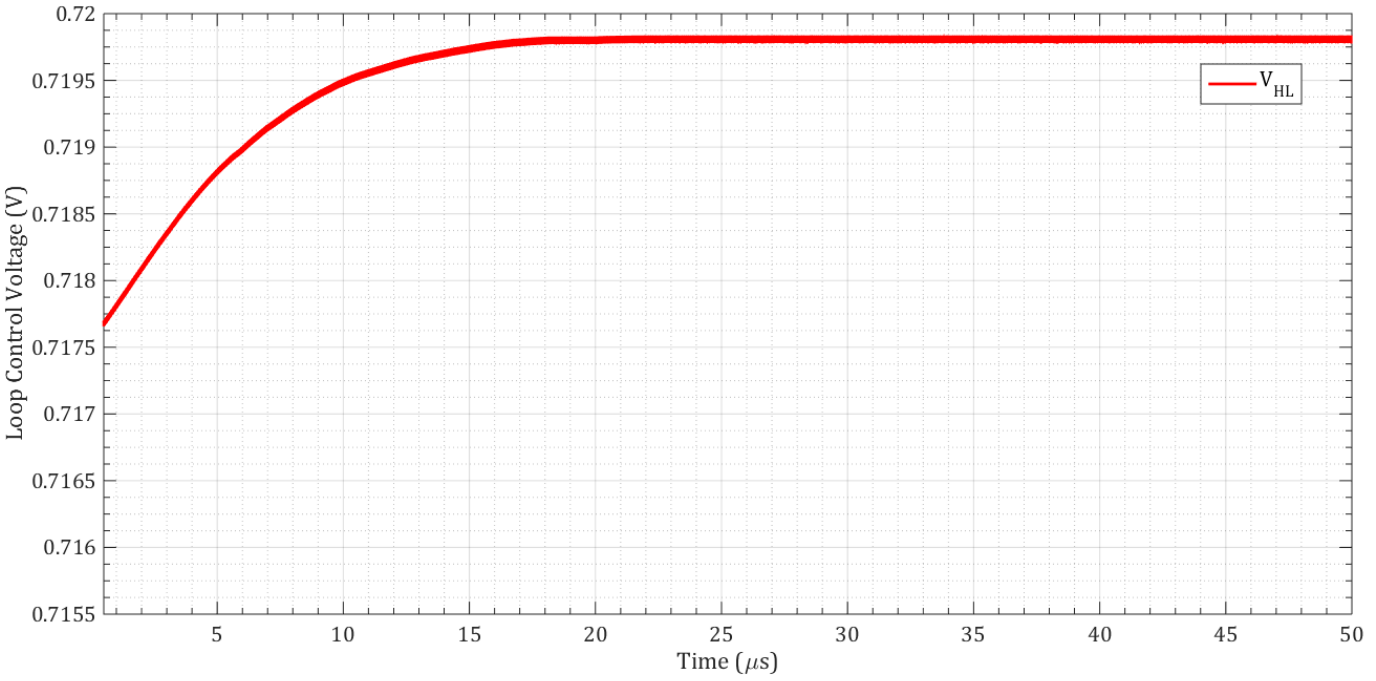


Figure 112 - Transient evolution of the control voltages during the locking of the DLL. TCLK equals 20 ns and the initial phase difference is 65ps. The final value for V_{HL} is 0.7198 V.

3.5.2 The Time Sampling Unit

The Time Sampling Unit (TSU) which is the central block of the hybrid TDC, it was designed to perform several tasks triggered by the arrival of a HIT signal indicating a detection event. These tasks include: 1) Acquiring the coarse time measurement results from the coarse counter. 2) Sampling the VCDL taps state at the detection moment, determining the VCDL stage where a '1' to '0' transition took place in order to acquire the fine time measurement. 3) Performing the ultra-fine time conversion based on the analog level in the '0' to '1' transition cell in order to obtain the ultra-fine time measurement. And 4) Transmitting the 20 bits total time conversion data where 12 bits represent the coarse time measurement data, 5 bits represent the fine time measurement data and 3 bits represent the ultra-fine time measurement data.

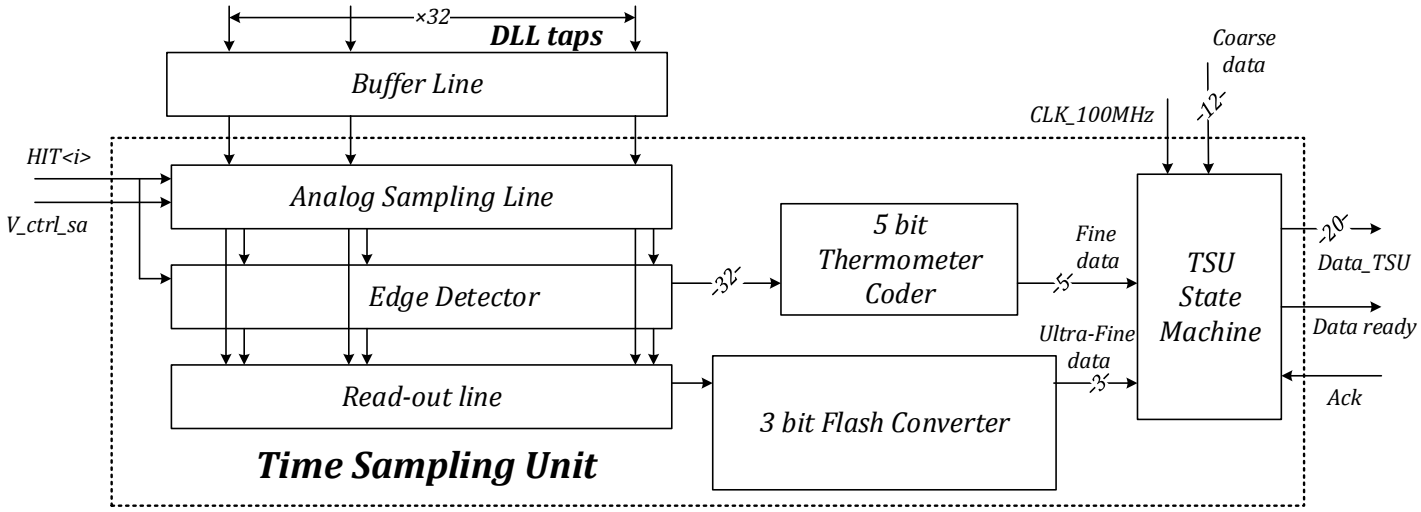


Figure 113 - Block diagram of the time sampling unit

The bloc diagram of the TSU is represented in Figure 113, it includes 4 main parts:

- The Analog Sampling Line (ASL) to perform the sampling and hold task
- The Edge Detector to detect the '1' to '0' transition position in the ASL followed by a 5 bit thermometer coder to generate the fine time measurement data with a fine resolution
- The Read-out line followed by a 3 bit flash converter to perform the ultra-fine time conversion with an ultra-fine time resolution T_{uf}
- The TSU state machine which collects all the measurements data and manages its communication to the following circuitry.

The main parts of the TSU will be presented and discussed in this section. The design process of the TSU aimed in particular to achieve a low dead time to ensure a high repetition rate, reasonable power consumption and a high scalability with a limited occupation area in order to enable the replication of the design in a multi-line array structure as we will demonstrate later. The simulations illustrated in this chapter were done with a 2.5ns input clock period resulting in $T_f \approx 80$ ps $T_{uf} \approx 10$ ps. This choice was made to ensure that the system is capable

of achieving a 10 ps time resolution however the system was designed to allow fine tuning through external signals allowing the scheme to be used for several input clock frequencies.

3.5.2.1 The Analog Sampling Line

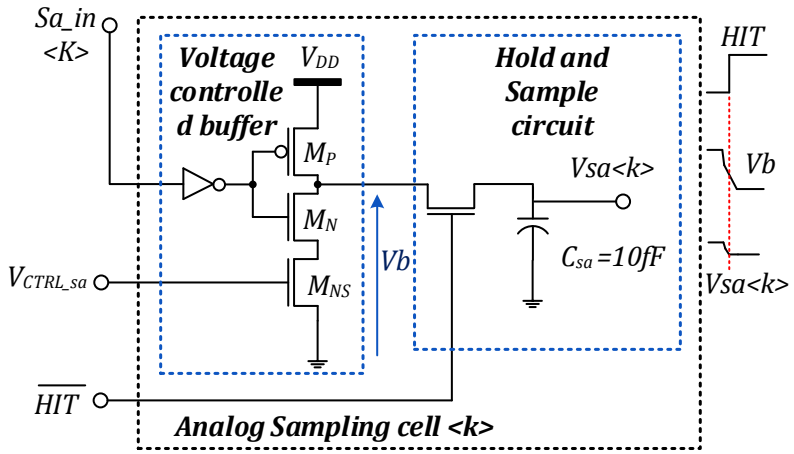


Figure 114 - Schematic of the analog sampling cell

The Analog sampling line enclosing 32 sampling cells was designed based on three major constraints: 1) The sampling cell should be compact enough to fit within the VCDL's delay cell-imposed width such that the Sampling line cells are exactly parallel to the VCDL's delay cells 2) the sampling cells should be fast enough to follow the variations of the maximum intended input signal 3) The low transition of the sampling cell output should be linearly stretched over a duration equal to the fine resolution T_f to allow a precise 8 sub-level ultrafine time conversion. Each sampling cell (Figure 114) incorporates an internal buffer made of an inverter connected to an NMOS current starved inverter with a shared external control signal V_{CTRL_sa}. The current starved internal buffer serves a dual purpose: 1) it makes the output signal independent of the fluctuation in the load capacitance value due to the change in the NMOS operating mode and 2) it allows us to adjust the high to low transition of the internal buffer's output via V_{CTRL_sa} (Figure 116) in order to fine-tune the time resolution of the TDC. The output of the current starved buffer is connected to a sample and hold circuit (S/H circuit) used to sample the VCDL's intermediate taps when the HIT signal goes from '0' to '1'. An S/H circuit can be seen as an analog memory cell that takes samples of the input signal and holds them for a fixed duration of time. These circuits operate in 2 consecutive phases: the sample phase where the output of the S/H circuit follows the input's variation and the hold phase where the input and the output are disconnected and the output's state remains stable for a fixed period of time. S/H circuits can be divided into open-loop circuits and closed-loop circuits. Open-loop S/H circuits are much more compact, they require less control signals, and provide high-speed operation. Closed-loop S/H circuits on the other hand employs operational amplifiers based techniques to remove most of the non-idealities associated with open-loop S/H circuits at the price of a higher occupation area and

a reduced operation speed [133]. Giving that the ultrafine time conversion scheme requires the sampling of an 80 ps fall time and assuming a first order behavior of the S/H circuit, the bandwidth of the sampling cell BW_S has to reach at least 4.5 GHz. It is clear that such high bandwidth can only be attained using an open-loop sampling scheme which is also the best choice giving the compact area and speed constraints.

The most straightforward way to sample a voltage in CMOS technology is to use a MOST switch and a Hold capacitor, such configuration is known as an open-loop parallel sampling since the input signal is parallel with the hold capacitor (Figure 115). A MOST switch can be built by an NMOS transistor, PMOS transistor or a complementary MOS switch or transmission gate (TGATE). A single MOST switch cannot conduct over the entire supply voltage range since at least a Vth is required across the Gate Source terminals to keep the MOST on, the use of a TGATE on the other hand allows the NMOS to compensate for the PMOS and vice versa, this solution may seem more attractive but it is not an appropriate choice for our application first because it will require the use of two complementary control signals and second because we are only interested in the signal transition between VDD/2 (0.9 V) and 0 V which is why an NMOS switch is more appropriate as it allows a better transmission of the targeted signal interval (Figure 117). Luckily the NMOS switch is also more compact and faster than the PMOS switch and is therefore also suitable for the area and speed constraints. Accordingly, an open-loop parallel S/H circuit with an NMOS switch was retained for the analog sampling cell (Figure 114).

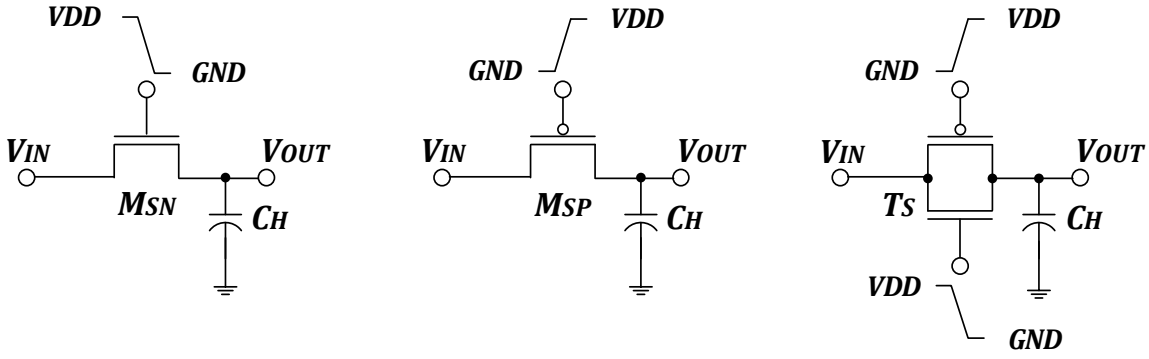


Figure 115 - Open-loop parallel S/H circuits.

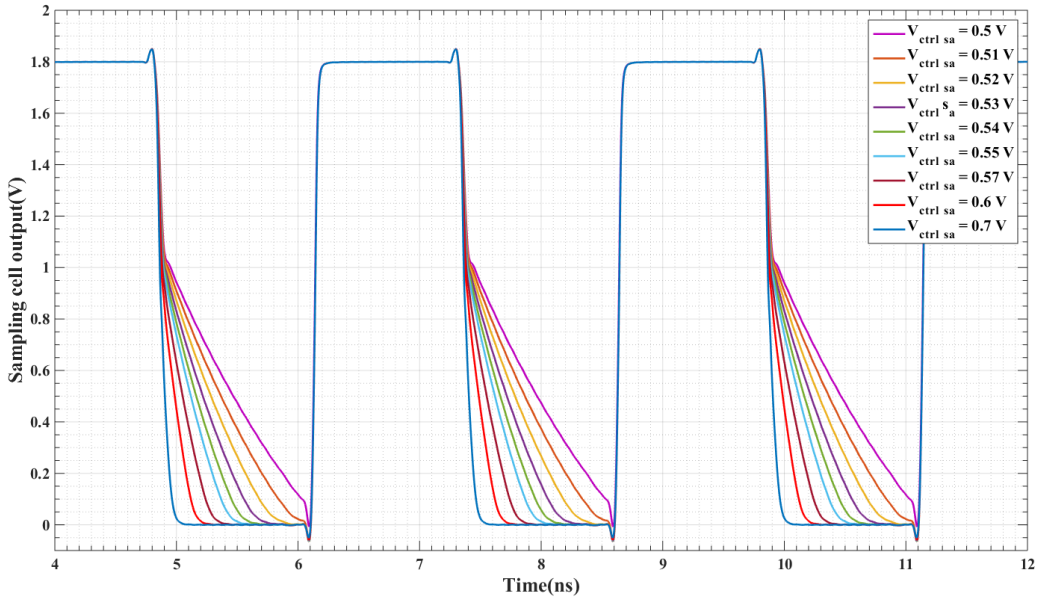


Figure 116 - Simulation results of the analog sampling cell starved buffer's output for several V_{CTRL_sa} input values

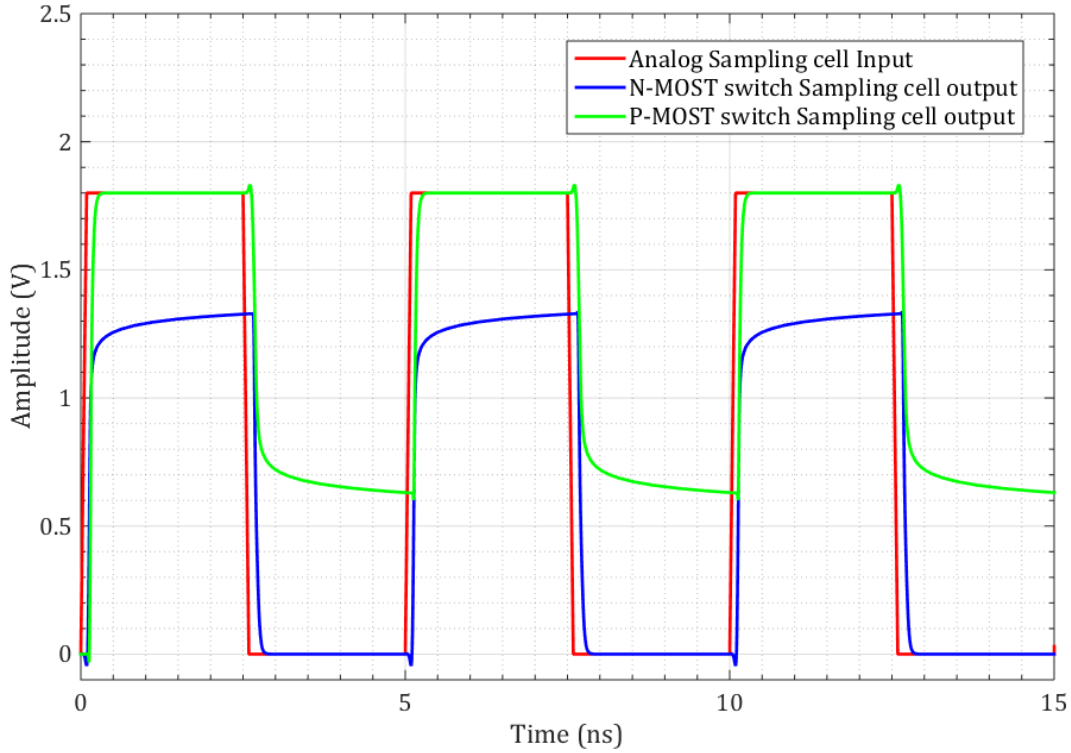


Figure 117 - Simulation results showing the input of the analog sampling cell and the associated outputs for the 2 type of MOST switch

For a given gate voltage, the values of W and L set the on-resistance RON of the MOST switch. Consequently, the circuits shown in Figure 115 can be seen as basic RC circuits. The channel resistance of a MOSFET is a source of thermal noise with a voltage power spectral density:

$$V_{n,R_{on}}^2 = 4KTR_{on} \quad (32)$$

The noise at the output of the S/H circuits of Figure 115 is given by the square root of the noise power within the equivalent noise bandwidth, defined by R_{on} and C_H:

$$V_{rms\,(n,R_{on})} = \sqrt{KTR_{on}f_c.{}^{\pi}/_{2}} = \sqrt{{}^{KT}/_{C_H}} \quad (33)$$

Where *fc* is the -3 dB bandwidth of the circuits in Figure 115, given by:

$$f_c = \frac{1}{2\pi R_{on}C_H} \quad (34)$$

Based on these equations, the value of the sampling capacitor determines the noise performance of the circuit. The subsequent slope quantification will result in a time resolution $\Delta t = 10\ ps$ leading to a time domain quantization error:

$$\varepsilon = \frac{\Delta t}{\sqrt{12}} = 3\ ps \quad (35)$$

By adjusting our system such that an 80 ps time interval is equivalent to a low transition Voltage step of 500 mV, the resultant conversion gain will be equal to 160 fs/mV and the 3 ps temporal noise will result in an rms noise of 19 mV which means that the hold capacitor value must be chosen such that $\sqrt{{}^{KT}/_{C_H}} \ll 19\ mV$ i.e. $C_H \gg 0.01\ fF$. Based on this argument, we chose a hold capacitor value of 10 *fF* to make sure that the noise performance of the S/H circuit will not create a sensible distortion compared to the quantification error.

The MOST switch resistance must be chosen in a way that the time constant $R_{ON}{\cdot}C_H$ is coherent with the highest frequencies contained in the input signal, the simulations showed that choosing a NMOS width $W_n = 5\ \mu m$ with a minimum length $L_n = 0.18\ \mu m$and is definitely sufficient to ensure an Ron small enough to meet the bandwidth requirement.

Figure 118 shows the simulation results of the Analog sampling cell output for 8 consecutive HIT signal arrivals separated by 10 ps. The analog sampling line in the TSU follows the signal of the VCDL taps and each analog sampling cell<k> fetches the input signal $Sa\_in < k >$ and stores it across the hold capacitor C_{sa}. When the *HIT* signal commanding the STU rises to '1', the NMOS in each analog sampling cell becomes off and the output $V_{sa} < k >$ is frozen until

the HIT signal goes to '0' and the NMOS switch is closed again (Figure 118). The low transition time of the internal starved buffer (Figure 116) and as a result the low transition time of the sampling cell output is controlled via the V_{CTRL_sa} signal (Figure 120). This scheme was devised to ensure that the TSU can be used with various input frequencies although the design of the subsequent TSU blocks were based on the supposition that, regardless of the V_{CTRL_sa} value, the falling edge transition is linear between 0.8 V and 0.3V as observed in the simulation results (Figure 120).

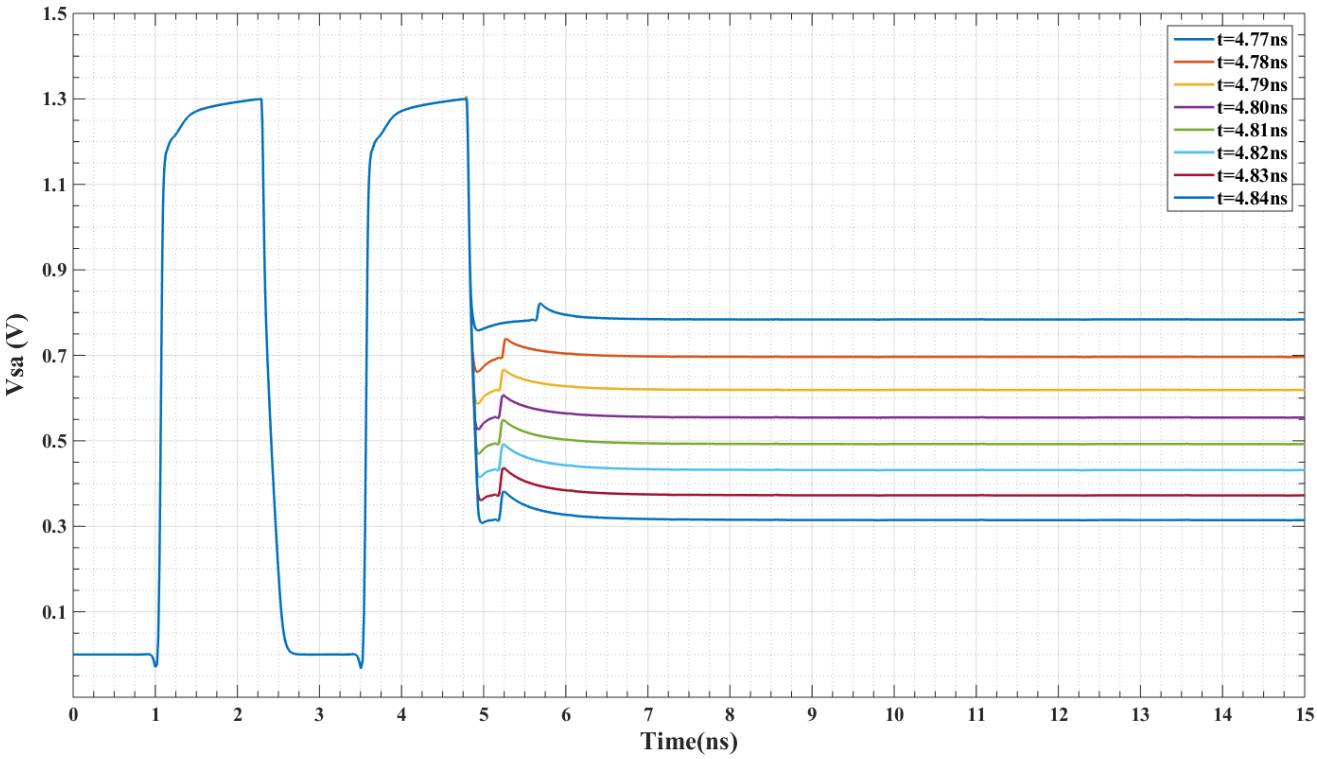


Figure 118 - Simulation results of the Analog Sampling cell's output sa<k>(V) for V_{CTRL_sa}=0.62V and f_{CLK}=400 MHz

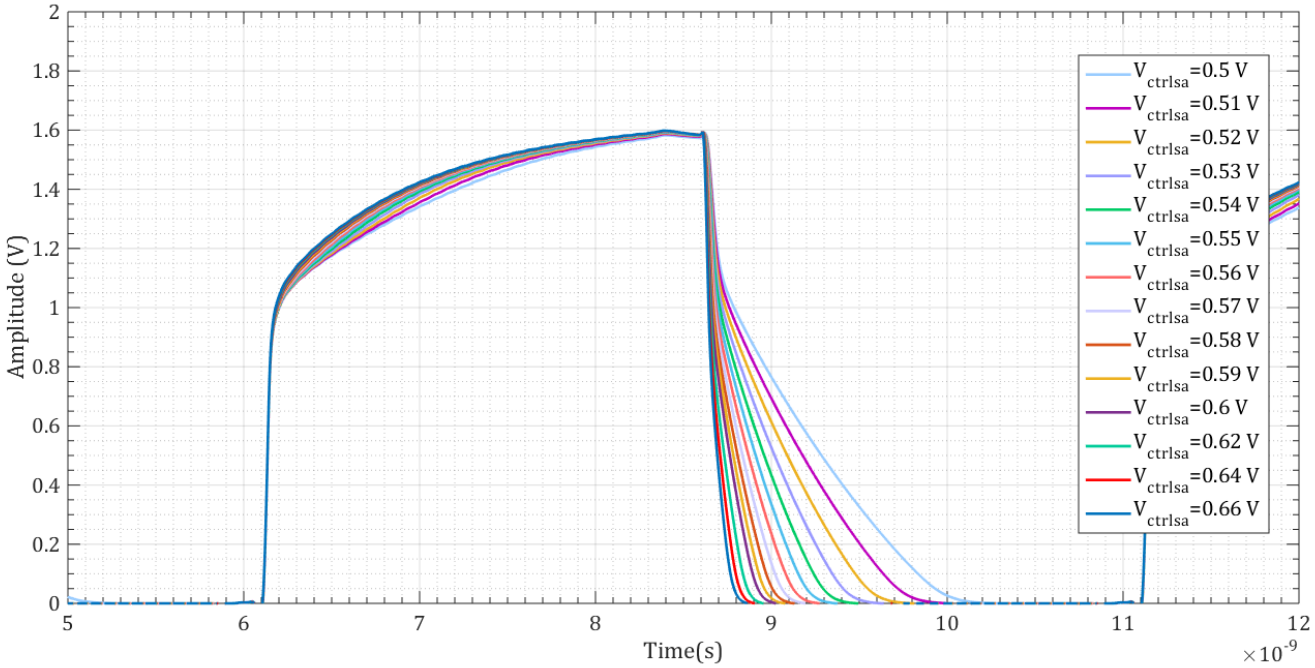


Figure 119 - Simulation results of the analog sampling cell output ($V_{sa} < k >$) for several V_{CTRL_sa} input values

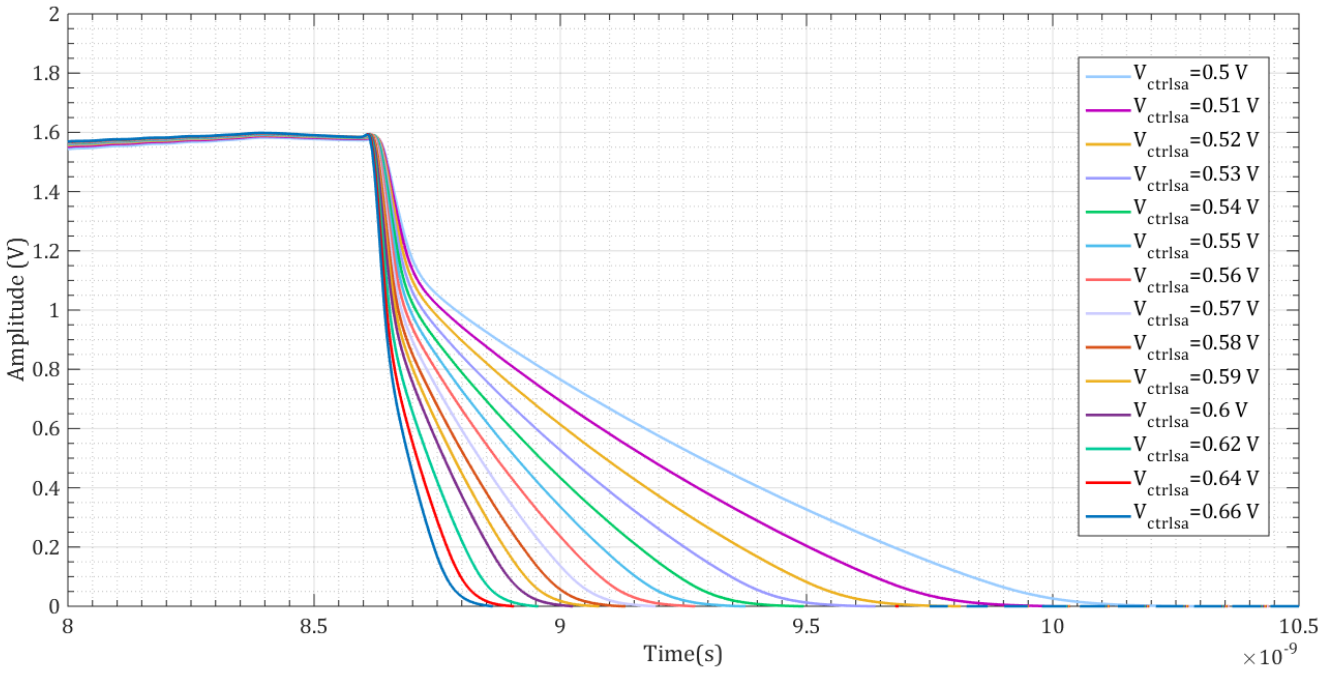


Figure 120 - Close-up of the analog sampling cell output's falling edge simulation for several V_{CTRL_sa} input values

3.5.2.2 The Edge Detector Line

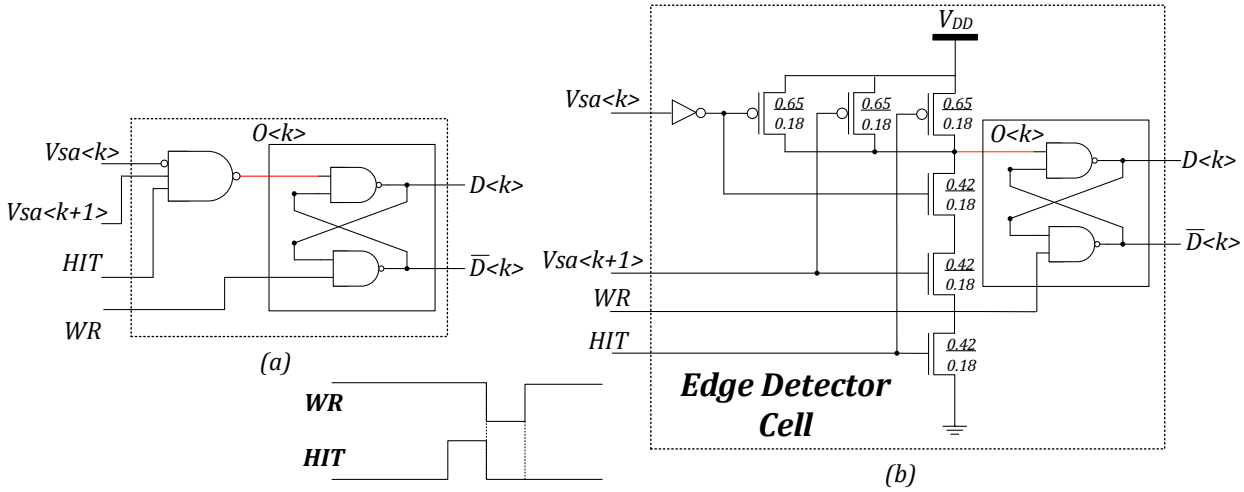


Figure 121 - Schematic cell of the standard edge detector (a) and the full custom edge detector (b)

The second stage of the time sampling unit (TSU) is the edge detector line which includes 32 cells that detect the stage where a high to low transition took place based on the 32 sampling unit outputs *Vsa<1:32>.* The schematic diagram of an edge detector cell is shown in (Figure 121), each cell<k> is connected to the parallel sampling cell's output *Vsa<K>* and the output of the *K+1* sampling cell Vsa<K+1>. A low transition is detected when Vsa<K>='1' and *Vsa<K+1>*='0'. The edge detector cell was designed with full custom cells instead of the provided standard cells after simulations showed a double detection for a 33.5 mV input value interval between 818.5 mV and 785 mV. In other terms, the simulation results showed that supposing Vsa<k-1> = '1' and Vsa<k+1> = '0', the 2 consecutive edge detector cells <k> and <k+1> will output a '1' indicating a low transition if the analog level of Vsa<k> is between 818.5 mV and 785 mV as shown in Figure 123. In order To limit occurrences of "double edge detection" cases, custom NAND cells were used to ensure more symmetric '1' and '0' switching thus reducing the previous 33.5 mV double detection input interval to the 13 mV interval [796mV - 809 mV] (Figure 123). Although a double detection scenario is still possible, the use of a custom ED makes it less recurrent and when it does the conflict will be resolved by the subsequent blocks of the TSU. Another conclusion to be drawn from these simulations is that the internal signal O<k> will switch to '0' for Vsa<k>=820 mV thus setting the '0' detection limit to 0.82V and not 0.9 V, this is a key information that will affect several aspects of the following blocks in charge of the Ultra-fine time conversion process. Finally, a Set and Reset NAND memory was added to ensure that the output remains stable when *Vsa<K>* is fluctuating around the low transition detection threshold value (818 mV) due to the charge injection noise source and the process is controlled by means of a WR signal which is a delayed inverted version of the HIT signal. The edge detection line outputs 32 bits consisting of 31 '0' bits with a '1' bit at the stage location where the low transition was detected, these outputs are transcoded using a 5 bit thermometer coder to limit the time

conversion output data rate. In the case of a double detection in cells <k> and <k+1, the thermometer coder was designed to output the bit code for <k>.

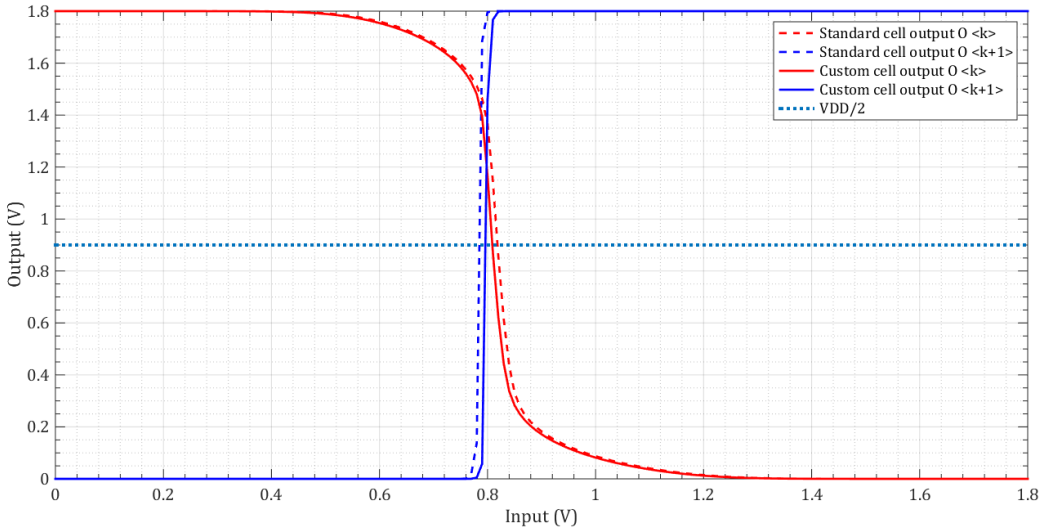


Figure 122 - DC simulation results of the internal O<k> signal in two successive edge detector cells using standard cells and custom cells

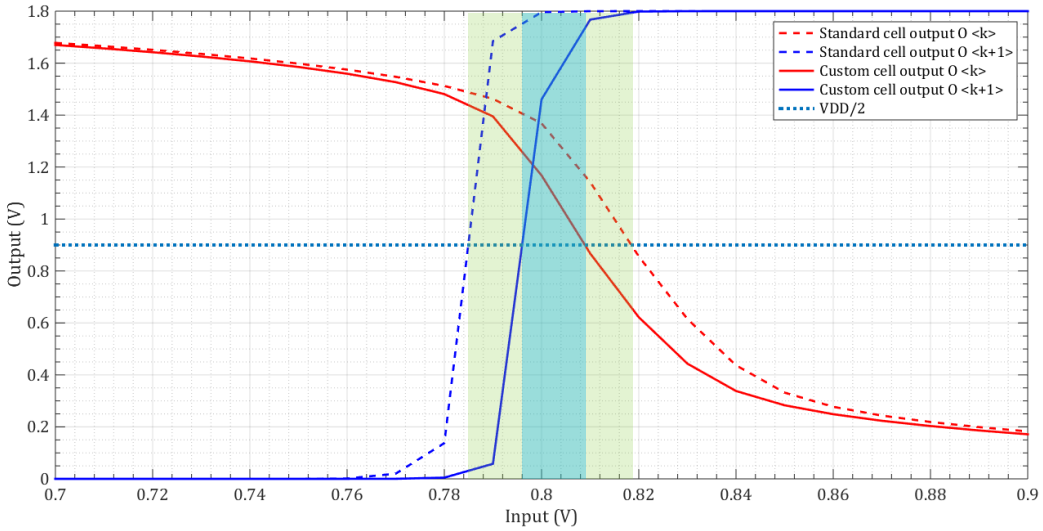


Figure 123 - Close-up of the DC simulation results of the internal O<k> signal in two successive edge detector cells, the Green area shows the original double detection interval using standard digital cells and the blue area shows the final interval obtained using custom digital cells

3.5.2.3 The Readout Line

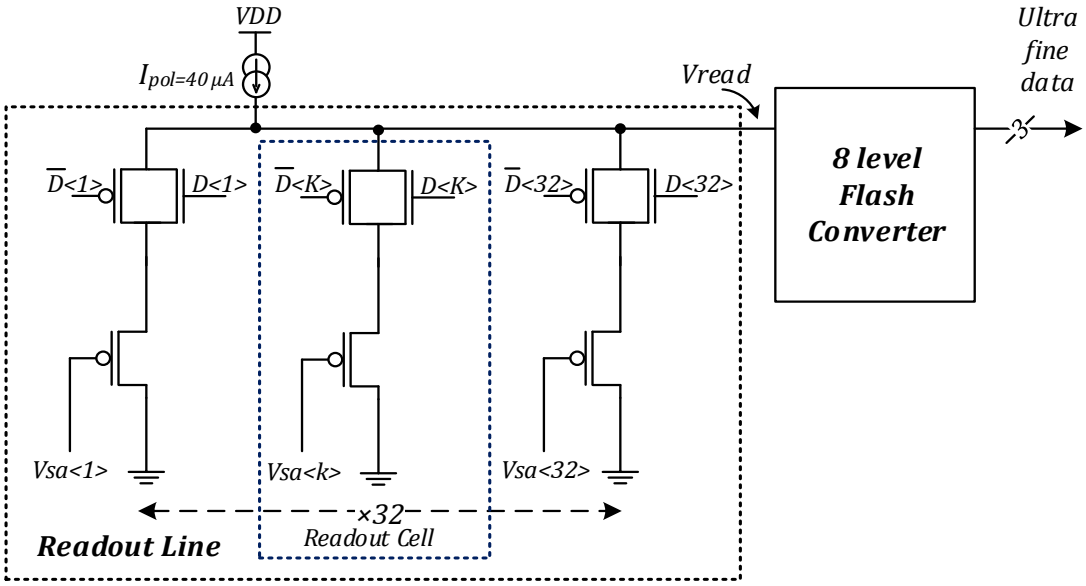


Figure 124 - Schematic bloc diagram of the 32 cells Readout line and the subsequent converter

The Read-out line (Figure 124) includes 32 Read-out cells consisting each of a PMOS source follower with a T-gate controlled by the Edge detector outputs thus connecting the Readout cell of the stage where a low transition was detected to the subsequent data converter's input. The Readout cell in this circuit serves as a voltage buffer allowing the transmission of the analog sampling cell output where a '1' to '0' transition was detected to the subsequent data converter. A PMOS source follower with an aspect ratio of 2.0/0.18 has been used in each Readout cell, the decision of using a PMOS source follower rather than an NMOS source follower was especially driven by the fact that a positive shift is more suitable for the targeted voltage interval i.e. [300mV, 800mV]as it will create a positive shift of approximately (Vth=~830 mV) thus shifting the voltage interval from to [1.13 V, 1.61 V] as shown in the DC simulation results (Figure 125). Despite a higher occupation surface, A PMOS source follower allows a near unity voltage gain, a better linearity due to reduced body effect and a better dynamic range. The small size body tied PMOS was thus chosen to limit dynamic errors caused by source switching and the presence of parasitic capacitance leading to parasitic effects such as charge coupling, glitches and noise contribution. Another problem that can be especially detrimental to the precision of the fine time conversion results is charge injection that leads to an increase of the output's amplitude thus causing distortions in the time conversion results was handled by using a TGATE. Although the TGATE requires two complementary control signals this was not an inconvenience in this case since the edge detector provides the 2 complementary signals needed ($D < k >$ and $\overline{D < k >}$). Figure 126 shows a the evolution of the output and the Readout cell's output simultaneously with the Analog sampling cell internal buffer's output and the Analog sampling cell's output (The TGATE was switched on for the sake of better clarity).

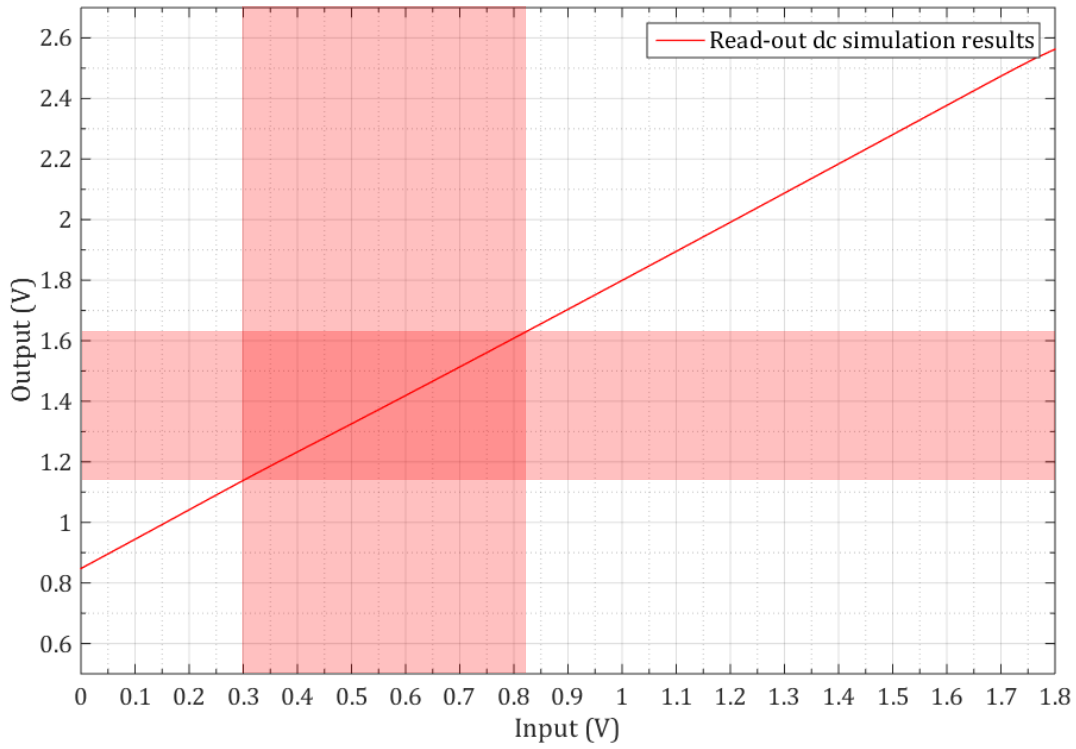


Figure 125 - DC simulation results of the Readout cells. The red zone represents the targeted linear low transition output interval in the analog sampling cell.

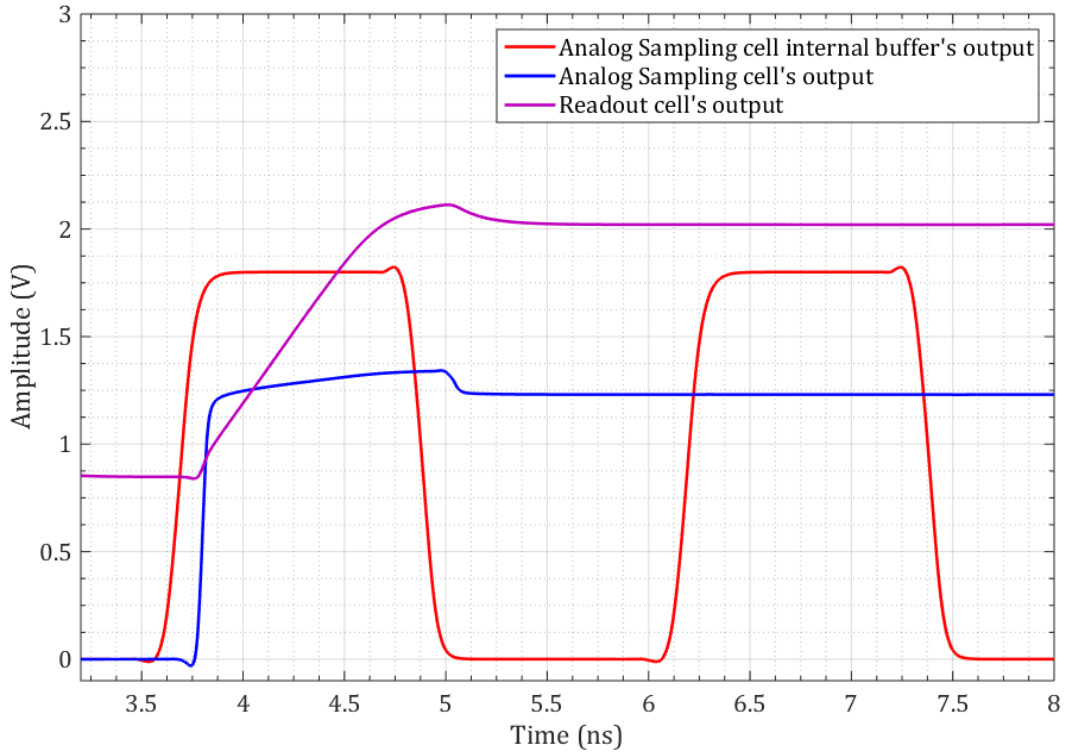


Figure 126 - Transient simulation showing the evolution of the Analog sampling cell internal buffer's output, the Analog sampling cell's output and the Readout cell's output

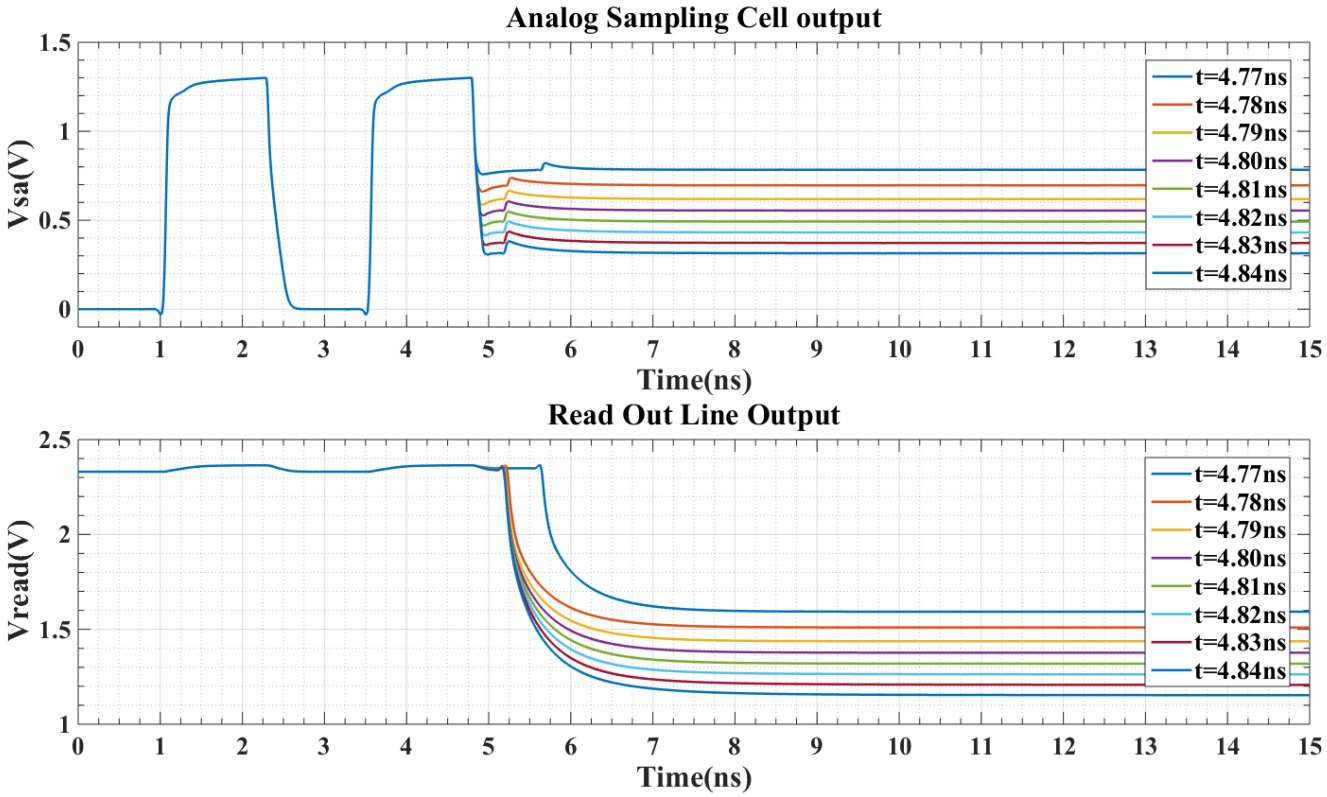


Figure 127 - Simulation results of (top) the Analog Sampling cell output sa<k>(V) and (bottom) the ReadOut line output Vread (V) for 8 HIT consecutive signal's arrival time with V_{CTRL_sa}=0.62V and f_{CLK}=400 MHz

Figure 127 shows the simulations results of the analog sampling cell output Vsa and the Readout Line output (Vread) for 8 consecutive HIT signal's arrival time separated by 10ps each and for a V_{CTRL_sa}=0.62V and a DLL clock input frequency of 400 MHz. The resulting signal levels are separated enough to obtain 10 ps time resolution via the following data converter. The results show also that the outputs in both figures become stable in less than 5ns thus showing that the ultra-fine time resolution does results in a reasonable dead time that allows a maximum measurement repetition rate of 200MHz.

3.5.2.4 The 3 Level Flash Converter

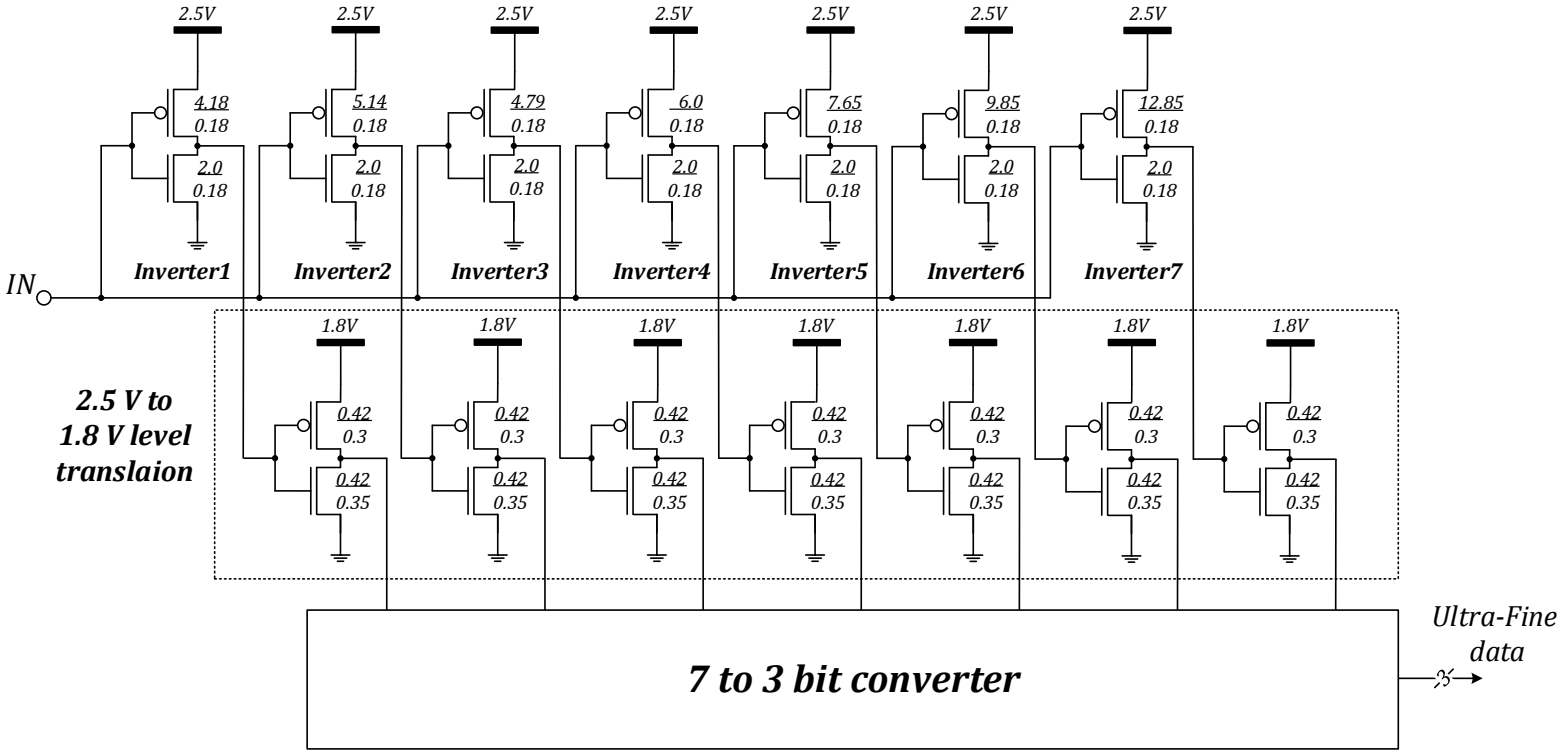


Figure 128 - Bloc diagram of the 3 level flash converter

The ultra-fine conversion scheme is based on determining more precisely the analog level of the digital '0' held in the analog sampling cell selected by the edge detector. The analog level of the detected '0' can be anywhere between 300 mV and 820 mV and giving that the slope between these 2 bounds is linear , dividing the 480 mV interval into 8 sub-intervals of 60 mV each, will allow us to increase the resolution obtained via the fine time conversion stage by 8 times. The former fine conversion result defined as

$$T_1 = (M \times 80)\ ps \tag{36}$$

Where M is the number of the stage where the low transition was signaled is therefore improved into

$$T_2 = \left[\left((M-1) \times 80\right) + (N \times 10)\right] ps \tag{37}$$

Where N represents the interval number comprising the analog voltage of the detected '0'. Assuming a 2.5 ns Clock period, the subsequent fine time resolution of 80 ps demonstrated earlier can thus be amplified into a 10 ps ultra-fine time resolution. This process achieved by means of the 3-level flash converter that outputs a 3 bit word representing the sub-interval number of the analog '0' level transmitted from the low transition stage through the Read-

out block resulting in an signal amplitude between 1.13 V and 1.61 V. The block diagram of the 3-level flash converters is shown in Figure 128. The analog level detection is obtained by 7 CMOS inverters operating with a 2.5 V power supply. The dimensions of their transistors were chosen to obtain 7 switching thresholds equal to the inner bounds of the 8 sub-intervals. As a result, when a detection occurs, only the inverters with a threshold value lower than the signal outputted by the read-out block will switch from '1' to '0' (Figure 129) Thus leading to a more precise evaluation of the digital '0' transmitted through the Read-out block.

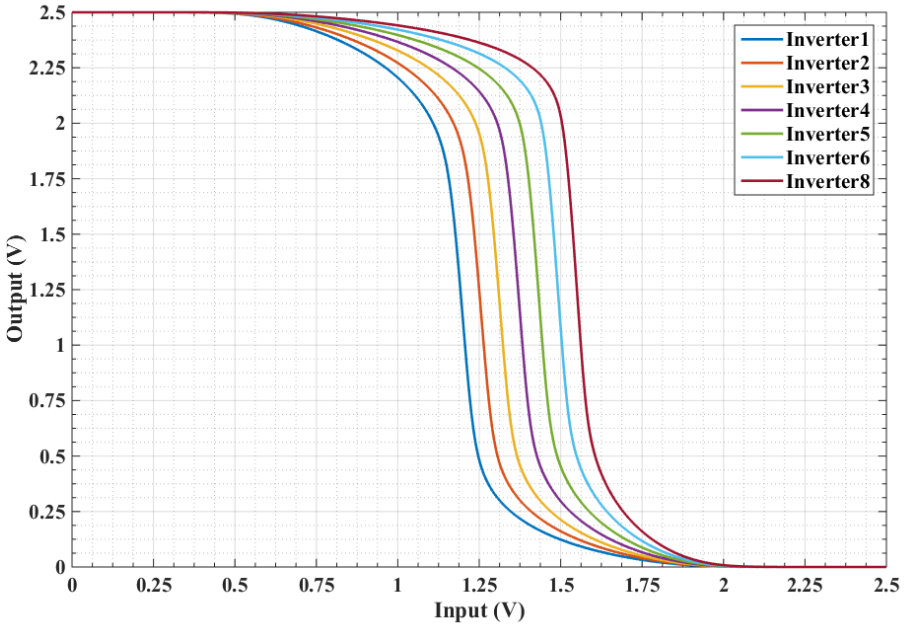


Figure 129 - DC simulation results of the 7 inverters used in the first stage of the 3-level flash converter

The 7 outputs of first stage are connected to a second stage consisting of 7 CMOS inverters powered by a 1.8 V supply source. The second stage inverters were designed such that their switching threshold is equal to 1.25 V instead of VDD/2 thus serving as an intermediate between the 2.5 V inverters and the subsequent digital 7 to 3 bit coder operating with a 1.8 V power supply and used to limit the ultra-fine data rate by transcoding the 7 bits obtained via the 7 inverters into 3 bits representing the actual value between 0 and 7. The employed schemed allows a fast digital conversion of the analog signal connected to the FC input with reasonable power and area occupation at the price of an additional 2.5 V power supply. The switching thresholds of the 7 inverters are summed in (Table 7) with their respective transistors dimensions. These transistors were not designed with minimum width value in order to limit possible variations affecting their threshold values. Monte-Carlo simulations were performed to determine the statics of variations due to device mismatch, the simulation results obtained for 1000 runs are represented in Figure 130-Figure 136 and show standard deviations of less than 3.5 mV which represents a reasonable variation ratio compared to the step size of 60mV.

Table 7 - Switching thresholds of the 7 inverters used in the first stage of the 3-level flash converter

	W/L (PMOS)	W/L NMOS)	Switching threshold (V)
Inverter 1	4.18/0.18	2.0/018	1.19
Inverter 2	5.14/0.18	2.0/018	1.25
Inverter 3	4. 79/0.18	1.5/018	1.31
Inverter 4	6/0.18	1.5/0.18	1.37
Inverter 5	7.65/0.18	1.5/0.18	1.43
Inverter 6	9.85/0.18	1.5/0.18	1.49
Inverter 7	12.85/0.18	1.5/0.18	1.55

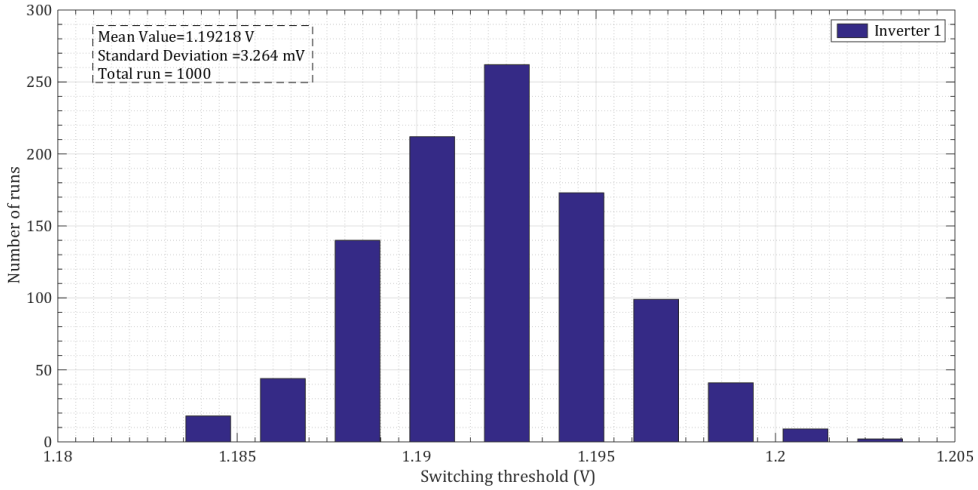


Figure 130 - Distribution of the switching threshold of Inverter1 based on device mismatch Monte-Carlo simulation

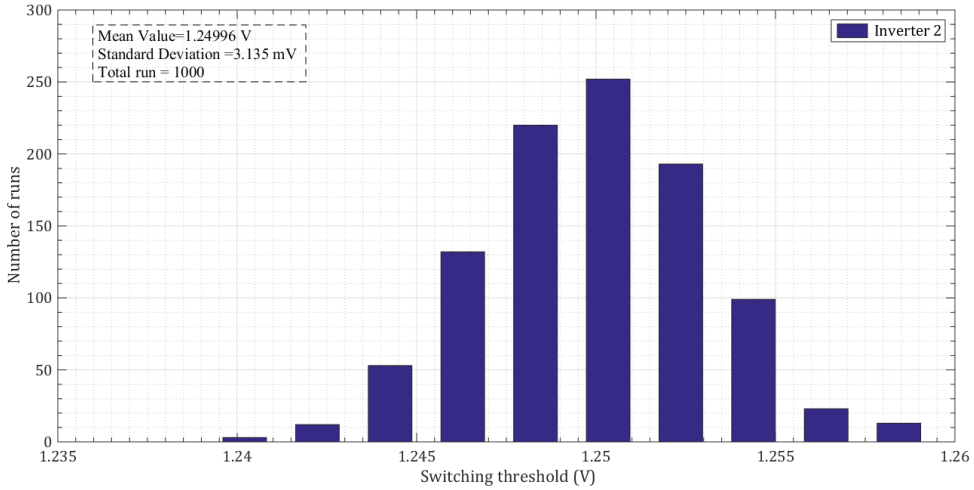


Figure 131 - Distribution of the switching threshold of Inverter2 based on device mismatch Monte-Carlo simulation

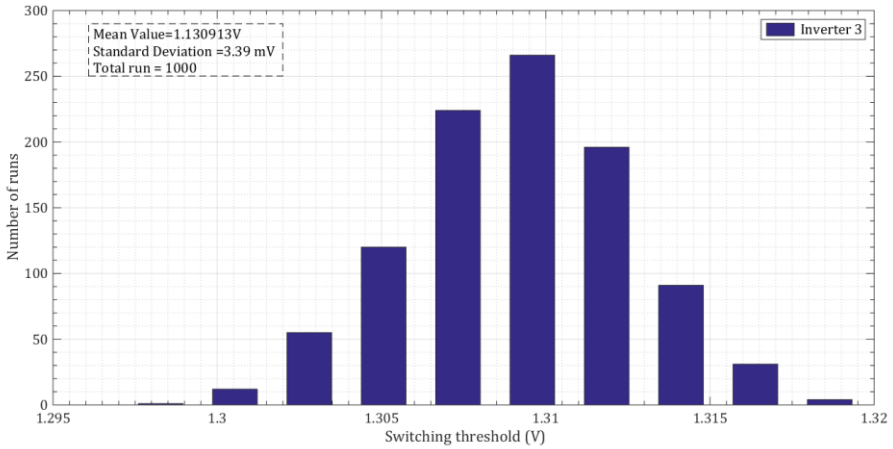


Figure 132 - Distribution of the switching threshold of Inverter3 based on device mismatch Monte-Carlo simulation

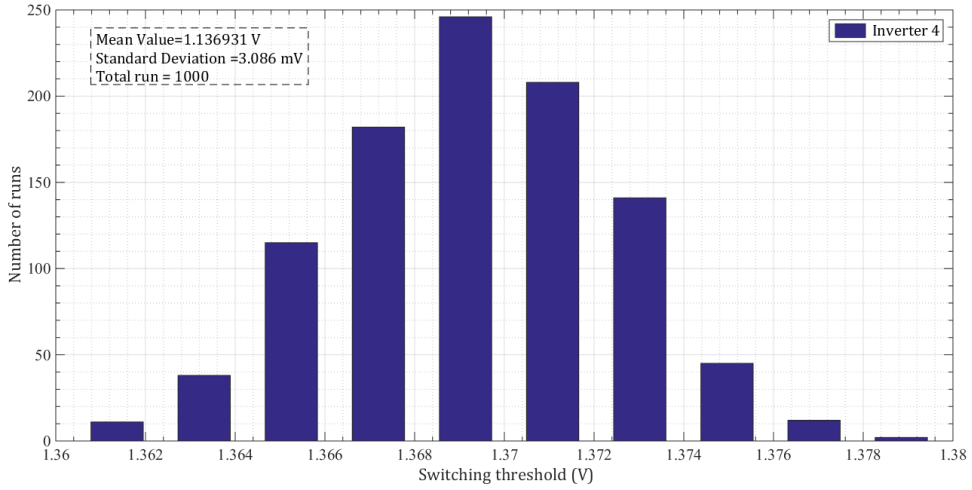


Figure 133 - Distribution of the switching threshold of Inverter4 based on device mismatch Monte-Carlo simulation

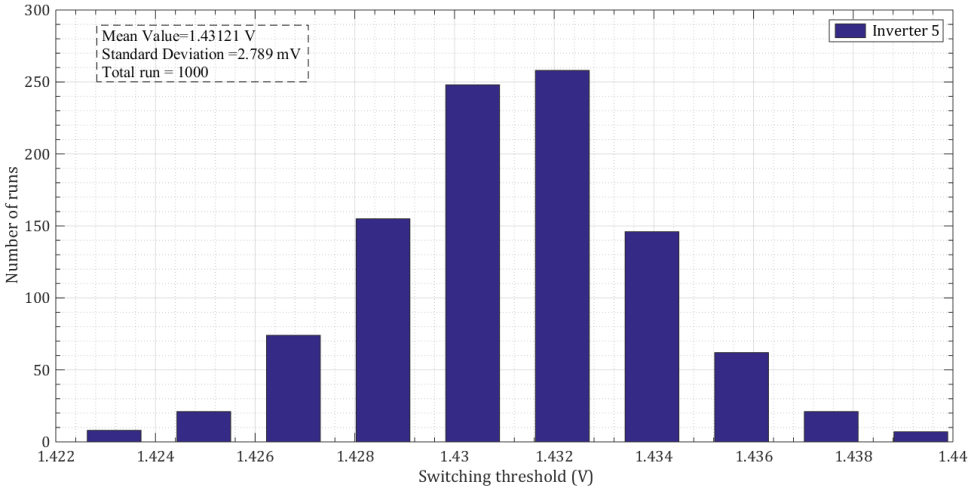


Figure 134 - Distribution of the switching threshold of Inverter5 based on device mismatch Monte-Carlo simulation

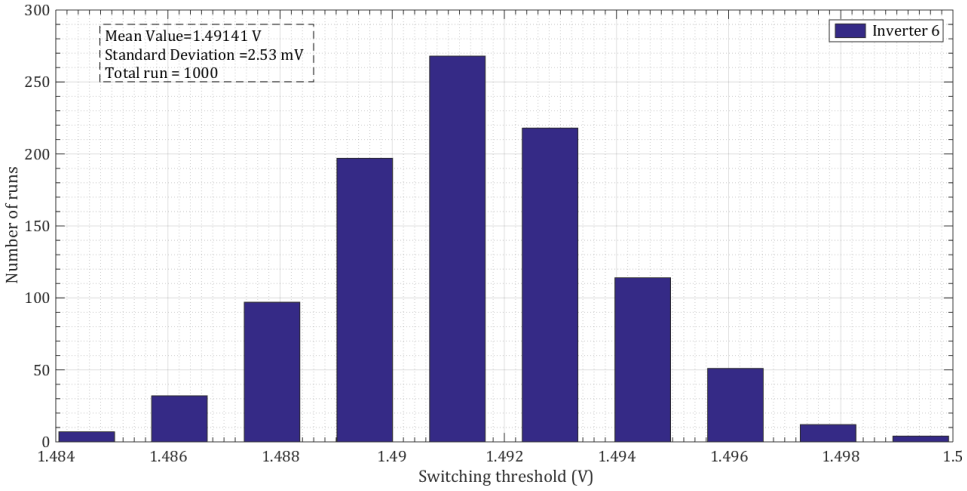


Figure 135 - Distribution of the switching threshold of Inverter6 based on device mismatch Monte-Carlo simulation

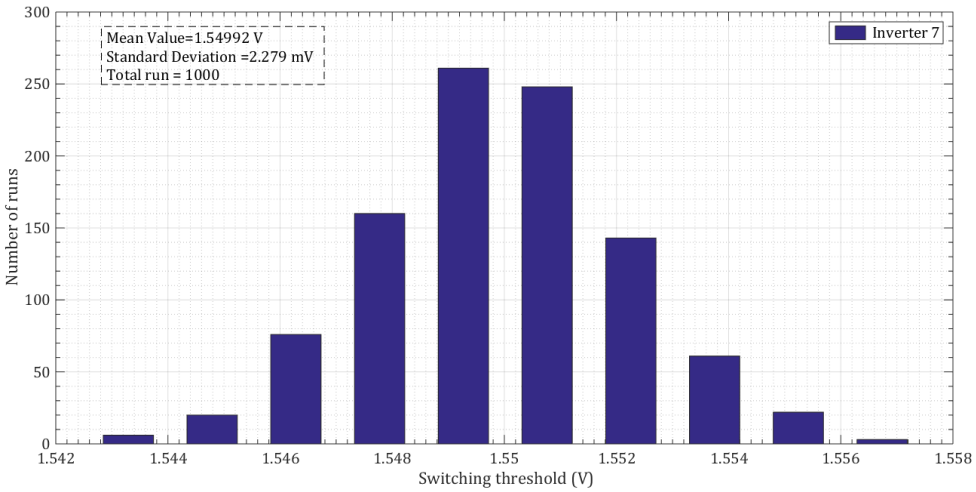


Figure 136 - Distribution of the switching threshold of Inverter7 based on device mismatch Monte-Carlo simulation

3.5.2.5 TSU State Machine

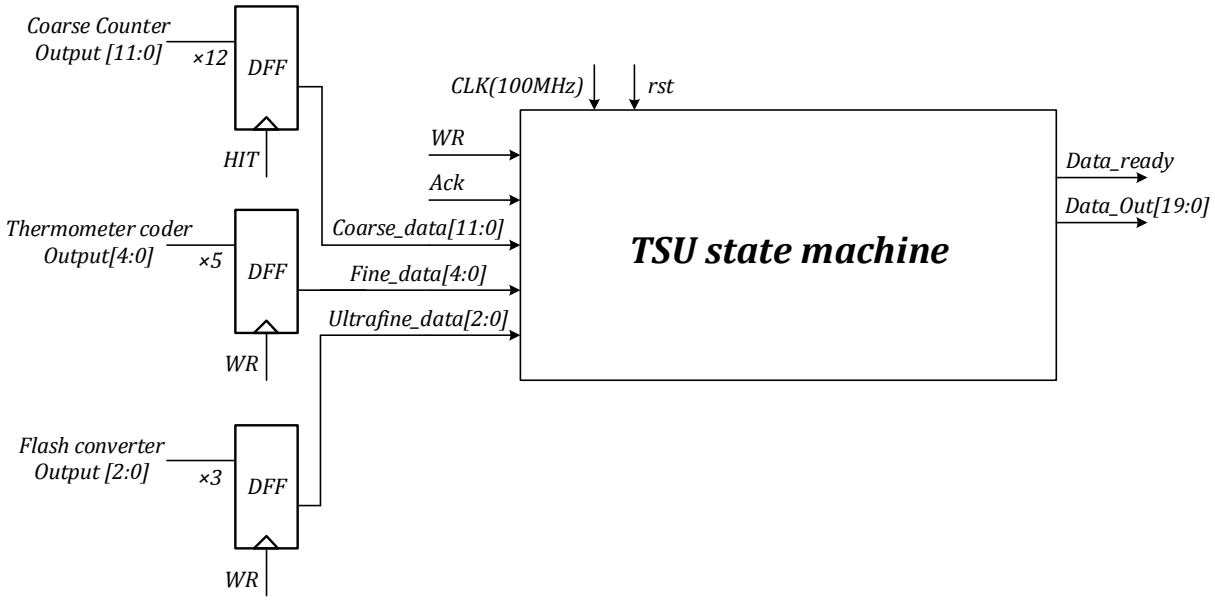


Figure 137 - Bloc diagram of the Conversion data transmission scheme

The presented conversion scheme results in 20 bit results where 12 bits represent the coarse time measurement results outputted by the coarse counter, 5 bits represent the fine time measurement results outputted by the thermometer coder and 3 bits represents the ultrafine time measurement results outputted by the 3 bit Flash converter. The 20 bit unit time conversion results are collected and transferred by the TSU state machine to the ensuing data processing unit, an intermediate block of 12 DFF is to collect the results as following: the coarse counter outputs are copied upon the arrival of the Hit signal, a second signal 'WR' generated after 10 ns of the Hit signal's arrival is connected to the clock input of the remaining 8 DFFs thus allowing the registration of the fine and ultra-fine measurement results once the Flash converter's output has stabilized (Figure 127). This 2 step intermediate data collections scheme was devised because unlike the fine time conversion results and the ultrafine time conversion results that requires a minimum wait of 5 ns before they are collected by the TSU state machine, the coarse measurement results must be immediately registered upon the arrival of the Hit signal otherwise waiting could lead to a huge conversion error since the coarse counter is connected to the 2.5 ns VCDL's output. The communication with the data processing unit is done through 2 signals: the output Data_ready signaling to the data processing unit that a new result is ready to be transmitted and the input "Ack" indicating that the data was taken into account by the following Memory block otherwise the TSU machine will remain in wait until the data transfer is acknowledged meanwhile the intermediate DFF block will ensure that the following time conversion result is saved until the TSU is free again.

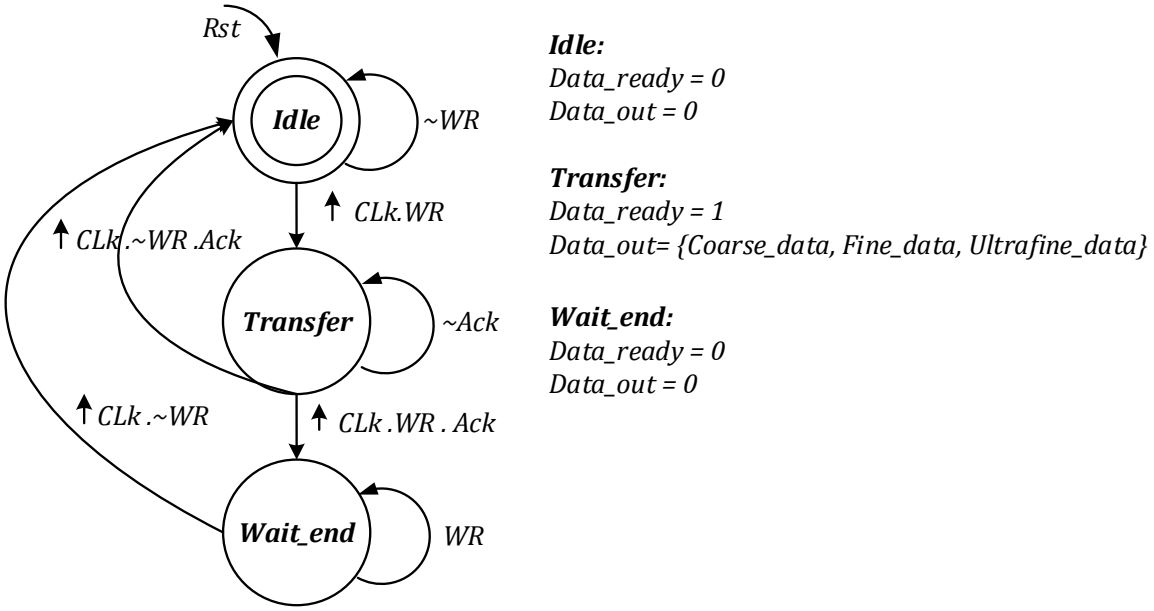


Figure 138 - state transition diagram of the TSU state machine

The state transition diagram of the TSU state machine is illustrated in Figure 138. Initially the TSU machine is in the "Idle" state and the 2 outputs "Data_out" and "Data_ready" are set to low. The TSU state machine is connected to a 100 MHZ clock signal, when the WR signal rises to '1' 10 ns after the arrival of the 'Hit' signal, The state machine moves to the 'transfer' state, the 20 bit time conversion total results are outputted with the output signal 'Data-ready' signaling that a valid data is presented. the data transmission is ended once the signal 'Ack' that goes to '1' signaling that the data was taken into account by the following Memory block otherwise the TSU machine will remain in the 'transfer' stating. If the transmission was finalized before the end of the HIT signal the TSU state machine will evolve to the "wait_end" state before going into the "idle" state again otherwise it will transitioned directly to the "idle" state.

3.5.3 Simulation Results

The described TSU block implemented in a 180 nm standard CMOS was simulated with cadence for a DLL input clock of 2.5 ns, $V_{ctrl\text{-}sa}$ = 607 mV and while varying the '*HIT*' time arrival with a 1 ps step in order to better evaluate the 3 stages conversion scheme conducted by the UFC in terms of precision and linearity. Figure 139 shows the simulation results of the fine and ultrafine conversions over a range of 90 ps and Figure 140 shows the time conversion results obtained at the end of the total conversion process. By combining the fine and ultrafine conversions outputs, the initial fine time resolution of 80 ps was in indeed increased to 10 ps without the need of using a higher clock input. The simulations also showed a non-linearity level estimated at ~2 ps resulting from the nonlinearity of the stretched descending slope outputted by the analog sampling cell in the chosen voltage interval [0.82 V, 0.3 V]. Ideally, by stretching the 0.52 V intervals over an 80 ps time interval, each 65 mV drop should be equivalent to a time interval variation of 10 ps. However the minor divergence between the real and the ideal falling slopes (Figure 141) leads to variations in the time interval lengths equivalent to each voltage subintervals resulting in nonlinearities in the ultrafine conversion results and as a result the total time conversion results. One way of correcting this divergence is by implementing a nonlinear Flash converter instead of the used 3 bit Flash converter with a uniform step size of 65 mV thus ensuring that each voltage sub-interval will be equivalent to a time interval of 10 ps.

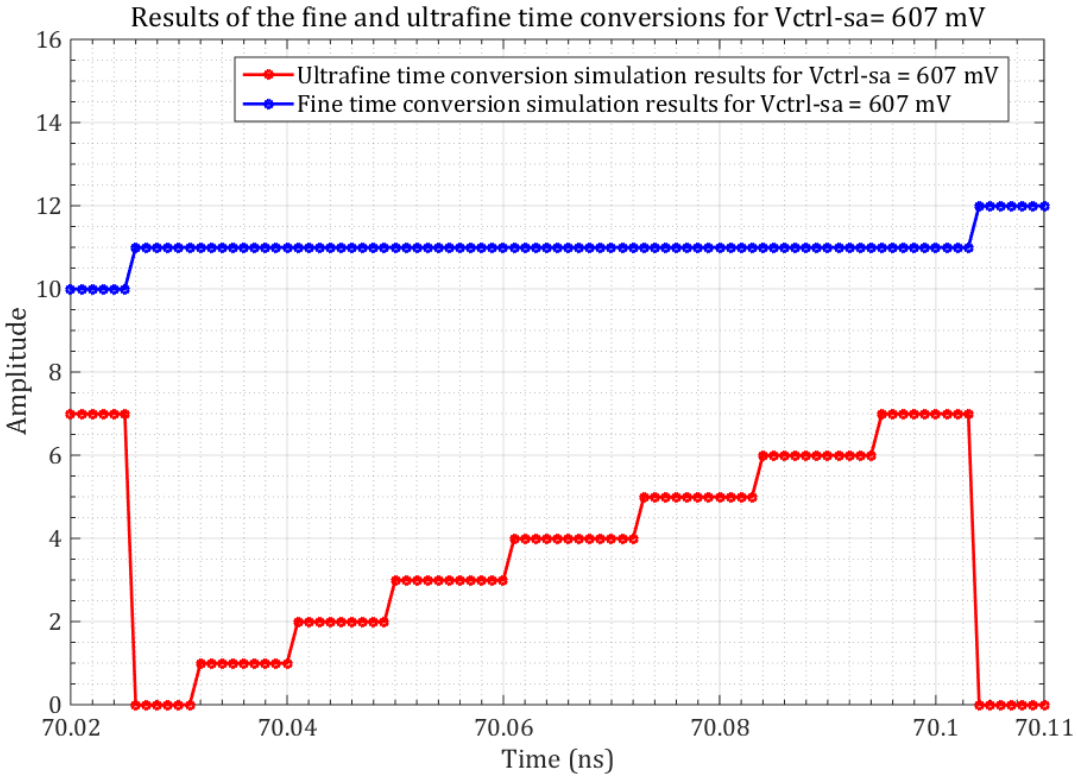


Figure 139 - Simulation results of the fine and ultrafine conversions over a range of 90 ps

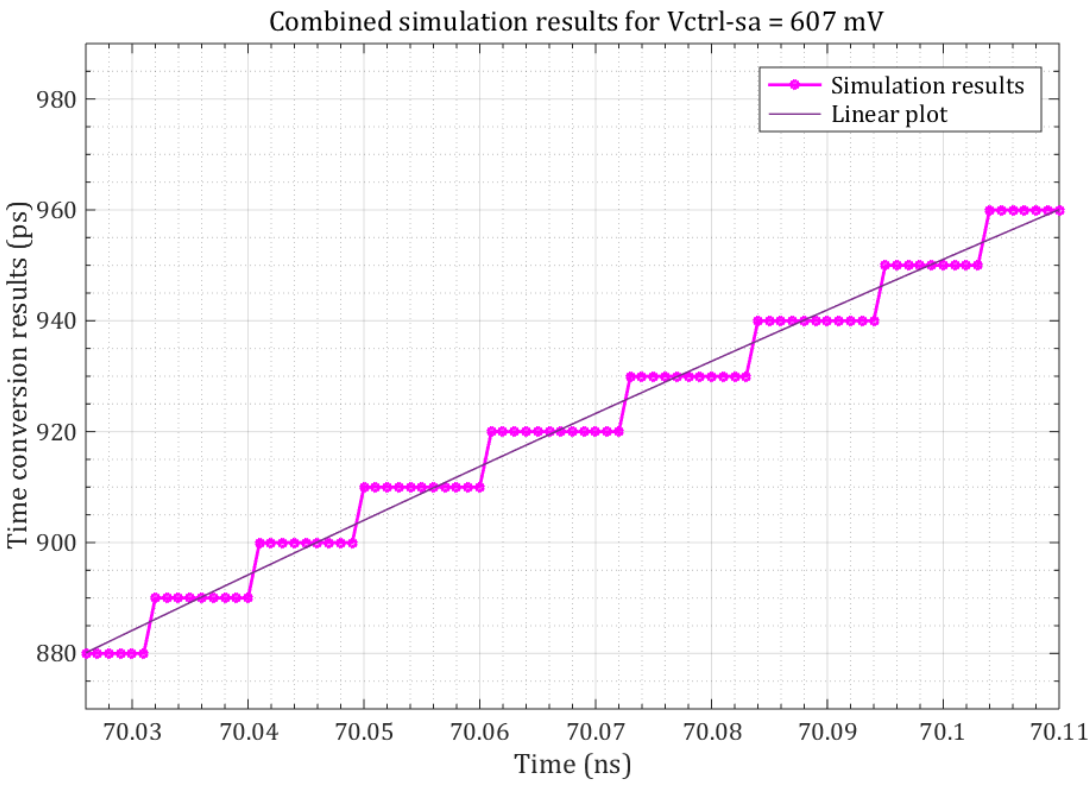


Figure 140 - Simulated results of the total time conversion scheme over a range of 90 ps

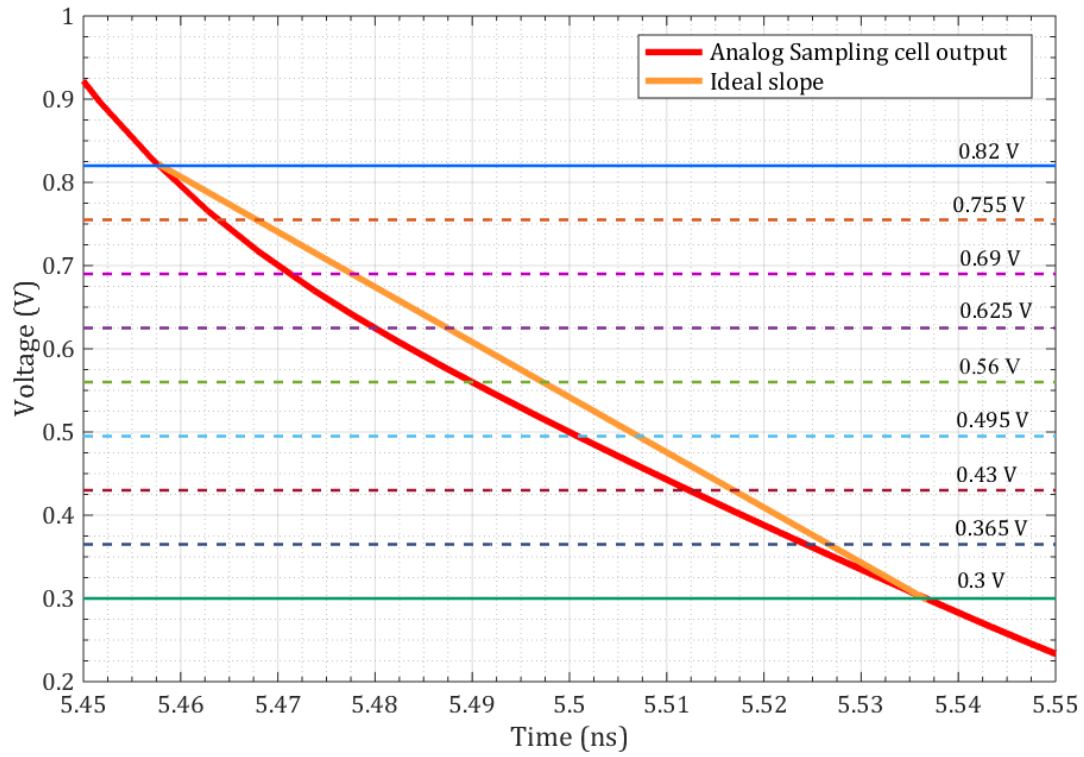


Figure 141 - Divergence between the falling slope of the analog sampling cell and an ideal slope with the 8 sub-intervals voltages bounding the 8 time sub-intervals

3.5.4 Characterization Results

The hybrid TDC was fabricated in a standard 180 nm CMOS technology and measurements were conducted to evaluate the fine and ultrafine time conversion operations accomplished by the TSU block of the TDC. A "Stanford research system DG645" pulse/delay generator was used to perform the measurements. The nominal operation should be to apply a clock signal to *CLK_In* and to trig the delay generator with the *Start* Signal generated by the TDC chip. However the DG645 delay generator is specified with a jitter of less than 25 ps in respect to the signal applied at its input trigger so in order to evaluate the 10 ps resolution of the TDC, this jitter has to be reduced. The trigger jitter of the DG645 can be canceled by operating the delay generator with its own internal trigger and generating both a *Clock_In* pulse to be propagated through the VCDL and a delayed version of this pulse to be connected to the *HIT* input of the TSU by two outputs of the delay generator. The residual jitter obtained between these two signals using this setup was evaluated by means of a "LeCroy wave master 13Zi" oscilloscope and it was found to be less than 7 ps rms. The averaged results for 130 measurements of the fine time conversion, ultrafine time conversion and the total time conversion outputs are illustrated for a measurement range of 2800 ps in Figure 142. The improvement in the achieved time resolution after combining the ultrafine time conversion results is clear despite fluctuations in the UFC results (Figure 143) that increase for higher stages. The fluctuations are most likely a consequence of the circuit's layout. The problem can be rectified by better power distribution to ensure a proper voltage levels in all the UFC stages for correct operation of the logic blocks, by adding decoupling capacitors to limit parasitic supply voltage variations and by using wider wires to limit the voltage drop due to wire resistance. A close up view on the measurements over a range 500 ps of with the fine time conversion results, the ultrafine time conversion results and the total results obtained by combining the fine and ultrafine conversions data is shown in Figure 144, the improvement obtained using the 3 stages conversion scheme is quite noticeable although the results characteristics were smoothed due the 7 ps rms jitter (40 ps peak to peak) associated to the measurements. Furthermore the INL (Figure 145) is improved from 21 ps rms using the fine time conversion results to ~ 5.6 ps rms using the complete conversion scheme. Furthermore, the minimal delay step of the Stanford DG645 is 5 ps which is quit high compared to the TDC resolution, thus the measured TDC performance are limited by the experimental characterization setup and the INL could be even better.

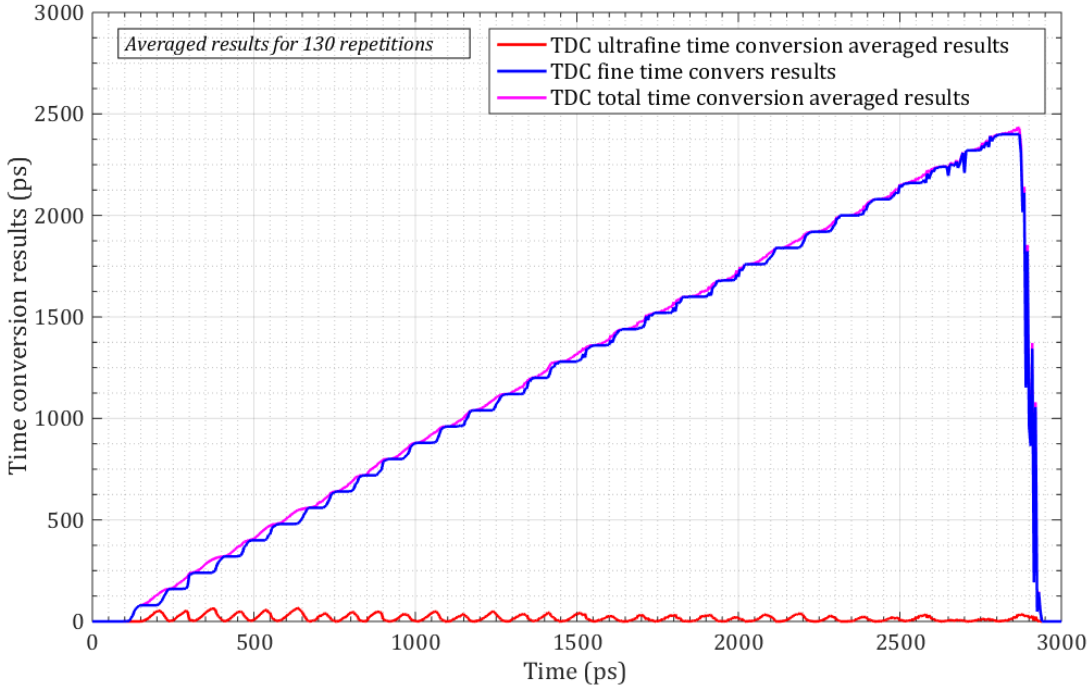


Figure 142 - Averaged measurements results over a range of 2800 ps for the fine time conversion, ultrafine time conversion and the total time conversion outputs

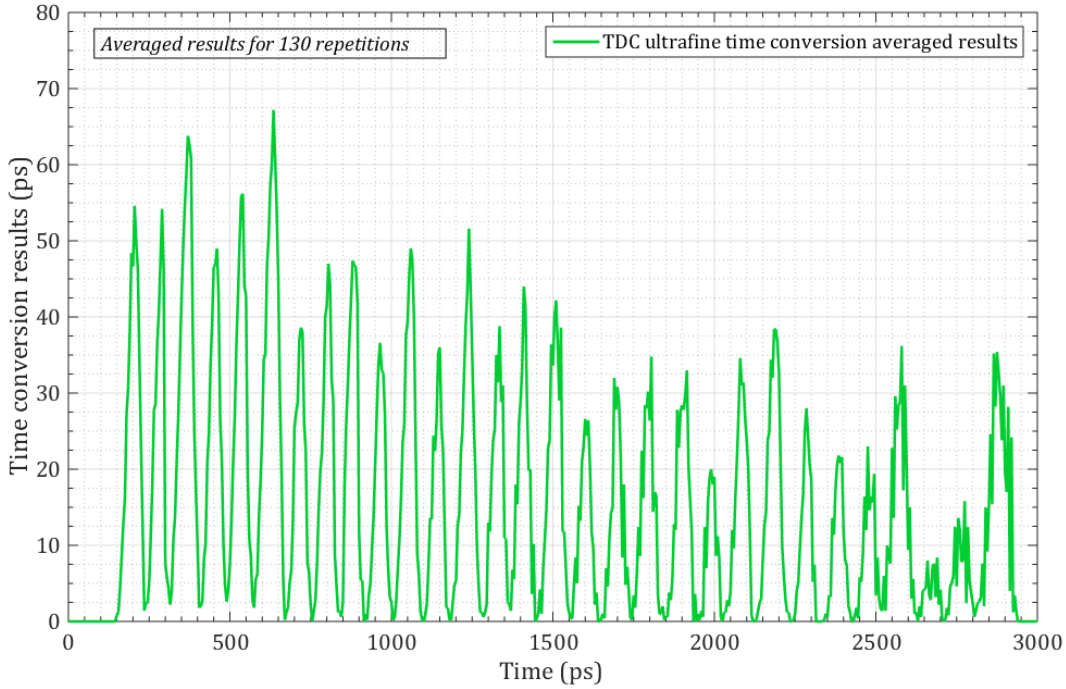


Figure 143 - Averaged measurements results over a range of 2800 ps for the ultrafine time conversion output

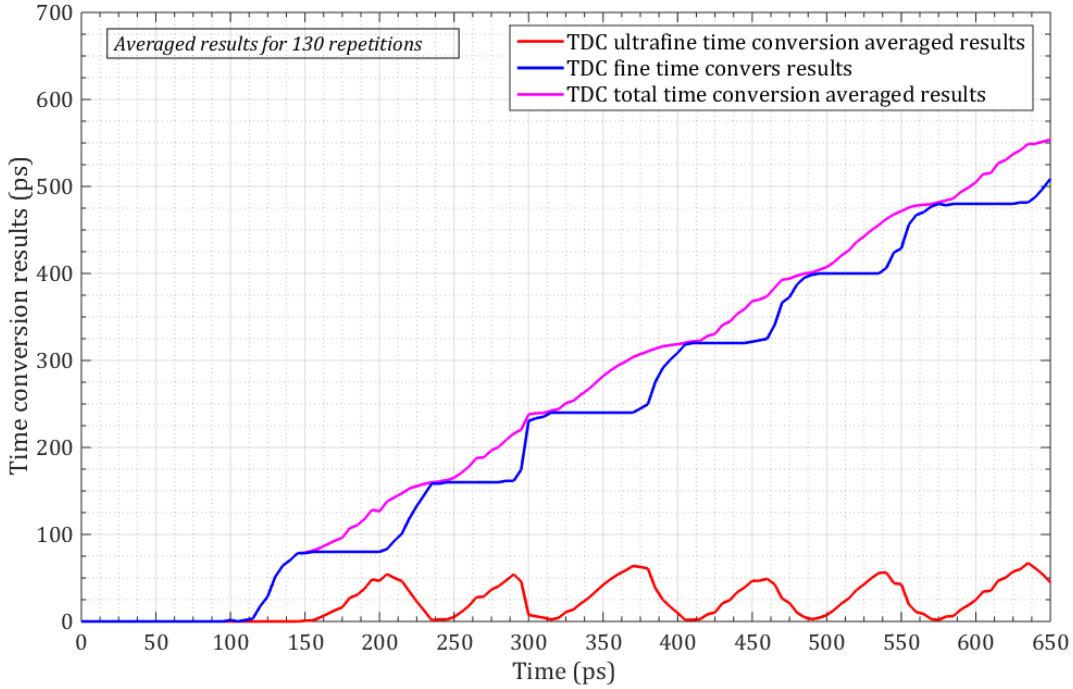


Figure 144 - Close-up of the averaged measurements results for the ultrafine, fine and total time conversions measurements over a range of 500 ps

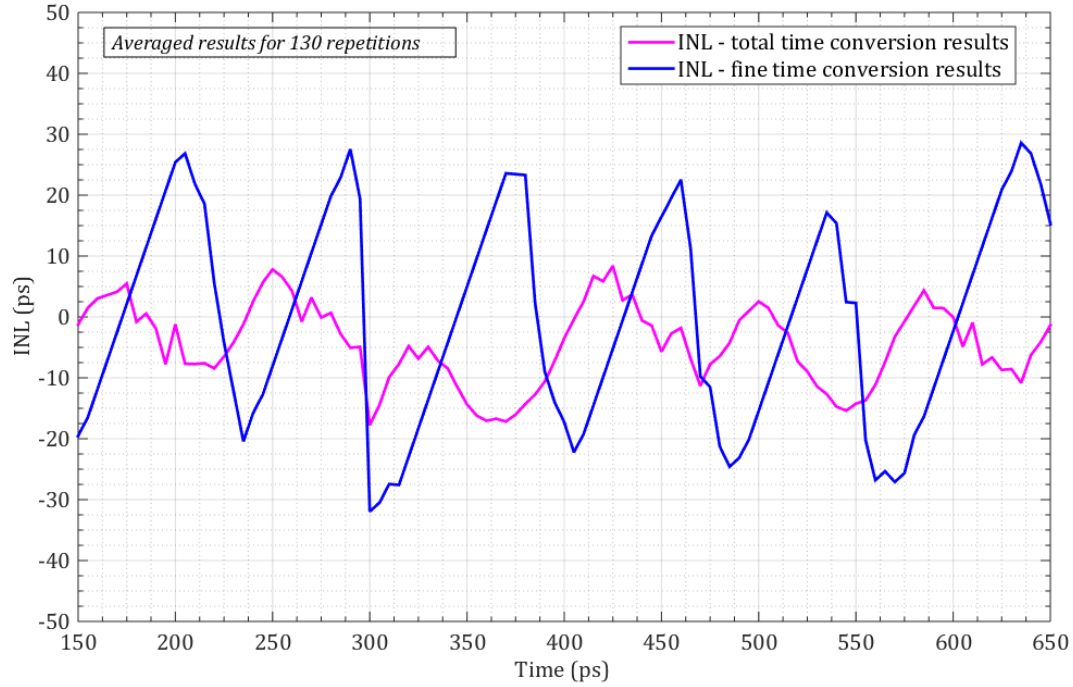


Figure 145 - Close-up of the estimated INL for the fine and the total time conversions measurements over a range of 500 ps

3.6 Conclusion

A hybrid TDC design capable of achieving high adjustable time resolutions with large dynamic range was presented and demonstrated. The design employs a time conversion scheme that combines traditional Analog Time to Amplitude Converter (TAC), Digital DLL-based and counter-based TDC techniques to obtain a high adjustable time precision resulting in a 3 bits enhancement of the least significant bit resolution (LSB). The hybrid TDC was especially designed to be used in a TCSPC system that incorporates an array of TDCs thus requiring a careful design to limit power consumption and occupation area. This was made possible by means of a 3 stages time conversion scheme that decentralizes the process by dividing it among the 3 parallel stages of the hybrid TDC. The common Delay locked loop (DLL) and coarse counter approach makes the design flexible and easily scalable allowing the conception of larger TDC arrays without the need to implement several DLLs and coarse counters. A TDC array englobing 8 TSU with a common DLL and coarse counter (

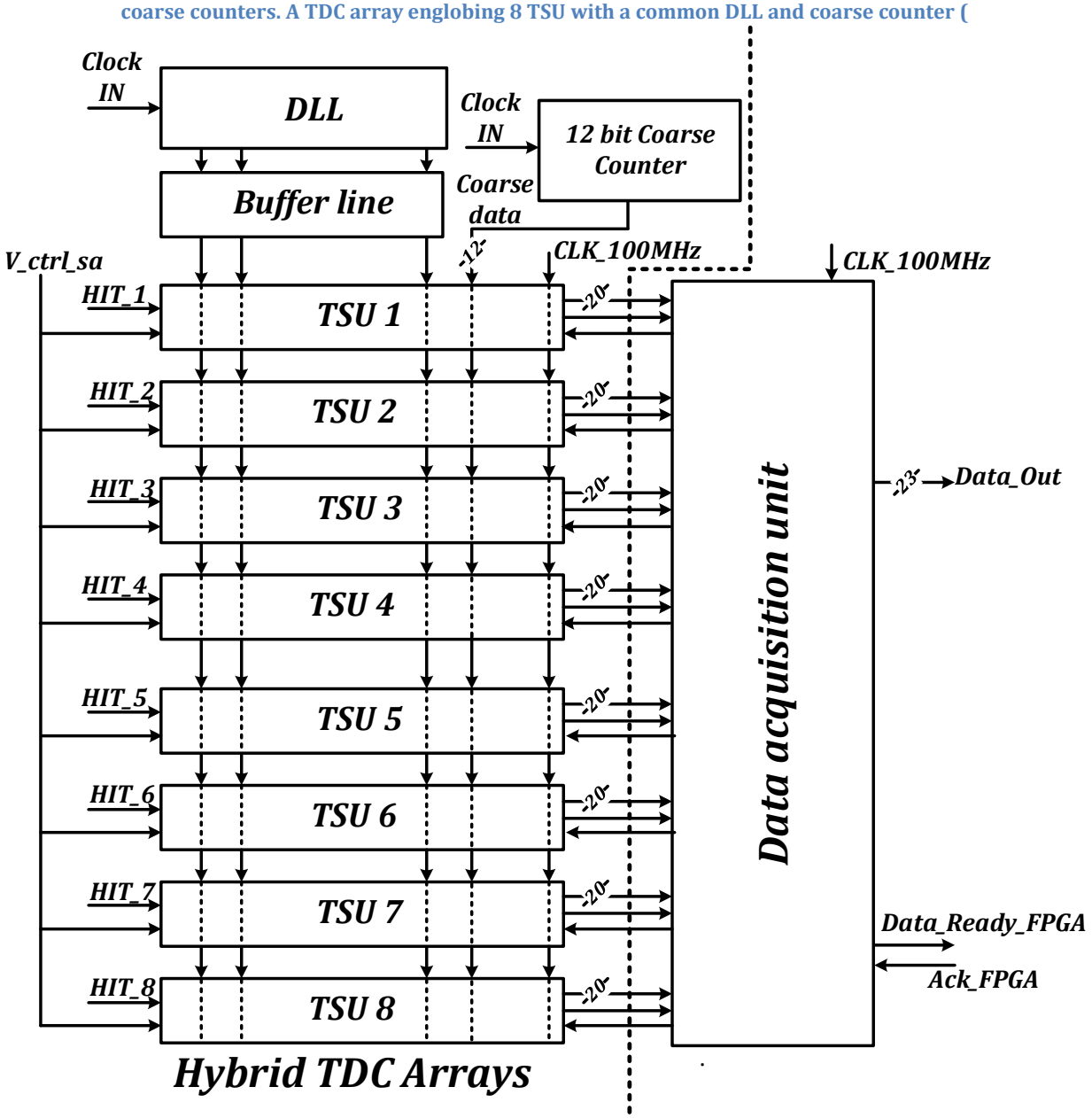


Figure 146) was realized based on this architecture. Each TSU operates independently by following the common DLL taps and sampling them when its corresponding HIT signal

rises to '1'. The hybrid TDC implemented in a 180 nm Standard CMOS technology (Figure 147) was specially designed to operate with variable input frequencies; the maximal speed operation using a 400 MHz clock input yields a 10 ps LSB and an estimated INL of ~ 5.6 ps rms compared to an 80 ps LSB and an INL of 21 ps rms sing the DLL based TDC alone. The use of the counter based TDC technique resulted in a maximum dynamic range of 10 µs instead of 2.5 ns. The maximum estimated power consumption of the hybrid TDC is 44 mW for a 20 MHz conversion rate with the global DLL consuming 35 mW resulting in 9 mW power consumption per TDC channel resulting in an important decrease in the total power consumption of the TDC arrays owing to the shared DLL approach.

Table 8 - Performance figures of the hybrid TDC compared to several state of the art integrated TDCs

	This work	[134]	[45]	[135]	[136]	[43]
Resolution	10 ps	400 ps	62.5 ps	97.5 ps	5 ps	17 ps
Dynamic range	10 µs	400 ns	64 ns	256 ns	13 ns	160 ns
Dead time	15 ns	-	-	-	150 ns	150
DNL	< 0.5 LSB	4.9%	< 4 LSB	0.2 LSB	-	-
INL	<5.6 ps	11.7%	< 8 LSB	0.3 LSB	-	< 4 peak
Average power	44 mW [1]	6mW	26.4 mW	175 mW	160 mW	< 15 mW
Area (mm^2)	0.6 × 0.36	-		1.6×1.8	1.2×1.1	0.3

[1] with 35 mW consumed by the DLL alone

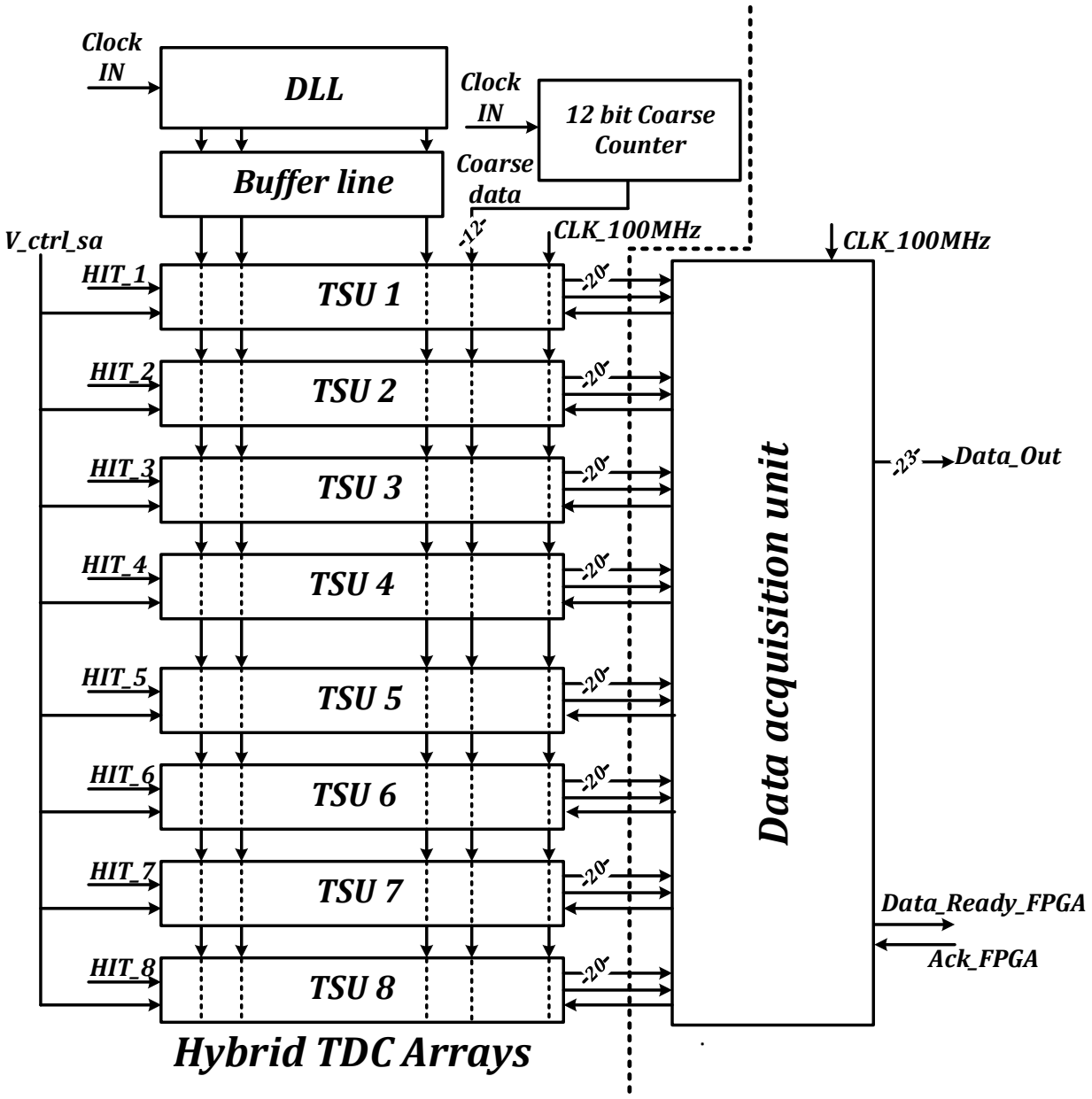


Figure 146 - Block diagram of the TDC array incorporating 8 independent TSUs with common DLL, coarse counter and a Data acquisition unit

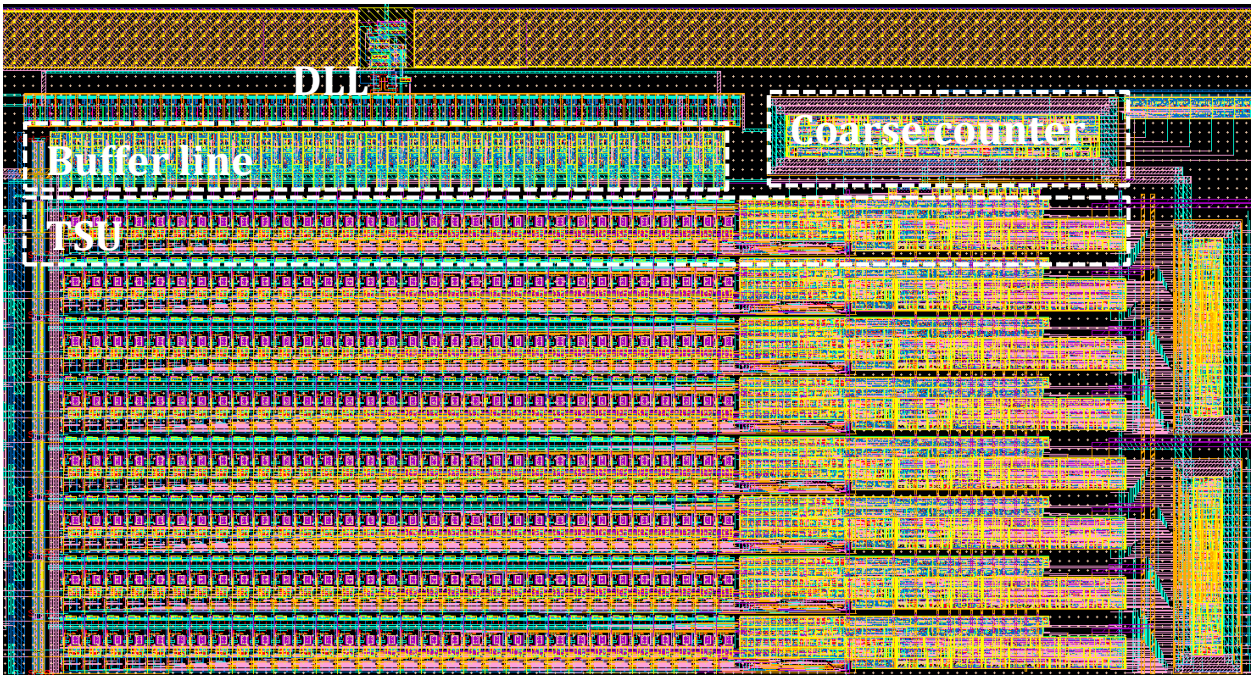


Figure 147 - Layout of the 8 TDCs array englobing 8 TSUs with shared DLL and coarse counter

Chapter 4
The TCSPC Streak Camera

Following the emergence of Silicon Photo Multiplier (SiPM) based on single photon avalanche diodes (SPADs) integrated in almost standard CMOS technology in the 2000's, several integrated 2D-TCSPC systems have been demonstrated lately. These 2D systems consist of SPAD arrays integrated in 2D with their associated electronics to form smart pixels resulting in a tradeoff between high photon detection efficiency and advanced electronic functionalities [65] [62] [68]. The use of a 3-dimensional (3D) heterogeneous integration with deep-submicron CMOS readout electronics represents a good solution to the previously mentioned tradeoffs but the 3D technology is still new, immature and highly costly. In this chapter we present another alternative to design TCSPC systems with both high efficiency and developed integrated functionalities by means of a photon counting integrated streak camera architecture that takes advantage of the streak mode imaging to overcome the space limitation inherent to 2D sensor arrays. This cost-effective solution allows the integration of complex functionalities in the pixel without the inconvenience of low fill factor that leads to low detection efficiency. A test structure with 8 units (or Macropixels) was developed in a 180 nm standard CMOS technology using this streak architecture but the final aim is to design larger structures with more units. The presented TCSPC streak camera test structure was designed to demonstrate the feasibility of a SPAD based streak camera but another aim was to design a TCSPC system for high throughput fluorescence life time applications. Consequently, the design was especially conceived to optimize the SNR and increase the efficiency by limiting counting losses probability due to the random nature of the photon arrivals in these measurements. This section is thus organized as following. We start by introducing the architecture of the elementary units, then we elaborate on the operation process following a photodetection, next we introduce the TCSPC Streak camera global architecture and finally we elaborate on the system optimization for High throughput screening of biomolecules.

4.1 TCSPC Unit

4.1.1 Architecture

The TCSPC unit is the building block of the TCSPC system; it was designed such that it can be easily used to implement larger arrays by assembling multiple units with a vertical pitch of 32 µm. A simplified bloc schematic of a TCSPC unit is represented in Figure 148; it consists of two main parts: a Macropixel for photodetection and a time sampling unit for time measurement of the detection event. The sub-blocks of the TCSPC unit are arranged horizontally with the SPAD at the beginning followed by the quenching electronics then the TSU block, this allows us to regroup the photon sensitive area in the form of a line thus ensuring that the overall photo sensitive area is not separated by the rest of the electronics thus increasing the effective Fill Factor compared to an array structure. It also allows us to easily regroup the TSU in the form of array sharing a common DLL and a coarse counter as discussed in

The simplified bloc schematic of the Macropixel operating as a single photon detector is illustrated in Figure 149; it consists of a Single Photon detector (SPAD) with its associated quenching electronics. The SPAD was designed with a hexagonal form and an active area diameter of 15 µm resulting in a fill factor of ~69%. The SPAD structure is similar to the structure E (Figure 150) represented and fully characterized in (0) with a DCR of less than 2 KHz and a PDP of 20% @ 450 nm for an excess bias of 600 mV. A voltage controlled PMOS resistor *Mp* is used for passive quenching thus allowing more flexibility in setting the quench resistor value by means of the analog input *Vctrl*. For faster quenching and reset process, an active quenching is provided by the NMOS transistor *Mq* and an active reset is provided by the PMOS transistor *Mr*. *Mq* and *Mr* are respectively controlled by the 2 *Quench_control* Block outputs *Quench_ctrl* and *Reset_ctrl* through 2 level translators allowing the 1.8V level digital inputs to control the analog electronics operating in 3.3V.

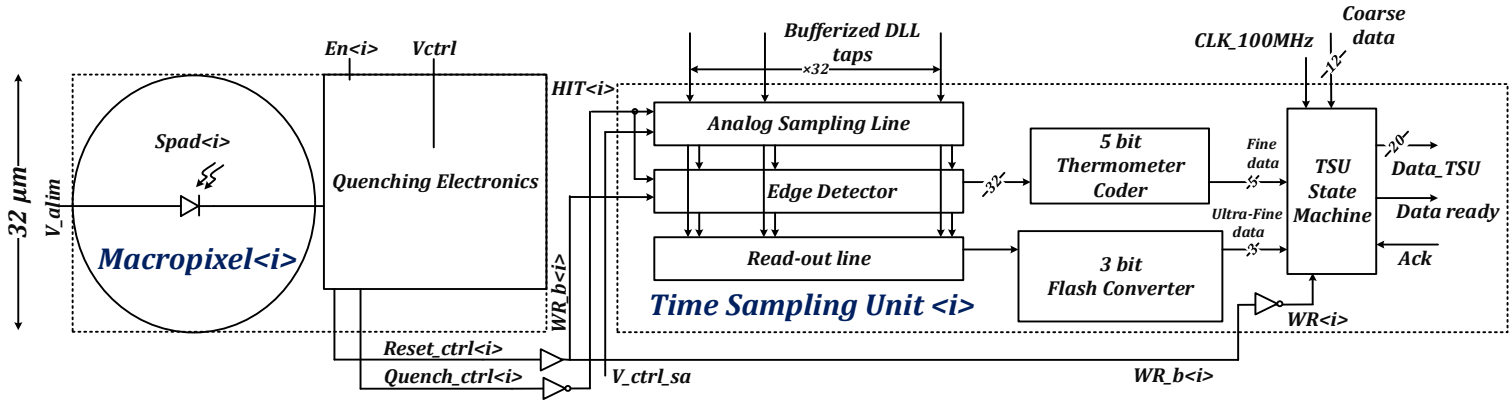


Figure 148 - Simplified bloc schematic of a TCSPC unit incorporating a Macropixel for photodetection with Time Sampling Unit for time measurement

The *Quench_control Block* is similar to the one used in the SPADs characterization circuit previously represented in (2.2.2), in this setup the quenching and reset mode is set to active quenching and reset and the quench and reset duration are fixed to 20 ns each. The three transistors *Mp Mq* and *Mr* operates with a 3.3 V power supply while the quench control block and the digital ports operates with a 1.8 V power supply and a group of 3 level translators are used as intermediaries between the quenching electronics and the *Quench control Block* operating with different voltage levels. Finally each Macropixel<i> can be individually controlled through the digital input *En<i>* which allows us to deactivate it if necessary.

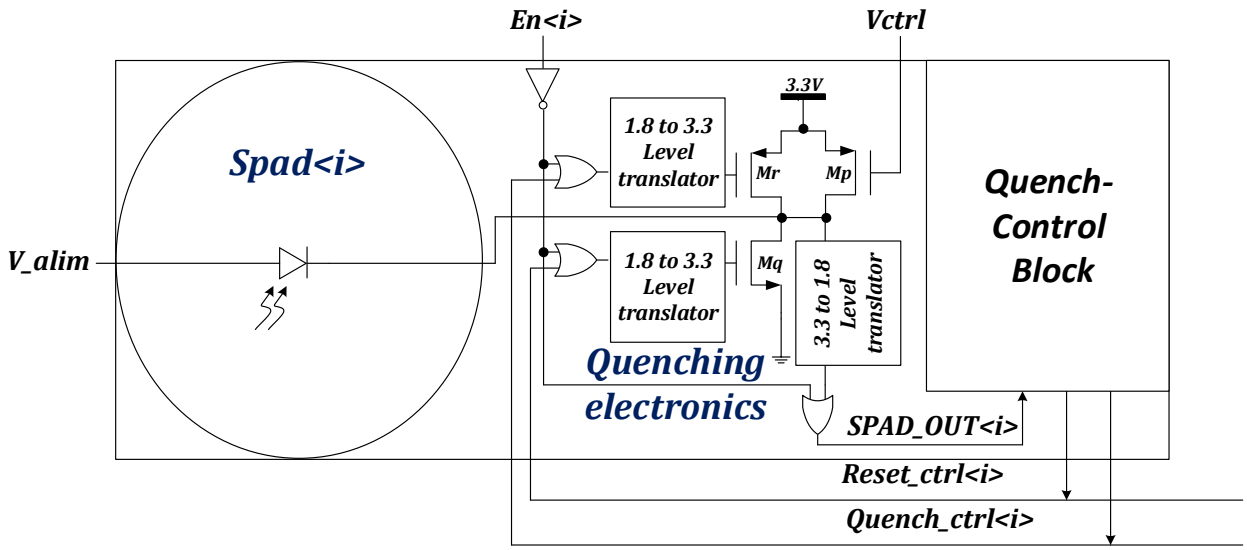


Figure 149 - Simplified bloc schematic of the Macropixel including a SPAD and the associated quenching electronics

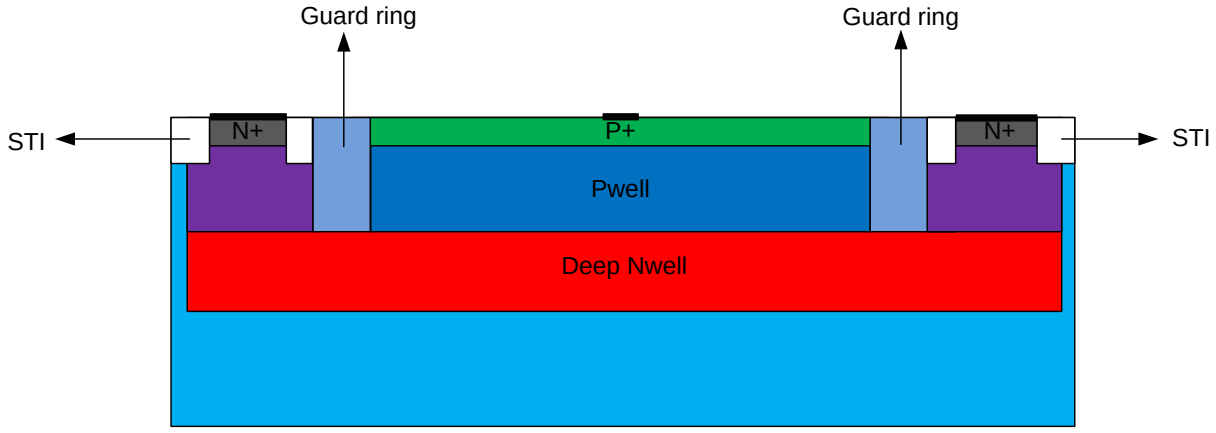


Figure 150 - Cross-section of the SPAD structure employed as a single photon detector in the Macropixel.

4.1.2 Operation Principle

Originally the *Quench_ctrl* signal is set to low and the *Reset_ctrl* signal is set to high and the 2 MOS transistors Mq and Mr are off , when a Photon detection occurs an avalanche ensues, The quenching process is started swiftly by means of the voltage controlled PMOS operating as a passive quenching resistor while the *Quench control Block* require a small waiting time before it detects the avalanche and reacts by activating the active quenching by means of the *Quench_ctrl* signal that goes from '0' to '1' setting *Mq* on for 20 ns then activating the active reset by setting *Mr* on for the following 20 ns by means of *Reset_ctrl* signal that goes from '1' to '0' thus resetting the SPAD to its original state before the photon detection with the *Reset_ctrl* signal set back to '1' again at the end of the process.

The *Quench_ctrl* and *Reset_ctrl* signals used to orchestrate the SPAD quenching process are also used to command the TSU block, with $\overline{Quench\_ctrl}$ used as a HIT signal indicating to the TSU that a detection event took place, *Reset_ctrl* is then used to stabilize the output of the edge detector while $\overline{Reset\_ctrl}$ is used as *Wr* signal signaling to the TSU state machine when it rises to '1'that the time conversion result is stable and ready to be transmitted to subsequent memory block. Giving that the time conversion requires a maximum of 15 ns to be completed and stabilized the 20 ns waiting time while *Quench_ctrl* rises to '1' is more than enough to ensure that the conversion output is stable and ready. As a result, the TSU state machine copies the 20 bit total time conversion result and its *data_ready* output rises to '1' indicating that a new valid set of data is present. Of course this data must be read in less than 40 ns to ensure that the possible subsequent data set is not lost.

4.2 Streak Architecture

The vast majority of TCSPC systems are regrouped in a form of pixel array where each pixel englobes a SPAD with its associated electronics and a TDC. The major problem with such setup is that it limits the achievable fill factor as well as the complexity of the associated electronics. These 2 limitations are avoided in the case of the TCSPC-SC structure thanks to the streak architecture that allows the integration of complex electronics in each unit without the price of lowering the fill factor as the photosensitive parts of the system are regrouped together instead of being scattered like in TCSPC arrays.

The block diagram of TCSPC-SC test structure is illustrated in Figure 151, the designed structure integrates a single line of Macropixels with a SPAD operating as a single photon detector in each unit. Each Macropixel is coupled with a time measurement unit based on the building block represented in (4.1). The 8 time measurement unit shares a common DLL and a 12 bit coarse counter also used to generate the 'START' signal triggering the laser source with a fixed repetition rate determined by the 12 bit input "*Laser_rate*". This setup allows a high time resolution over a wide dynamic range TCSPC measurements with reduced power

and area consumption owing to the TSU structure designed to allow the easy and compact assembly of a large number of TDCs. The devised architecture allows a fully parallel operation for all the units with the ensuing time measurement data processed through a shared data acquisition unit specifically designed to limit measurement loss without the need to use extremely high output rate as will demonstrate in 4.3.

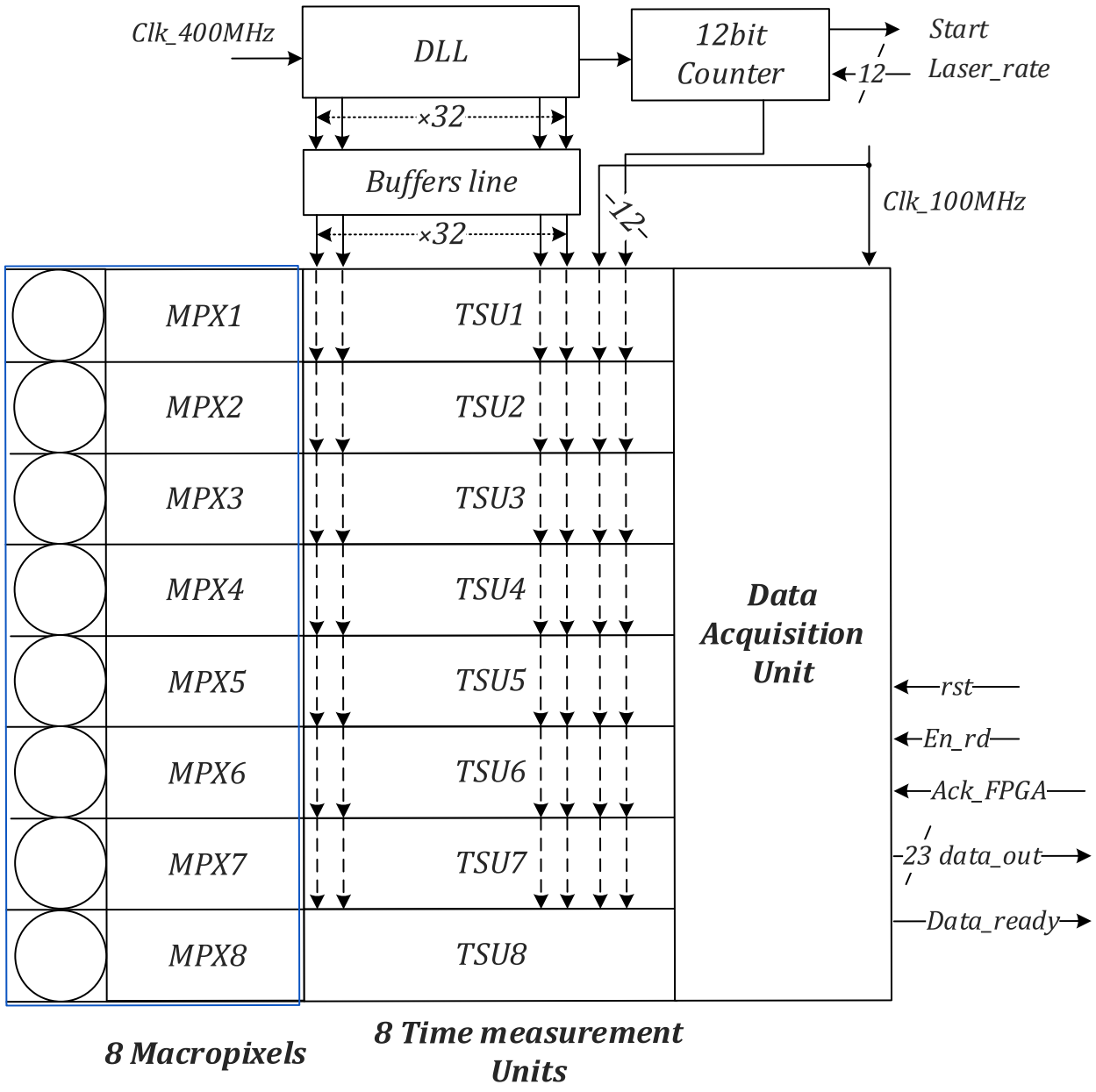


Figure 151 - Simplified bloc schematic of the designed TCSPC–SC test structure

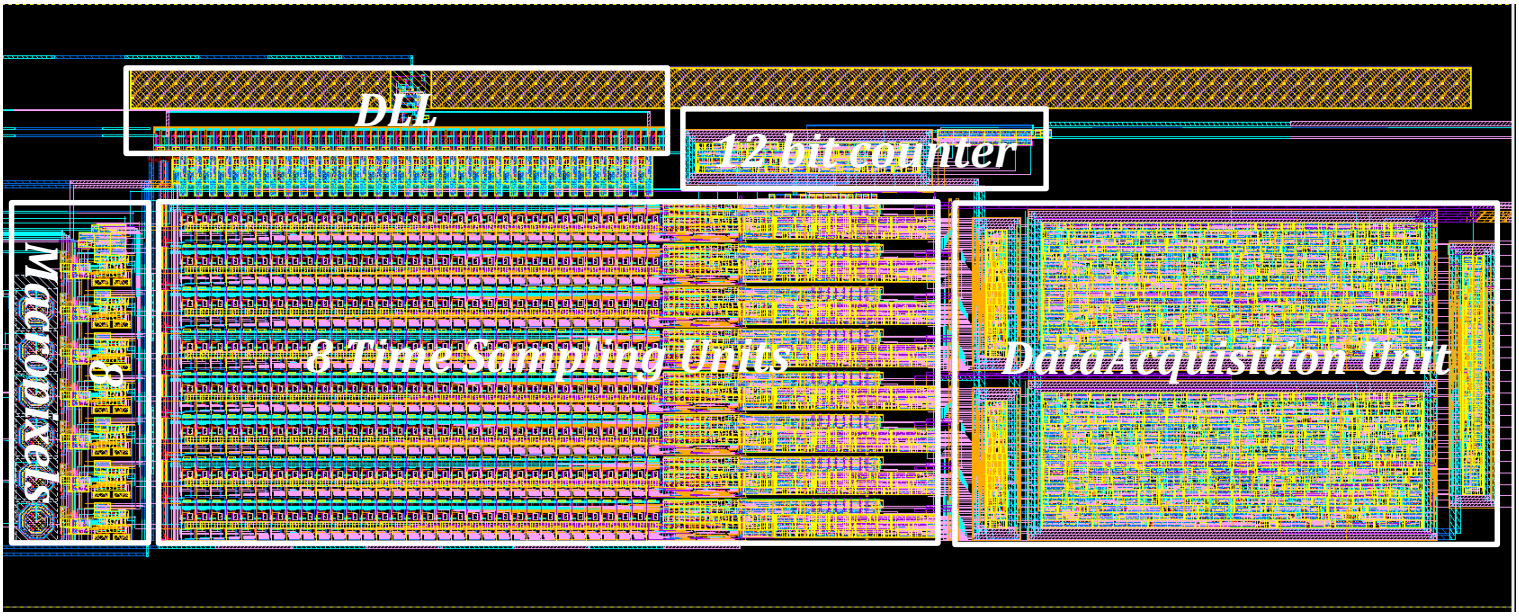


Figure 152 - Screen shot of the designed TCSPC–SC test structure layout

4.3 Data Acquisition Unit

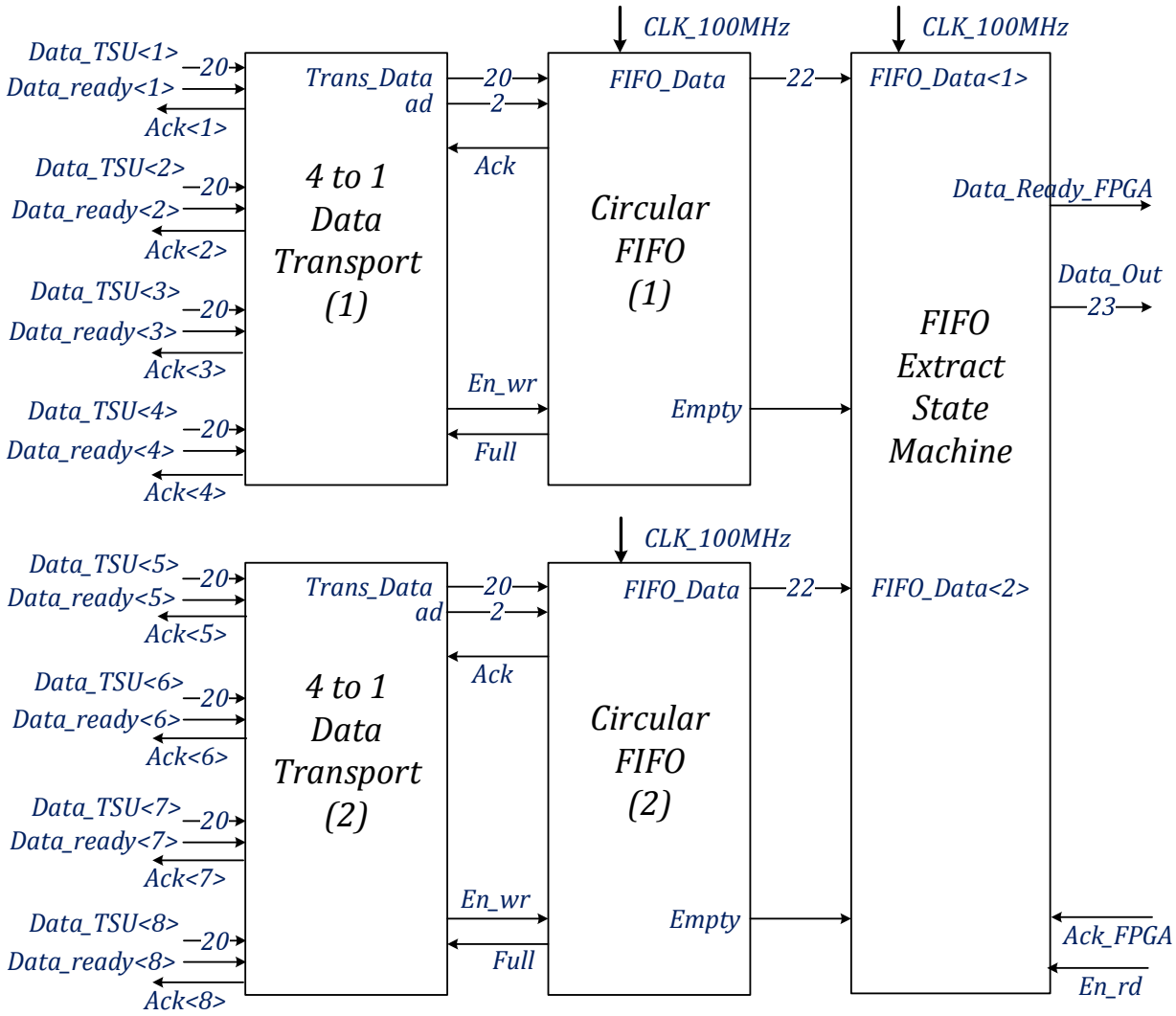
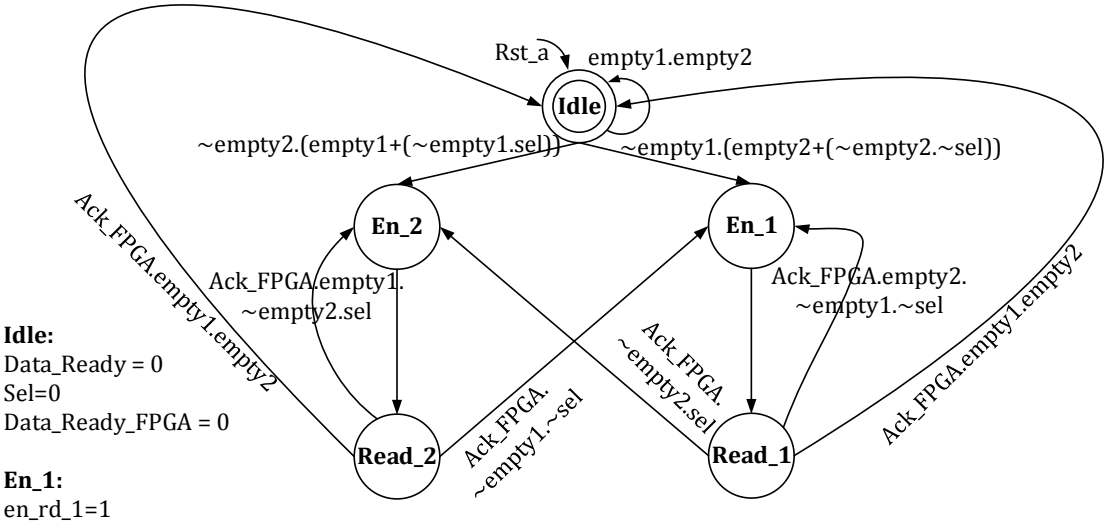


Figure 153 - Block diagram of the data acquisition unit

At the end of the time conversion process following a photon detection, the TSU delivers the result in the form of a 20 bit word, with a dead time of 40 ns this results in a data rate that could reach 500Mb/s per TSU and a global rate of 4Gb/s for the overall 8 TSCPC units. This requires a careful design of the Memory scheme to ensure a smooth and organized data transfer between the 2 blocks. Figure 153 shows a block diagram of the data acquisition unit devised to register and transfer the measurement data towards an external FPGA board. Each group of 4 TSUs is connected to a shared circular FIFO through a "4 to 1 Data Transport" combinatorial logic block working as an intermediary between the TSU state machine and its corresponding FIFO. Each TDC array is given a priority order so that in case of simultaneous HIT events in the same group, the data coming from the one with the highest priority order is processed first while the other wait for their turn. When the Time conversion process

following a detection event is complete, the TSU state machine signals it via its corresponding "*data_ready*" output and presents the resultant data. Depending on its priority and giving that the circular FIFO is not full, the "*4 to 1 data transport*" block will copy the DATA_TSU output, deliver it to the circular FIFO along with an additional 2 bits address referring to the data source and free the TSU state machine via its *Ack<i>* signal. While the data is continuously being recorded, the 2 circular FIFOs will communicate it to an external FPGA board via the "*FIFO Extract State Machine*" block which alternates reading between the 2 FIFOs as long as they are not empty (Figure 154). The state machine is connected to a 100 MHz clock input and requires 3 clock cycles to finish the transfer which ensure that the data outputted by each TSU would be read before it's substituted by a new set of data.



Idle:
Data_Ready = 0
Sel=0
Data_Ready_FPGA = 0

En_1:
en_rd_1=1
en_rd_2=0
Data_Ready_FPGA = 0
data_out= data_1 & 0

Read_1:
Sel=1
en_rd_1=0
en_rd_2=0
Data_Ready_FPGA = 1
data_out= data_1 & 0

En_2:
en_rd_1=0
en_rd_2=1
Data_Ready_FPGA = 0
data_out= data_2 & 1

Read_2:
Sel=0
en_rd_1=0
en_rd_2=1
Data_Ready_FPGA = 1
data_out= data_2 & 1

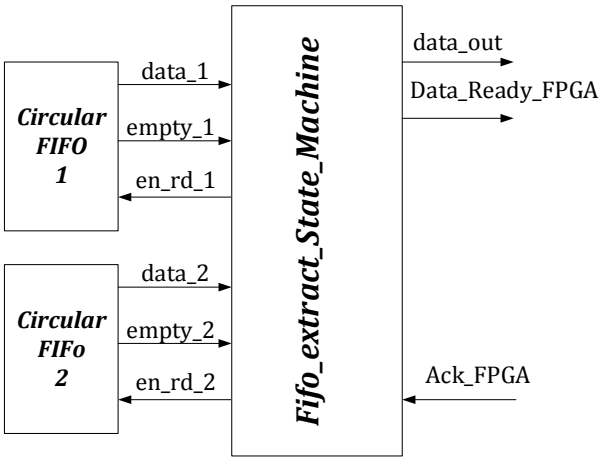


Figure 154 - State transition diagram of the data transmission scheme via the FIFO extract state machine

4.4 High Throughput Time Correlated Single Photon Counting

In [137] a commercial single SPAD, with specifications comparable to those of the SPADs integrated in the TCSPC-SC, was used in combination to detect Time Resolved Fluorescence from one single excitation spot on a microfluidic channel. The measurements showed that up to a few 1000 photons may be detected by time-correlated single photon counting (TCSPC) in individual droplets circulating through the excitation spot at a rate of one thousand droplets per second, and that a fluorescence lifetime may be inferred with sub-100 ps time resolution. This demonstrates the potentiality offered by a 1D array of such detectors to be used in applications such as FLIM and HTS which is why the TCSPC – SC was specially designed to achieve an optimum efficiency. This was made possible by two means: first by devising a scheme allowing the optimization of the SNR to ensure it will not be greatly affected by undesirable effects such as an abnormal DCR or uneven illumination. Second, the data acquisition scheme was devised to maximize the system's efficiency in dealing with the random photon detection. In this chapter we will present and discuss these two schemes based on mathematical models and Matlab simulations.

4.4.1 SNR Optimization

One feature in the Macropixel is that it is possible to activate it or deactivate if necessary. This option was added to ensure that the SNR is not affected by an undesirable effect that could decrease the detector's efficacy. Indeed, the signal delivered by a photon counting detector is affected by temporal fluctuations that are expressed as a Poisson distribution. If N is the average number of detected pulse, it includes a fluctuation expressed in the shot noise $n = \sqrt{N}$ while the other electronic noise can be ignored thanks to the infinity gain of the SPAD. The total signal N is given by $N=N_{ph}+N_d$ where N_{ph} is the total of detected photon and N_d is the number of counts caused by the dark count. The associated shot noises are $n_{ph} = \sqrt{N_{ph}}$ and $n_d = \sqrt{N_d}$.

The number of photons is measured by subtracting the results of two measurements: one for the total number of counts $N_{ph}+N_d$ and the second for the dark ones N_d. In this case, the total noise is given by:

$$n_{tot} = \sqrt{N_{ph} + 2N_d} \tag{38}$$

If N_d is considered as a constant equal to the mean value $\overline{N_d}$ instead of being measured each time, the variance of the term comes to zero and thus, the number of photons and its associated noise are given by:

$$N_{ph} = N_{tot} - \overline{N_d} \tag{39}$$

$$n_{tot} = \sqrt{N_{ph} + \overline{N_d}} \tag{40}$$

Therefore the signal to noise ratio is:

$$SNR = \frac{N_{ph}}{n_{tot}} = \frac{N_{ph}}{\sqrt{N_{ph} + \overline{N_d}}} \tag{41}$$

Let N_{phi} and N_{di} be the number of detected photons and the dark count rate of the ith SPAD (SPADi) in the ith Macropixel, the SNR of the TCSPC_SC structure is then given by the sum of each SPAD photon count divided by the total noise component:

$$SNR = \frac{\sum_i^8 N_{ph,i}}{\sqrt{\sum_i^8 (N_{ph,i} + N_{d,i})}} \tag{42}$$

And thus by switching the $SPAD_i$ on/off the signal to noise ratio can be optimized according to a specific algorithm.

A. Hot pixel elimination algorithm

By turning off all the SPADs in the detector except one, its individual DCR Nd_i can be measured in the dark. These measurements can be carried out in a phase of calibration of the sensor. Then, the idea of the hot pixel elimination (HPE) algorithm is to disable the SPAD that present a high DCR. In order to evaluate the benefit of the HPE algorithm, we assume that the Macropixels are uniformly lighted, i.e. all the Nph_i are equal to Nph, and all the SPAD's DCR are equal to Nd except for one $SPAD_j$ that presents a DCR m times higher than the rest of the SPADs. Thus, the signal to noise ratio is given by:

$$SNR_{(7+m)} = \frac{8.\overline{Nph}}{\sqrt{8.\overline{Nph} + (7+m).\overline{Nd}}} \tag{43}$$

By turning off the noisy SPAD, the SNR becomes:

$$SNR_{(7)} = \frac{7.\overline{Nph}}{\sqrt{7.\overline{Nph} + (7).\overline{Nd}}} \tag{44}$$

Consequently, disabling the noisy SPAD leads to a signal to noise improvement of:

$$\frac{SNR_{(7+m)}}{SNR_{(7)}} = \sqrt{\frac{7}{8}\left[1 + \frac{(m-1)}{8(\alpha+1)}\right]} \quad (45)$$

With $\alpha = \overline{N_{ph}}/\overline{N_d}$ is the mean photon count on the mean DCR ratio.

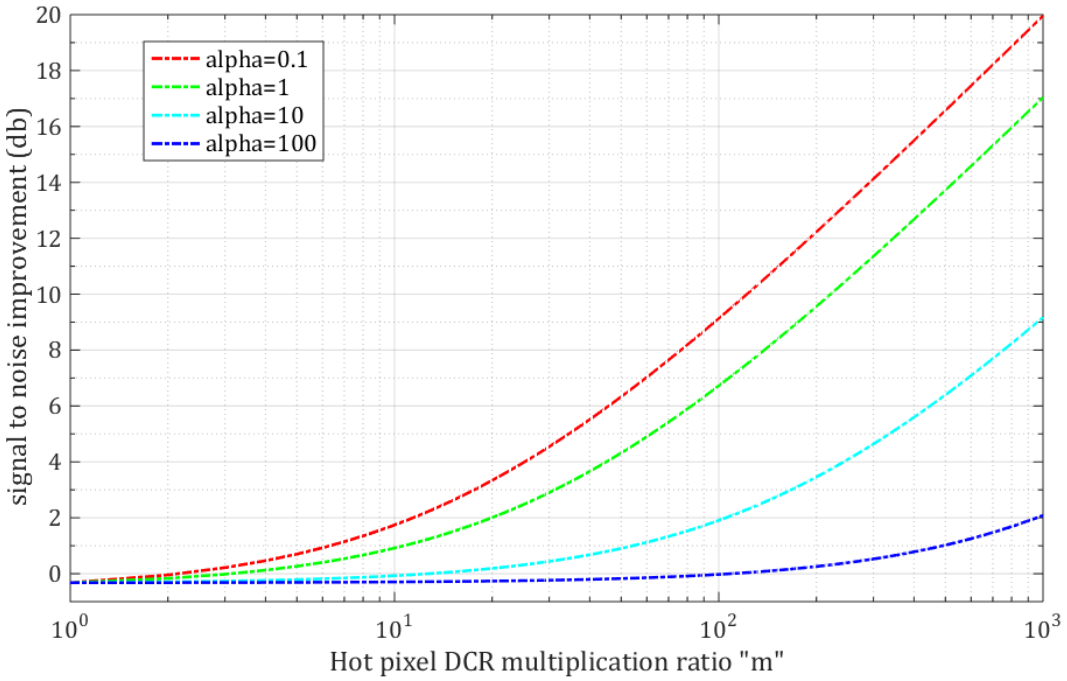


Figure 155 - Signal to noise ratio improvement obtained by a hot pixel elimination.

Figure 155 shows the SNR gain versus the hot pixel DCR multiplication *m* for different α ratio. For a weak signal measurement (α=0.1), the gain can be as high as 20 dB. Nevertheless, this assessment clearly states that the SNR may be slightly lowered if the *m* coefficient is too low. Thus it is not pertinent to remove for example all the SPADs which have a DCR above the mean DCR. As an efficient rule of thumb: only the SPADs Presenting an *m* coefficient above the α one, with obviously *m*>1, have to be disabled. Previous works report that about 20% of the SPADs integrated in an array have a dark count about 10 to 1000 times higher than the 80% other diodes [138] [66]. Consequently, there is a high probability to have a hot SPAD among the 8 SPADs. Therefore, the proposed TCSPC-SC architecture can leads to a significant SNR improvement ranging from 0 to 20 dB.

B. Dark pixel elimination algorithm

In the scenario where the sensibility of some SPADs was reduced, due to a manufacturing defect, a dust or simply because the SPADs are not uniformly lightened, the SNR ratio will also be affected. In order to evaluate the SNR gain, we assume the worst case, i.e. the best case for SNR improvement, where n SPADs are completely blind. Thus, the SNR is:

$$SNR_{(8-n_{blinded})} = \frac{(8-n).\overline{Nph}}{\sqrt{(8-n).\overline{Nph} + 8.\overline{Nd}}} \quad (46)$$

If all blinded SPADs are turned off, the SNR becomes:

$$SNR_{(8-n_{blinded})} = \frac{(8-n).\overline{Nph}}{\sqrt{(8-n).\overline{Nph} + (8-n).\overline{Nd}}} \quad (47)$$

Consequently, for $n \neq 8$, the SNR gain is given by:

$$\frac{SNR_{(8-n_{blinded})}}{SNR_{(8-n_{off})}} = \sqrt{\frac{(8-n)\alpha + 8}{(8-n)\alpha + (8-n)}} \quad (48)$$

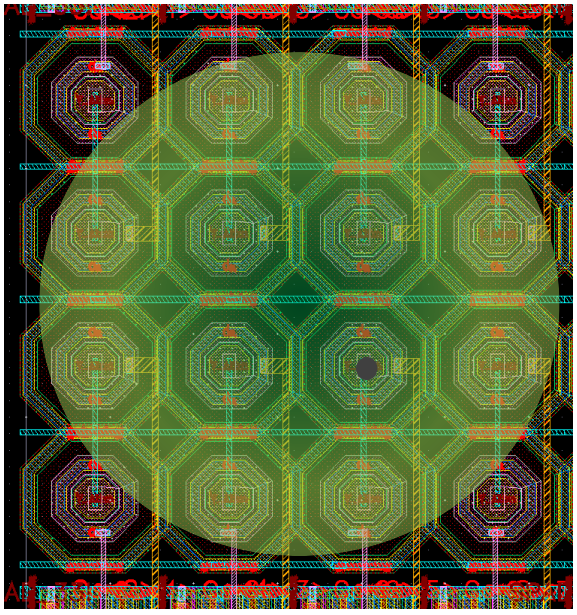


Figure 156 - Illustration of a 4×4 SPAD array with dark pixels as a result of irregular illumination or the signal being blocked due to the presence of an impurity

Figure 157 shows the signal to noise gain versus the mean photon count rate to the dark count rate ratio α. We clearly see that the SNR improvement rapidly decrease with α. That indicates that this correction is useless for a photon to dark count rate ratio above 10. For weaker signal, the gain is more important and increases with the number of SPADs turned off. Nevertheless, it is limited to $\sqrt{8/(8-n)}$ in the best case. In order to use this feature, one must measure the signal of each SPAD individually during illumination. This phase of calibration must be carried out at each measurement, explicitly if the measurement conditions change. This constraint must be placed within sight of the low SNR improvement. The relatively low efficiency of this algorithm makes it useful only for very critical and time constrained measurements.

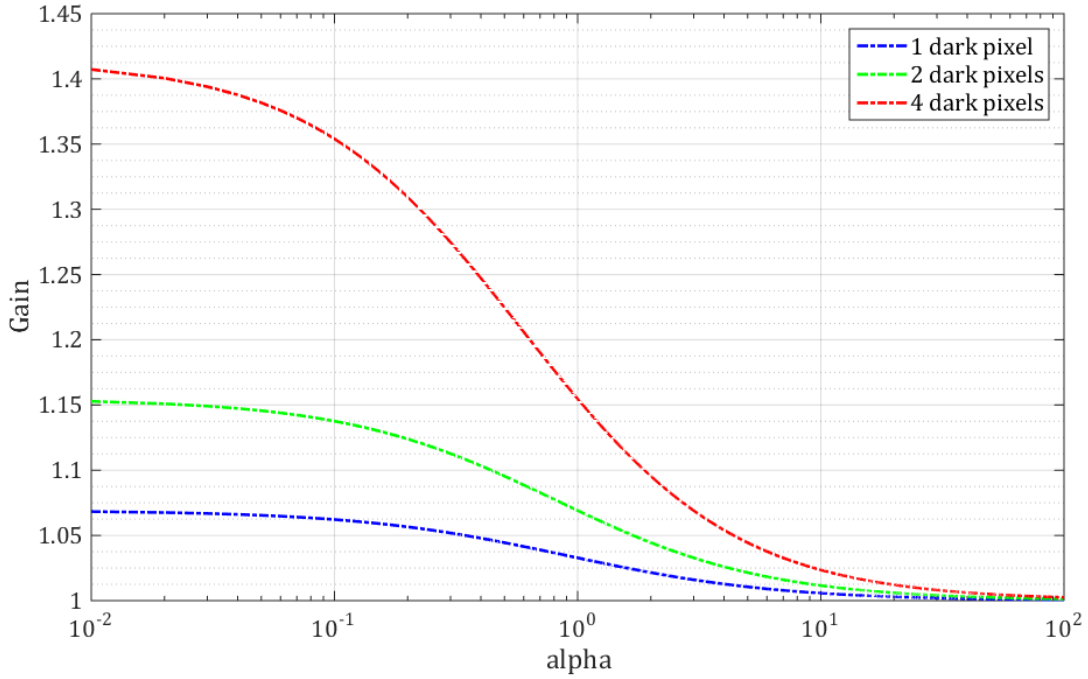


Figure 157 - Signal to noise ratio improvement obtained by some dark pixel elimination.

4.4.2 Efficiency Improvement

TCSPC consists of detecting single photon events, measuring their arrival times within a period of the light signal and reconstructing the temporal profile of the signal after repeating the measurements for enough times. Applications such as FLIM and HTS usually deal with a high counting rate which requires the use of multichannel systems [45] [139]in order to limit the counting loss, these systems usually involve a high output data rate and as a result high output frequencies are used [45]. In this section we study the impact of integrating a FIFO to limit the counting loss in high rate TCSPC systems and increase the system efficiency without the need to use very high output frequencies. The efficiency of this scheme is assessed according to the FIFO depth for a giving photon rate to output rate ratio using a queueing system model.

4.4.2.1 Counting loss in a TCSPC system

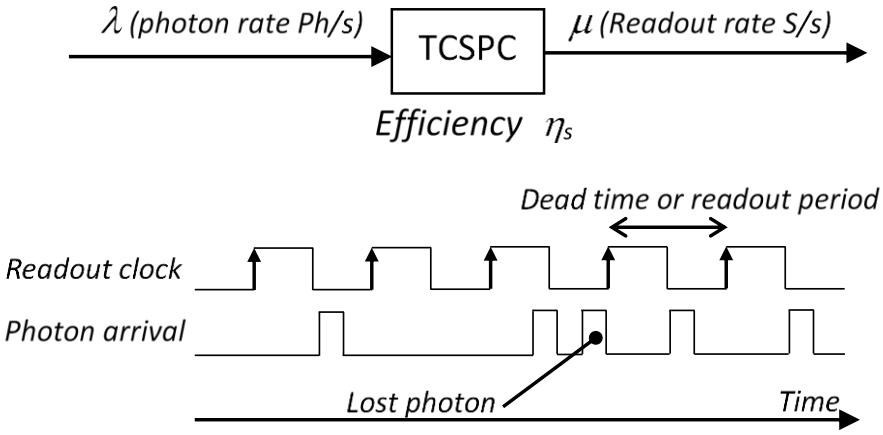


Figure 158 - Counting loss in a TCSPC model

Typical TCSPC setup includes a pulsed optical laser source, a discrete detector such as a silicon photon multiplier (SiPM) or a SPAD, a time to digital converter (TDC) or time to amplitude converters (TAC) for time measurements and an external CPU to compute the measurement results. When a photon is collected and detected a certain time is required for data processing; such time is referred to as a dead time. During the dead time the system is blind and a photon collected by the SPAD in this interval can't be detected by the circuitry resulting in a counting loss (Figure 158), and a loss of SNR as a result of the decreased counting efficiency which is at best equal to:

$$CV = \frac{1}{\sqrt{n}}. \quad (49)$$

In order to improve the signal quality many parallelized structure and imaging array have been reported to increase the rate of detected photons, however this trend lead to an

increased data bottleneck requiring complex readout circuitry [66] and the use of very high output frequencies in order to keep a reasonable dead time [45]. Another solution for the high output data is the use of an embedded FIFO to store the measurement results while they are been read, however FIFOs are very demanding in terms of surface and power and to our knowledge there has been no study done to determine the exact FIFO length required for optimum results.

4.4.2.2 TCSPC system as a queuing model

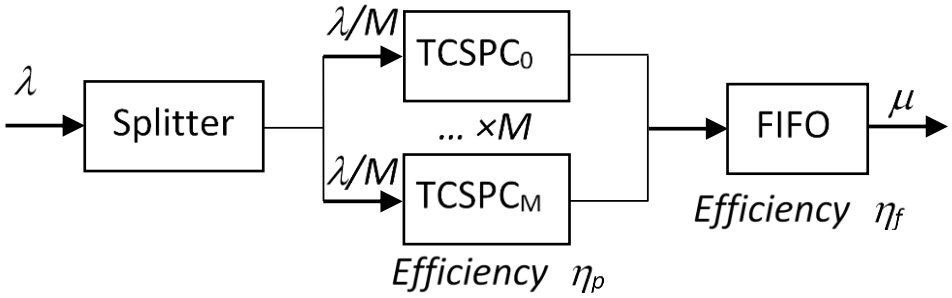


Figure 159 - Parallelization principle of several TCSPC modules

TCSPC systems are based on measuring arrival times of single photon events. Processing such a measurement required several operation steps such as quenching the photon detector, shaping the regenerated signal, converting the time to a digital value and sending this last into a process unit or memory. This overall operation leads to a dead time for which the system is unavailable for another measurement. To simplify the explanation, we assume that the readout period of the system is equal to its dead time. The stochastic nature of the single photon detection process along with the system's dead time limit the system efficiency due to the counting loss because the system is busy processing a previous arrival. The assessment of this loss, evaluated in this section, can be carried out by modeling the TCSPC system as a queuing system with an arrival rate λ, i.e. the average number of photon interacting with the sensor per second, and a departure rate μ, i.e. the readout data rate given in sample per second. Figure 158 illustrates this phenomenon: even if the arrival rate λ is equal or less than the departure rate μ, the random process of the arrival of photon leads to a quiet period followed by a peak of arrivals of photon which is a well know characteristic of a Poisson process. During this peak of activity, the system can lose some photon. A first approach to limit this loss is to reduce both the dead time and the readout period. Obviously reducing these times is limited by physical and electrical constraints to tens of nanosecond Another approach is to use a structure with parallel modules in which the incoming light is uniformly split (Figure 159). Thus, assuming an equal distribution of the photon arrivals this is similar to dividing the arrival rate λ into M equal parts, where M is the number of parallel structures. The counting loss, in addition to the pile up effect, is decreased for each single TCSPC module. However this approach creates a data bottleneck at the end of

the processing chain as a result the use of high data rate bus to process the resultant high counting rate is still necessary. Indeed, the loss problem is only shifted to the overall output which is limited to the data rate μ. This may be avoided by integrating a FIFO in the TCSPC system which allows a better flexibility in processing the stochastic arrival events. A TSCPC system without a FIFO can be modeled as a one buffer queuing system, similarly a TCSPC system integrating a FIFO with a depth of N cells can be modeled as an N cell queuing system. We will assume that the FIFO's input data follows a Poisson process, a plausible assumption when the average Photon arrival rate is significantly lower than the TCSPC's operating frequency. Giving the stochastic nature of the measured phenomena i.e. the photon arrival Poisson process, it is important to study the system's behavior in terms of the traffic intensity in and out of the FIFO in order to determine the impact of its limited capacity on the sensor's sensibility due to missed arrivals when the FIFO is full. The FIFO can be described as a size N queuing system where the input is a Poisson arrival process with a mean arrival rate λ and the probability function of n arrivals occurring during the time interval [t ,t+τ]:

$$P[N(t+\tau)-N(t)=n]=\frac{e^{-\lambda\tau}.(\lambda\tau)^n}{n!} \qquad n=0.1.2... \qquad (50)$$

The FIFO's output follows a periodic departure process, the departure rate is equal to μ and the time needed for one departure to be finished is $T_d = \mu^{-1}$ is the readout period. The system can be modeled as a semi-Markov chain with $Q_n=Q(t=t_n)$ is the number of occupied cells in the FIFO immediately after departure moments $\{t_n, n=0,1,2...\}$ [140]. Since the FIFO's capacity is limited to N cells, the number of occupied cells in the system cannot exceed N-1 and the embedded Markov chain contains N states labeled according to the number of occupied cells left soon after a departure $S=\{n , n=0,1,2...N-1\}$. Figure 160 shows the embedded Markov chain with all the possible transitions from a random state 'i'.

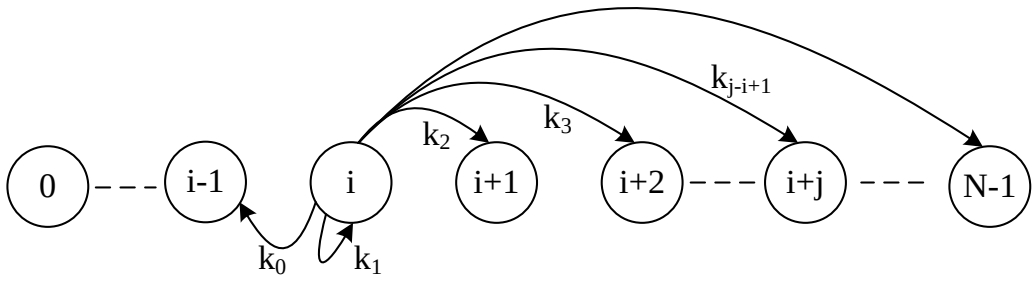


Figure 160 - Markov chain states and possible state transitions from and into a state 'i'

Steady State Probabilities evaluation

Let X_n be the number of arrivals during the readout period T_d giving the Poisson arrival property, the probability of j arrivals occurring during the readout period is:

$$k_j = P[X_n=j] = \frac{e^{-\rho}.(\rho)^j}{j!} \qquad j=0.1.2... \tag{51}$$

Where ρ defined as :

$$\rho = \lambda.T_d = {}^{\lambda}/_{\mu} \tag{52}$$

represents the photon rate to the readout rate ratio. The number of occupied cells after the *n+1th* period is increased by the number X_{n+1} of photon arrivals during this period and is reduced by one readout. If the number of photon arrivals overloads the FIFO, the number of occupied cells is clipped to *N-1* and a loss of measurement occurs. If the FIFO is empty, i.e. $Q_n = 0$, no readout occurs. Therefore, the relation between Q_n and Q_{n+1} is defined as:

$$\begin{gathered} Q_{n+1} = min(Q_n + X_{n+1} - 1\,, N-1)\; if\, Q_n > 0 \\ min(X_{n+1}\,, N-1)\; if\, Q_n = 0 \end{gathered} \tag{53}$$

And the transition probability from the state *i* to the state *j* after *m* transitions is:

$$P_{ij}^{(m)} = P(Q_{n+m} = j / Q_n = i) \qquad i, j \in S \tag{54}$$

In particular the 1-step transition probability is:

$$\begin{aligned} &P_{ij}^{(1)} = P_{ij} = P(Q_{n+1} = j / Q_n = i) \\ &P_{ij}^{(1)} = P(i + X_{n+1} - 1 = j) \qquad if\; i > 0 \\ &P(X_{n+1} = j) \qquad if\; i = 0 \end{aligned} \tag{55}$$

Which allows us to define the *K×K* transition probability Matrix *'P'* of the 1-step transition probabilities $P_{i,j}$ [140]:

$$P = \begin{bmatrix} k_0 & k_1 & k_2 & \dots & k_{N-2} & 1 - \sum_{i=0}^{N-2} k_i \\ k_0 & k_1 & k_2 & \dots & k_{N-2} & 1 - \sum_{i=0}^{N-2} k_i \\ 0 & k_0 & k_1 & \dots & k_{N-3} & 1 - \sum_{i=0}^{N-3} k_i \\ \vdots & \vdots & & & & \dots \\ 0 & 0 & 0 & \dots & k_0 & \dots 1 - k_0 \end{bmatrix} \quad (56)$$

Where the *i,j* element $P_{i,j}$ of the matrix represents the probability of being in the state '*j*' giving that the system was in the state '*i*'. These probabilities describe the transient behavior of the system, however as the system evolve it will converge into a state of equilibrium known as the steady state with time independent distribution [130] represented as a vector $\pi = (\pi_0, \pi_1, \pi_2 \dots \pi_{N-1},)$ where π_i is the probability to be in the state *'i'* once the system has reached its equilibrium.

The steady state distributions satisfy the following equations:

$$\pi(j) = \sum_{i=0}^{N-1} \pi(i).P(i,j) \quad (57)$$

And

$$\sum_{i=0}^{N-1} \pi(i) = 1 \quad (58)$$

Furthermore the vector π is the solution to the set of the linear equation [141]:

$$\pi.P = \pi \quad (59)$$

Resulting in a system of N equations with N variables π_i:

$$\begin{gathered} k_0.\pi_0 + k_0.\pi_1 = \pi_0 \\ k_1.\pi_0 + k_1.\pi_1 + k_0.\pi_2 = \pi_1 \\ k_2.\pi_0 + k_2.\pi_1 + k_1.\pi_2 + k_0.\pi_3 = \pi_2 \\ \dots \end{gathered} \quad (60)$$

Blocking probability

The main goal of our study is to evaluate the system efficiency based on the probability of an arrival finding the FIFO full and as a result being lost, such a probability is referred as the blocking probability P_B. In order to evaluate P_B we need to have the state distribution at all time and not only at departure moments. Let us define the following system probabilities:

P_k: Probability of the system containing k registered arrivals *(k=0…N)*
π_k: State probabilities at departure instants *(k=0…N-1)*
$\pi_{a,k}$: State probabilities at arrival instants regardless whether the arrival join the queue or not *(k=0…N)*

An important property of the Poisson arrival process is the Poisson Arrival See Time Averages [140] which implies that the distribution of occupied cells seen at arrival instants is the same as the distribution seen by a random observer:

$$\Rightarrow P_k = \pi_{a,k} \tag{61}$$

On the other hand the probability that an arrival finds $k<N$ occupied queue in the system is equal to the probability that a departure leaves k occupied cell giving that the new arrival is admitted

$$\Rightarrow P_k = \pi_k(1-P_B) \tag{62}$$

In particular for k=0 we have:

$$P_0 = \pi_0(1-P_B) \tag{63}$$

Furthermore arrivals entering the system occur at rate λ as long as they are admitted into the queue and we define the effective arrival rate as:

$$\lambda_e = \lambda(1 - P_B) \tag{64}$$

Simultaneously departures out of the system continue to occur with a rate μ as long as the system is not empty which allows us to define the effective departure rate as:

$$\mu_e = \mu(1 - P_0) \tag{65}$$

Given that in equilibrium the traffic entering the queue system is equal to the one leaving the queue [66], we have:

$$\lambda(1 - P_B) = \mu[1 - \pi_0(1 - P_B)] \tag{66}$$

And the blocking probability is:

$$P_B = \frac{1}{\rho + \pi_0} \tag{67}$$

The described method was used to determine the blocking probability and the system efficiency η:

$$\eta = 1 - P_B = \frac{\pi_0}{\rho + \pi_0} \tag{68}$$

Where π_0 is defined in (60).

Figure 161 shows the system efficiency using a buffer and FIFO with *N=2, 4, 8, 16*. It is obvious that the system efficiency increases with bigger FIFO depth however the amount of augmentation decreases and giving the resources needed for a FIFO we can say that a FIFO depth of 8 is enough to reduce the arrivals input loss due to the blocking phenomenon.

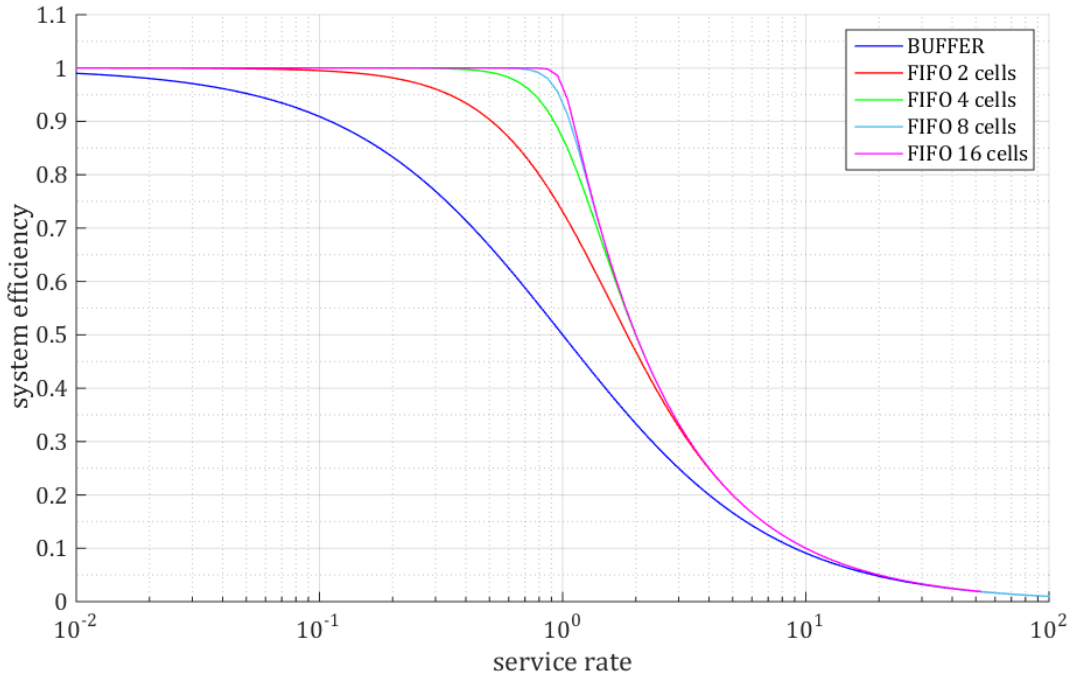


Figure 161 - Simulation results of the system efficiency for a FIFO depth of 2, 4, 8 and 16 cells

The TCSPC system illustrated in Figure 151 was especially designed to be used for an HTS application that requires counting rates up to several MHz per channel. With a TDC dead time of 40 ns, the maximum data rate is equal to 25 MS/s. According to Figure 161, the use of a unique TCSPC module would lead to an efficiency η of respectively 98, 90 and 50% for a photon rate of 0.25, 2.5 and 25 MHz, i.e. a service rate of 0.01, 0.1 and 1. Obviously, for a service rate ρ>1, the system's efficiency would tend to 1/ρ regardless of the use of a FIFO. A photon rate of λ=25 Mega photons/s is therefore not reasonable in the configuration of a single TCSPC module but if the arrival rate is divided among the 8 TCSPC (Figure 162) and assuming the arrival process is equally distributed among the 8 units, each $TCSPS_i$ receives an arrival rate:

$$\lambda_i = \frac{\lambda}{8} \tag{69}$$

Resulting in a service rate ρ_i =0.0125 and an efficiency $\eta_{ph} = 90\%$ i.e. an expected departure rate TCSPC $\mu_{TCSPCi} = 2.8\ MHz$ out of each TCSPC unit which is similar to the value obtained in [137].

As discussed in (4.3), the data acquisition scheme ensures the admission of all the data presented by the 8 TCSPC modules into the FIFO which transfers the measurements to an external FPGA board through the FIFO extract block.

Giving the low service rate of each TCSPCi, the output of each TCSPC unit will have a distribution very similar to the Poisson process and the resulting process is the sum of 8 Poisson process with their respective arrival rate $\lambda_i, i = 1,2, \dots 8$ and is therefore also a Poisson process with an arrival rate:

$$\lambda = \sum_{i=1}^{8} \lambda_i = 8 \times 2.8\ MHz = 22.4\ MHz \tag{70}$$

Assuming an output frequency of only 33.33MHz, the service rate will be $\rho_f = 0.67$. In the absence of a FIFO the system can be assimilated to a buffer resulting in Memory Block efficiency $\eta_M = 0.6$ (Figure 161) and a total efficiency:

$$\eta_{no\ FIFO} = \eta_{ph} \times \eta_M = 0.9 \times 0.6 = 54\% \tag{71}$$

The efficiency of the system is therefore not improved by the parallelization of the TCSPC even with the reduction of the pile up effect. However, using the 8 FIFO cells leads to a memory block efficiency of $\eta_M \cong 100\%$, the overall TCSPC system efficiency is maintained at about 90%. Such efficiency level can only be achieved with a 3 GHz output frequency without the use of the FIFO which shows the significance of the FIFO solution.

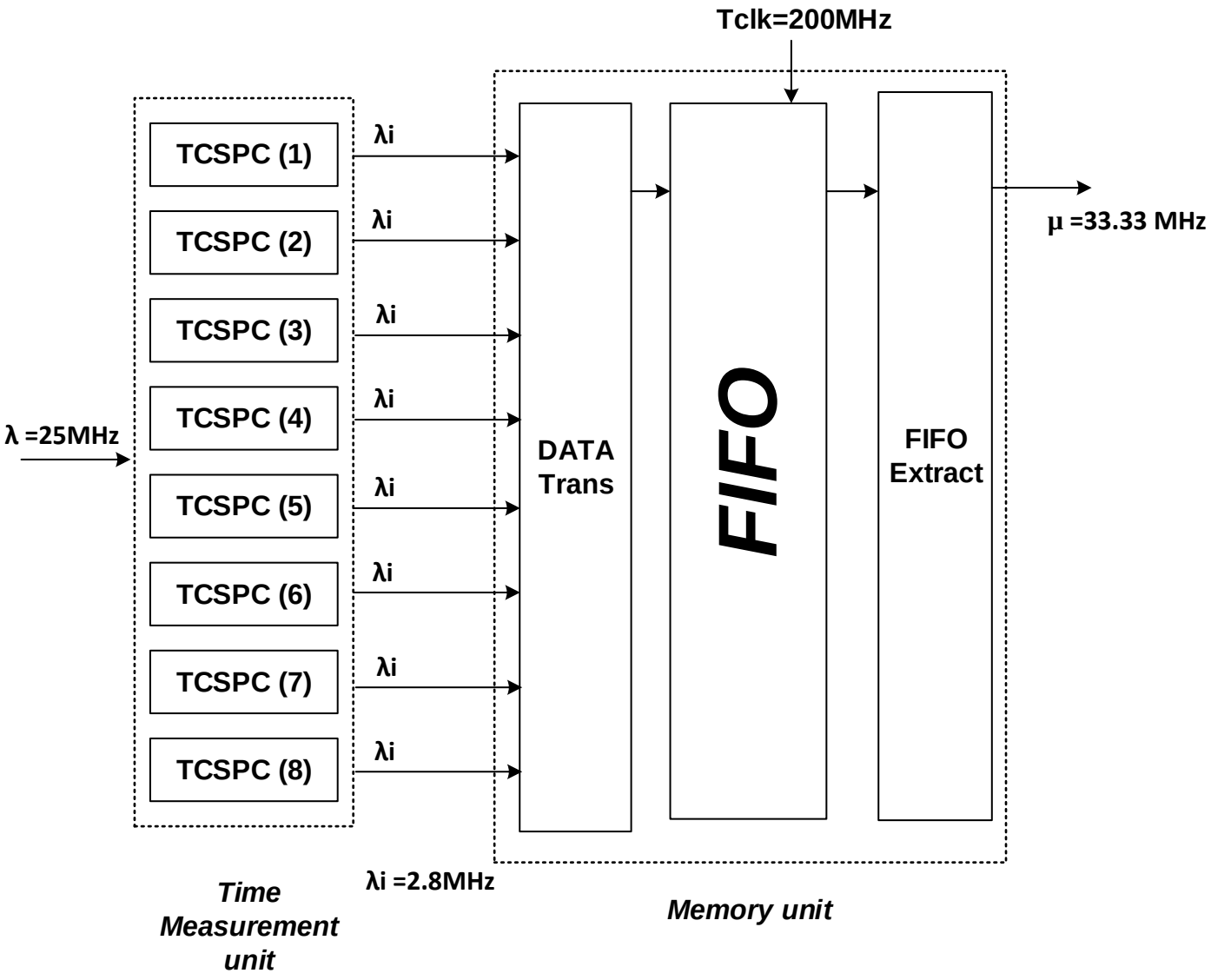


Figure 162 - Parallelization scheme of the high repetition rate TCSPC with the embedded FIFO

Chapter 5
Conclusion and Future Work

This thesis aimed to demonstrate a test structure of a sub-nanosecond temporal resolution integrated Streak Camera based on a TCSPC system, a task that was carried out in three phases. The first phase consisted of designing a Single Photon Avalanche Diode To in standard technology with good performance figures to be integrated in the TCSPC system. To do this, 6 SPAD Structures with 8 diameter variations were implemented in a 180 nm Standard CMOS technology. Following a preliminary characterization process, the most performing SPAD structure was selected and fully characterized in terms of noise, time resolution and sensitivity. The measurements showed a DCR below 5 KHz at 15°C with an afterpulsing probability of ~0.2%, a peak PDP of 20% around 430 nm and a timing jitter less than 100 ps FWHM for λ =450 nm at an excess bias of 500 mV measured with the SPAD featuring an active area diameter of 40 µm. Based on the characterization results, a SPAD with an active area diameter of 15 µm, a DCR of less than 2 KHz and a PDP of 20% @ 450 nm for an excess bias of 600 mV was selected to be used as the Single Photon detector in the TCSPC-SC. Next we demonstrated a hybrid TDC architecture especially designed to be used in a system that incorporates an array of TDCs which. This required a careful design to limit the power consumption and occupation and was achieved through a time conversion scheme that combines the traditional Analog Time to Amplitude Converter (TAC), Digital DLL-based and counter-based TDC techniques to overcome their limitations and achieve high adjustable time resolutions over a large dynamic range. Based on this approach a TDC array englobing 8 time-measurement channels was realized, the 8 channels are fully independent and share a common DLL and coarse counter. The hybrid TDC was designed to operate with variable input frequencies, the maximal speed operation using a 400 MHz clock input yielded 10 ps LSB, an estimated INL of ~ 5.6 ps rms, a maximum dynamic range of 10 µs and a dead time of 15 ns. The maximum estimated power consumption was 44 mW for a 20 MHz conversion rate with the global DLL consuming 35 mW resulting in 9 mW power consumption per TDC channel resulting in an important reduction in the total power consumption of the TDC array. Following the design of the two main building blocks of a TCSPC system, a TCSPC-SC test structure was realized in a 180 nm standard CMOS technology. The final circuit integrates a single line of Macropixels with a SPAD operating as a single photon detector in each unit. Each Macropixel is coupled with a time measurement unit based hybrid TDC architecture The

devised architecture allows a fully parallel operation for all the units with the ensuing time measurement data processed through a shared data acquisition unit specifically designed to limit measurement loss without the need to use extremely high output rate.

While many 2D-TCSPC system have been demonstrated in recent years, the 1D-TCSPC structure is still a concept that hasn't been fully explored and only a limited number of linear structure have been reported in the literature [63]. In Table 9, the main characteristics of the TCSPC-SC structure are compared to those of linear SPAD arrays found in the literature. The majority of these systems rely on external processing and only [142] and [63] are complete TCSPC systems. In [142] a 60 × 1 linear array of single SPADs with 60 integrated time-to-digital converters (TDCs) was presented. The SPADs have a 100 µm active area diameter with 52% fill factor, high photon detection efficiency of 50% at 420 nm, and a dark count rate of 2.5 kcps at room temperature. The TDC has a time resolution of 250 ps, a single-shot precision of 200-ps rms, the DNL is 5% LSB rms and the INL = 30% LSB rms. [143] reports on a complete TCSPC system in the form of two separate chips, one with 32×1 SPAD detectors fabricated in custom technology and the other one with 32×1 channels of timing electronics fabricated in standard technology. The SPADs have a 50-µm diameter with only 15.7% fill factor and a dark count rate of 7.3 kcps. In [144], 2 large arrays with 1024×8 pixels each were designed for Raman Spectroscopy with a 1 bit gated counter integrated in each pixel. The first array showed a very low DCR of 6 cps, but also a very low fill-factor of 4.9% and a low PDE with only 6% peak value. The second array showed a higher fill-factor of 44.3%, but also a higher DCR with 5.7 kcps. In [145] a 64×1 SPADs array was presented for FLIM applications. The system provides only time gated counting and the SPADs has a 15.8 µm diameter, 34% fill-factor, 32% peak PDE and a DCR of 1k cps. Finally, in [146] an array of 32×2 SPADs was reported but with no on chip electronics. The SPADs had a 50 µm diameter and featured a 50% PDE and 150 cps DCR. Like [142] each SPAD in the TCSPC-SC presented in this document is combined with a TDC thus combining SPADs with good performance figures and high time resolution TDCs. The performance of our SPADs compares well with the other custom SPAD arrays in terms of fill factor and DCR with 69% fill factor and 2 Kcps and the hybrid TDCs achieve a far better timing performance.

The demonstrated TCSPC-SC was realized in a 180 nm standard CMOS technology with two objectives in mind. The first one was to demonstrate the potentiality offered by such 1D detector to be used in HTS and FLIM applications and in particular Resolved Fluorescence measurements through TCSPC in microfluidic droplets under high-throughput conditions. The demonstrated structure was especially adapted to this type of applications by means of associated smart pixel functionalities added to ensure good SNR and maximize the system's efficiency. The SNR improvement is achieved through an SNR optimization scheme that leads to a significant SNR improvement ranging from 0 to 20 dB. The systems efficiency on the other hand is improved through a parallelized architecture and an adapted data acquisition

scheme devised to limit the counting loss. In [46] the authors demonstrated the achievability of Time Resolved Fluorescence detection by TCSPC high throughput screening (HTS) of biomolecular interactions in microfluidic droplets. In this type of measurement, it is known that the count rate should not exceed about 1% of the excitation repetition rate in order to avoid distortions caused by the pile-up effect [147]. Assuming a similar experiment where a light laser diode pulsed at a repetition rate of 20 MHz was used to stimulate the droplet's fluorescence emission, the measurements set-up should be such that only 200 KHz could be detected by a SPAD. However by using an array of 16×1 SPADs with the TCSPC-SC architecture the count rate is increased up to 3.2 MHz without distorting the results and with 99% counting efficiency owing to the employed data acquisition scheme. Additionally, an acquisition period of 1 ms will result in a rise of 2000 to 32 000 detected photons, a DCR of 2 kcps per SPAD will thus lead to an SNR enhancement from 42 up to 151 mainly limited by the photon shot noise. This shows the benefit of using the devised structure in HTS TCSPC applications such as FLIM where long acquisition times are often required to compensate for poor statistics when keeping the count rate below the pile-up limit. The second objective of this project, which is still in progress at the time of writing this report, is to perform the characterization tests to validate the proposed architecture and do the necessary modifications to correct potential problems in order to devise a fully operating TCSPC-SC model.

Table 9 - Comparison of different linear SPAD arrays

	Number of pixels	Fill factor (%)	Pitch (µm)	DCR (cps)	Peak PDE (%)	Integrated electronics
This work	**8×1**	**69**	**32**	**2000**	**20**	**TDC (10ps)**
[148]	60×1	52	150	2.500	50	TDC (250 ps)
[142]	32×1	15.7	250	7.300	44	External processing
[143]	4×112		25	6	5,3	External processing
[144]	1024×8	4.9	24	80	6	Time gated counters
[144]	1024×8	44.3	24	5.700	23	Time gated counters
[145]	64×1	34	26	1000	32	Time gated counters
[146]	32×2	20	100	150	50	External processing

Bibliography

[1] H. W. Connell, "The Heape and Grylls Machine for High-speed Photography," *Journal of Instruments,* vol. 4, p. 82–87, 1926.

[2] R. Lambert, "A Rotating Drum Camera for Photographing Transient Phenomena," *Review,* vol. 8, p. 13–15, 1937.

[3] C. D. Miller.US Patent 2400887, 1946.

[4] B. Brixner, "One Million Frame per Second Camera," *Journal of the Optical Society of America,* vol. 45, no. 10, p. 876–880, 1955.

[5] "http://cordin.com," [Online].

[6] A. Bartolick and J. M. Frank, "Solid State Replacement of Rotating Mirror Cameras," *SPIE 27th International Congress on High-speed Photography and Photonics,* vol. 6279, 2007.

[7] D. Litwiller, "CCD vs. CMOS: Facts and Fiction," *Spectra,* 2001.

[8] S. Morrison, "A New Type of Photosensitive Junction Device," *Solid-State Electronics,* vol. 6, no. 5, p. 485–494, 1963.

[9] P. J. Noble, "Self-scanned Silicon Image Detector Arrays," *IEEE Transactions on Electron Devices,* vol. 15, no. 4, p. 202–209, 1968.

[10] G. F. Amelio, M. F. Tompsett and G. E. Smith., "Experimental Verification of the Charge-Coupled Semiconductor Device Concept," *The Bell System Technical Journal,* vol. 49, p. 593–600, 1970.

[11] W. S. Boyle and G. Smith,, "Charge-Coupled Semiconductor Devices," *The Bell System Technical Journal,* vol. 49, p. 587–593, 1970.

[12] G. E. Smith, "The Invention of the CCD," *Nuclear Instruments and Methods in Physics Section A,* no. 471, pp. 1-5, 2001.

[13] "Hamamatsu Learning Center, MOS Capacitor," [Online]. Available: http://hamamatsu.magnet.fsu.edu/articles/moscapacitor.html.

[14] A. Luštica, "CCD and CMOS image sensors in new HD cameras," *ELMAR 2011 Proceedings,* pp. 133-136, 2011.

[15] W. Uhring and M. Zaltanski, "Ultrafast Imaging in Standard (Bi)CMOS Technology," in *Photodetectors*, S. Gateva, Ed., 2012.

[16] P. W. Fuller, "Some Highlights in the History of High-Speed Photography and Photonics as Applied to Ballistics," in *High-Pressure Shock Compression of Solids VIII, The Science and Technology of High-Velocity Impact*, Springer, 2005, pp. 251-298.

[17] E. A. Igel and M. Kristiansen, *Rotating-Mirror Streak and Framing Cameras,* SPIE, 1997.

[18] . K. Scheidt and G. Naylor, "500 fs Streak Camera for UV-hard X-rays in 1 kHz Accumulating Mode with Optical-jitter Free-synchronisation," in *4th Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators*, 1999.

[19] Z. Jiang , G. G. Sun and X. C. Zhang, "Terahertz Pulse Measurement with an Optical Streak Camera," *Optics Letters,* vol. 24, no. 17, p. 1245–1247, 199.

[20] F. J., . H. J. Shin, J. R. Nasiatka, W. Wan, A. T. Young, . G. Huang, A. Comin, . J. Byrd and H. A. Padmore, "An X-ray Streak Camera with High Spatio-temporal Resolution," *Applied Physics Letters,* vol. 91, no. 13, p. 2007.

[21] E. K. Zavoisky and S. D. Fanchenko, "Image Converter High-speed Photography with 10-9–10-14 sec Time Resolution," *Applied Optics,* vol. 4, no. 9, pp. 1155-1167, 1965.

[22] . W. Uhring, M. Jung and P. Summ , "Image Processing Provides Low-frequency Jitter Correction for Synchroscan Streak Camera Temporal Resolution Enhancement," *Proc. of SPIE,,* vol. 5457, p. 245–252, 2004.

[23] E. K. Zavoisky and S. Franchenko, "On investigation of high-speed light processes," *Reports of the Soviet Academy of Sciences (DAN SSSR),* vol. 100 (4), pp. 661-663, 1955.

[24] E. K. Zavoisky and S. D. Fanchenko, "Physical principles of image-converter chronography," *DAN SSR,* vol. 108, pp. 2418-221, 1956.

[25] E. K. Zavoisky and S. Fanchenko, "Image Converter High-Speed Photography with 10−9–10−14 sec Time Resolution," *Applied Optics,* vol. 4, no. 9, pp. 1155-1167, 1965.

[26] T. J. Bensky, L. Clemo, C. Gilbert, B. Neff, M. A. Moline and D. Rohan, "Observation of nanosecond laser induced fluorescence of in vitro seawater phytoplankton," *Applied Optics,* vol. 47, no. 22, pp. 3980-3986, 2008.

[27] D. C. Swift, T. Tierney, R. A. Kopp and T. Gammel, "Shock pressures induced in condensed matter by laser ablation," *Physical Review E69,* 2004 .

[28] P. Montgomery, F. Anstotz, G. Johnson and R. Kiefer, "Real time surface morphology analysis of semiconductor materials and devices using 4D interference microscopy," *Journal of Materials Science: Materials in Electronics,* vol. 19, pp. 194-198, 2008.

[29] M. Zlatanski, W. Uhring, V. Zint, J.-P. Le Normand and D. Mathiot, "Technology, Architectures and Signal Reconstruction Methods for Nanosecond Resolution Integrated Streak Camera in Standard CMOS," in *Design and Architectures for Signal and Image Processing (DASIP), 2010 Conference*, Edinburgh, 2010.

[30] C. Lai, "A New Tubeless Nanosecond Streak Camera Based on Optical Deflection and Direct CCD Imaging," *SPIE Proceedings,* vol. 1801, pp. 454-468, 1999.

[31] C. Lai, D. R. Goosman, J. T. Wade and G. R. Avara, "Design and field test of a galvanometer deflected streak camera," *SPIE Proceedings,* vol. 4948, pp. 330-335, 2003.

[32] "Cordin Model 131-HD Camera," [Online]. Available: http://www.cordin.com.

[33] B. Casadei, J. P. Le Normand, Y. Hu and B. Cuni, "Design and characterization of a fast CMOS multiple linear array imager for nanosecond light pulse detections," *IEEE Tranactions on Instrumentaion and Measurement,,* vol. 52, no. 6, pp. 1892-1897, 2003.

[34] F. Morel, J. P. Le Normand, C. V. Zint, W. Uhring, Y. Hu and D. Mathiot, "A new spatiotemporal CMOS imager with analog accumulation capability for nanosecond low-power pulse detections," *IEEE Sensors Journal,* vol. 6, no. 5, pp. 1200-1208, 2006.

[35] M. Zlatanski, W. Uhring, L. N. Jean-Pierre, . C.-V. Zint and . D. Mathiot , "Integrated Circuit Architectures for High-speed Time-resolved Imaging," in *Sensor Technologies and Applications (SENSORCOMM)*, Venice, 2010.

[36] G. P. Weckler, "Operation of p-n Junction Photodetectors in a Photon Flux Integrating Mode," *IEEE Journal of Solid-State Circuits,* Vols. sc-2, no. 3, pp. 65-73, 1967.

[37] K. Kleinfelder, Y. Chen, K. Kwiatkowski, and A. Shah, "High-Speed CMOS Image Sensor Circuits With In Situ Frame Storage," *IEEE Transactions on Nuclear Science,* vol. 51, no. 4, pp. 1648-1656, 2004.

[38] M. Zlatanski, W. Uhring, . J. P. Le Normand, C. V. Zint and D. Mathiot, "12 × 7.14 GS/s rate time-resolved BiCMOS imager," in *8th IEEE International NEWCAS Conference,*, Montreal, QC, 2010.

[39] M. Zlatanski, W. Uhring and J. Le Normand, "Sub-500 ps Temporal Resolution Streak-mode Optical Sensor,," *2015,* vol. 15, no. 11, pp. 6570-6583, IEEE Sensors Journal,.

[40] W. Uhring and M. Zlatanski , "Streak-mode Optical Sensor in Standard BiCMOS Technology," *Sensors, 2011 IEEE,* pp. pp. 1604-1607, 2011.

[41] L. M. Bollinger and G. E. Thomas, "Measurement of the Time Dependence of Scintillation Intensity by a Delayed-Coincidence Method," *Review of Scientific Instruments,* vol. 32, p. 1044–1050, 1961.

[42] . D. O'Connor and D. Phillips, Time-Correlated Single Photon Counting, London: Academic Press, 1984.

[43] M. B., T. S., V. F, T. A and Z. F., "A High-Linearity, 17 ps Precision Time-to-Digital Converter Based on a Single-Stage Vernier Delay Loop Fine Interpolation," *Circuits and Systems I: Regular Papers, IEEE Transactions,* vol. 60, no. 3, pp. 557-569, 2013.

[44] J. Pichette, E. Lapointe and Y. Bérubé-Lauzière, "Time-domain 3D localization of fluorescent inclusions in a thick scattering medium," *Proc. SPIE,* vol. 7099, p. 709907, 2008.

[45] Field, R. M., Realov, S. and Shepard, K. L., "A 100-fps, Time-Correlated Single-Photon-Counting-Based Fluorescence-Lifetime Imager in 130-nm CMOS," *IEEE Journal of Solid-State Circuits,* vol. 49, no. 4, pp. 867 - 880, 2014.

[46] J. Léonard, N. Dumas, J. Caussé, S. Maillot, N. Giannakopoulou, S. Barre and W. Uhring, "High-throughput time-correlated single photon," *Lab on a Chip,* no. 14, pp. 4338-4343, 2015.

[47] M. Wahl, "Modern TCSPC Electronics: Principles and Acquisition Modes," in *Advanced Photon Counting. Applications, Methods, Instrumentation*, 1 ed., Springer, 2015, pp. 1-21.

[48] M. Wahl, "Time-Correlated Single Photon Counting," [Online]. Available: http://www.picoquant.com/images/uploads/page/files/7253/technote_tcspc.pdf.

[49] D. W. Becker, Advanced Time-Correlated Single Photon Counting Techniques, A. Castleman, J. Toennies and W. Zinth, Eds., Springer, p. 81.

[50] "Photomultiplier tubes – basics and applications, 3rd ed. Hamamatsu," [Online]. Available: https://www.hamamatsu.com/resources/pdf/etd/PMT_handbook_v3aE.pdf.

[51] A. Bulter, "Single-Photon Counting Detectors for the Visible Range Between 300 and 1,000 nm," in *Advanced Photon Counting. Applications, Methods, Instrumentation*, Springer, 2015, pp. 23-42.

[52] "Photosensor modules H7422 series. Hamamatsu," [Online]. Available: . http://www.hamamatsu.com/resources/pdf/.

[53] "PMA series photomultiplier detector assembly. PicoQuant.," [Online]. Available: http://www.picoquant.com/images/uploads/downloads/pma_series.pdf.

[54] R. E. Warburton, "Infrared Time-Correlated Single-Photon Counting. Thesis submitted for the Degree of Doctor of Philosophy. Heriot-Watt University," 2008.

[55] "High speed compact HPD (hybrid photo detector) R10467U-40/R11322U-40. Hamamatsu.," [Online]. Available: http://www.hamamatsu.com/resources/pdf/etd/R10467U-40_R11322U-40_TPMH1337E01.pdf.

[56] M. Ghioni, A. Gulinatti, I. Rech, F. Zappa and S. Cova, "Progress in Silicon Single-Photon Avalanche Diodes," *IEEE Journal of Selected Topics in Quantum Electronics,* vol. 13, no. 4, pp. 852-862, 2007.

[57] R. M. Field, S. Realov and K. L. Shepard, "A 100 fps, Time-Correlated Single-Photon-Counting-Based Fluorescence-Lifetime Imager in 130 nm CMOS," *IEEE Journal of Solid-State Circuits,* vol. 49, no. 4, pp. 867-880, 2014.

[58] M. Perenzoni, N. Massari, D. Perenzoni, L. Gasparini and D. Stoppa, ""A 160 x 120 Pixel Analog-Counting Single-Photon Imager With Time-Gating and Self-Referenced Column-Parallel A/D Conversion for Fluorescence Lifetime Imaging," *IEEE Journal of Solid-State Circuits,* vol. 51, no. 1, pp. 155-167, 2016.

[59] D. Palubiak, M. M. El-Desouki, O. Marinov, M. J. Deen and Q. Fang, "High-Speed, Single-Photon Avalanche-Photodiode Imager for Biomedical Applications," *IEEE Sensors Journal,* vol. 11, no. 10, pp. 2401-2412, 2011.

[60] L. Pancheri, N. Massari and D. Stoppa, "SPAD Image Sensor With Analog Counting Pixel for Time-Resolved Fluorescence Detection," *IEEE Transactions on Electron Devices,* vol. 6, no. 10, pp. 3442-3449, 2013.

[61] D. Bronzi, F. Villa, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers and W. Brockherde, "100 000 Frames/s 64 × 32 Single-Photon Detector Array for 2-D Imaging and 3-D Ranging," *IEEE Journal of Selected Topics in Quantum Electronics,* vol. 20, no. 6, pp. 354-363, 2014.

[62] F. Villa, S. Bellisai, D. Bronzi, A. Tosi , F. Zappa, S. Tisa, D. Durini , S. Weyers , U. Paschen and W. Brockherde, "SPAD Smart Pixel for Time-of-Flight and Time-Correlated Single-Photon Counting Measurements," *IEEE Photonics Journal,,* vol. 3, no. 4, pp. 795-804, 2012.

[63] M. Sergio, C. Niclass and E. Charbon, "A 128x2 CMOS Single-Photon Streak Camera with Timing-Preserving Latchless Readout," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, San Francisco, CA, 2007.

[64] I. Vornicu, R. Carmona-Galán and A. Rodríguez-Vázquez, "On the Calibration of a SPAD-Based 3D Imager with In-Pixel TDC Using a Time-Gated Technique," in *Circuits and Systems (ISCAS), 2015 IEEE International Symposium*, Lisbon, 2015.

[65] C. Niclass and E. Charbon, "single photon detector array with 64×64 resolution and millimetric depth accuracy for 3D imaging," *Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International, San Francisco, CA,* vol. 1, pp. 364-604, 2005.

[66] C. Veerappan , J. Richardson, R. Walker, D. U. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa , F. Borghetti , M. Gersbach, R. K. Henderson and E. Charbon, "A 160×128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, San Francisco, CA, 2011.

[67] F. Villa, D. Bronzi, S. Bellisai, G. Boso, A. B. Shehata, C. Scarcella, A. Tosi, F. Zappa, S. Tisa, D. Durini, S. Weyers and W. Brockherde, "SPAD imagers for remote sensing at the single-photon level," *Proc. SPIE,* vol. 8542, pp. 85420G-1–85420G-6, 2012.

[68] D. Perenzoni , L. Gasparini, N. Massari and D. Stoppa, "Depth-Range Extension With Folding Technique for SPAD-Based TOF LIDAR Systems," in *SENSORS, 2014 IEEE*, Valencia, 2014.

[69] H. Finkelstein, *Shallow-trench-isolation bounded single-photon avalanche diodes in commercial deep submicron CMOS technologies. PhD Thesis.,* University of California,San Diego, 2007.

[70] F. Zappa, A. Giudice, M. Ghioni and S. Cova, "Fully-integrated active-quenching circuit for single-photon detection," in *Solid-State Circuits Conference, 2002. ESSCIRC 2002. Proceedings of the 28th European*, Florence, Italy, 2002.

[71] Edoardo Charbon, Matt Fishburn, Richard Walker, Robert K. Henderson and Cristiano Niclass, "SPAD-Based Sensors," in *TOF Range-Imaging Cameras*, Springer, pp. 11-38.

[72] D. Bronzi, F. Villa, S. Bellisai, S. Tisa, G. Ripamonti and A. Tosi, "Figures of merit for CMOS SPADs and arrays," *Proc. SPIE,* vol. 8773, pp. 877304–1–877304-7, 2013.

[73] Gian-Franco Dalla Betta, Lucio Pancheri, David Stoppa, Robert Henderson and Justin Richardson, "Avalanche Photodiodes in Submicron CMOS Technologies for High-Sensitivity Imaging, Advances in Photodiodes," 2011. [Online]. Available: http://www.intechopen.com/books/advances-in-photodiodes/avalanche-photodiodes-in-submicron-cmos-technologies-for-high-sensitivity-imaging.

[74] M. A. KARAMI, "Deep-submicron CMOS Single Photon Detectors and Quantum Effects. PhD Thesis," 2011.

[75] 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,* vol. 20, no. 6, pp. 409-426, 2014.

[76] A. Gallivanoni, I. Rech and M. Ghion, "Progress in Quenching Circuits for Single Photon Avalanche Diodes," *IEEE Transactions on Nuclear Science,* vol. 57, no. 6, pp. 3815-3826, 2010.

[77] S. Cova, A. Longoni and A. Andreoni, , "Towards picosecond resolution with single-photon avalanche diodes," *Sci. Instrum.,* vol. 52, p. 408–412, 1981.

[78] S. Cova, "US Patent #4,963,727, Italian Patent 22367A/88, Industrial Licence to SILENA S.p.A., Milano, Italy.".

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

[80] A. Lacaita, S. Cova, M. Ghioni and F. Zappa, "Single photon avalanche diodes with ultrafast pulse response free from slow tails," *IEEE Electron Device Lett.,* vol. 14, no. 7, p. 360–362, 1993.

[81] A. Spinelli, M. Ghioni, S. Cova, and L. M. Davis, "Avalanche detector with ultraclean response for time-resolved photon counting," *IEEE J. Quantum Electron,* vol. 34, no. 5, p. 817–821, 1998.

[82] A. Rochas, M. Gani, B. Furrer, P. A. Besse, R. S. Popovic, G. Ribordy and N. Gisin, "Single photon detector fabricated in a complementary metal–oxide–semiconductor high-voltage technology," *Review of Scientific Instruments,* vol. 74, no. 7, pp. 3263-3270, 2003.

[83] A. Rochas , M. Gosch, P. A. Besse, R. S. Popovic, T. Lasser and R. Rigler, "First fully integrated 2-D array of single-photon detectors in standard CMOS technology," *IEEE Photonics Technology Letters,* vol. 15, no. 7, pp. 963-965, 2003.

[84] Niclass C, Rochas A, Besse PA and Charbon E, "A CMOS single photon avalanche diode array for 3D imaging," *IEEE Int. Solid-State Circuit Conf. (ISSCC),* p. 120–121, 2004.

[85] Stoppa D, Pancheri L, Scandiuzzo M, Malfatti M, Pedretti G and Gonzo L, "A single-photon-avalanche-diode 3D imager," *Proc. 31st European Solid-State Circuits Conf,* p. 487–490, 2005.

[86] E. Charbon, "Single-photon imaging in complementary metal oxide semiconductor processes," *Philos Trans A Math Phys Eng Sci.,* 28 Mar 2014 .

[87] Simona Privitera, Salvatore Tudisco , Luca Lanzanò, Francesco Musumeci , Alessandro Pluchino, Agata Scordino, Angelo Campisi, Luigi Cosentino, Paolo Finocchiaro, Giovanni Condorelli, Massimo Mazzillo, Salvo Lombardo and Emilio Sciacca , "Single Photon Avalanche Diodes: Towards the Large Bidimensional Arrays," *Sensors,* pp. 4636-4655, 2008,.

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

[89] E. A. G. Webster, L. A. Grant and R. K. Henderson, "A High-Performance Single-Photon Avalanche Diode in 130-nm CMOS Imaging Technology," *IEEE Electron Device Letters,* vol. 33, no. 11, pp. 1589-1591, 2012.

[90] C. Niclass , M. Soga and S. Kato, "A 0.18µm CMOS single-photon sensor for coaxial laser rangefinders," *Solid State Circuits Conference (A-SSCC), 2010 IEEE Asian, Beijing,* pp. pp. 1-4, 2010.

[91] N. Faramarzpour, M. J. Deen, S. Shirani and Q. Fang, "Fully integrated single photon avalanche diode detector in standard CMOS 0.18-µm technology," *IEEE Trans. Electron Devices,,* vol. 55, no. 3, p. 760–767, 2008.

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

[93] M. A. Karami, M. Gersbach, H.-J. Yoon and E. Charbon, "A new single photon avalanche diode in 90 nm standard CMOS technology," *Opt. Exp.,* vol. 18, no. 21, p. 22158–22166, 2010..

[94] E. A. G. Webster., J. A. Richardson, L. A. Grant, D. Renshaw and R. K. Henderson, "A single-photon avalanche diode in 90-nm CMOS imaging technology with 44% photon detection efficiency at 690 nm," *IEEE Electron Device Lett.,* vol. 33, no. 5, p. 694–696, 2012.

[95] C. Veerappan and E. Charbon, "A Substrate Isolated CMOS SPAD Enabling Wide Spectral Response and Low Electrical Crosstalk," *IEEE Journal of Selected Topics in Quantum Electronics,* vol. 20, no. 6, pp. 299-305, 2014.

[96] W. Uhring, C.V. Zint and J. Bartringer, "A low-cost high-repetition-rate picosecond laser diode pulse generator," *Proc. SPIE, 5452,* p. 583, 2004.

[97] "Very high longterm stability synchroscan streak camera," *Rev. Sci. Instrum. 74, 2646.*

[98] Danilo Bronzi, Federica Villa, Simone Bellisai,, Bojan, Simone Tisa, Alberto Tosi, Franco Zappa , Sascha Weyers, Daniel Durini, Werner Brockherde, and Uwe Paschen, "Low-noise and large-area CMOS SPADs with timing response free from slow tails," in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, Bordeaux, 2012.

[99] S. Isaak, Bull, S., Pitter, M.C. and Harrison, I., "Fully integrated linear single photon avalanche diode (SPAD) array with parallel readout circuit in a standard 180 nm CMOS process," *Enabling Science and Nanotechnology (ESciNano), 2010 International Conference,* pp. 1-2, 2010.

[100] H. Finkelstein, M.J. Hsu and S.C. Esener, "STI-Bounded Single-Photon Avalanche Diode in a Deep-Submicrometer CMOS Technology," *IEEE Electron Device Letters,* vol. 27, no. 11, pp. 887-889, 2006.

[101] F. Guerrieri, S. Bellisai, A. Tosi, G. Padovini, F. Zappa and S. Tisa, "SPAD arrays for parallel photon counting and timing," in *2010 23rd Annual Meeting of the IEEE Photonics Society*, Denver, CO, 2010.

[102] S. Tisa, F. Zappa and I. Labanca, "On-chip detection and counting of single-photons," *Electron Devices Meeting, 2005. IEDM Technical Digest,* 815-818.

[103] L. Pancheri and D. Stoppa, "Low-noise single Photon Avalanche Diodes in 0.15 μm CMOS technology," *Proceedings of the European Solid-State Device Research Conference (ESSDERC),* pp. 179-182, 2011.

[104] S. Henzler, Time-to-Digital Converters, Springer Netherlands.

[105] Wu Gao, Deyuan Gao, Christine Hu-Guo and Yann Hu , *Integrated High-Resolution Multi-Channel Time-to-Digital Converters (TDCs) for PET Imaging, Biomedical Engineering, Trends in Electronics, Communications and Software,* M. A. Laskovski, Ed.

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

[107] M. A. Z. Straayer, *Noise Shaping Techniques for Analog and Time to Digital Converters Using Voltage Controlled Oscillators, PhD thesis at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY,* 2008.

[108] N. R. Mahapatra, A. Tareen and S. V. Garimella, "Comparison and analysis of delay elements," in *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium*, 2002.

[109] J. M. Rabaey, A. P. Chandrakasan and B. Nikolic, Digital Integrated Circuits: A Design Perspective,, Prentice Hall Electronics and VLSI Series ed., 2003.

[110] D. Abramovitch, "Phase-Locked Loops: A Control Centric Tutorial," *Proceedings of the 2002 American Control Conference (ACC),* pp. 1-16, 2002.

[111] Yuwen Wang, Fan Ye and Junyan Ren, "A DLL based low-phase-noise clock multiplier with offset-tolerant PFD," in *ASIC (ASICON), 2013 IEEE 10th International Conference*, Shenzhen, 2013.

[112] C. Zhang and M. Syrzycki, "Modifications of a Dynamic-Logic Phase Frequency Detector for extended detection range," in *Circuits and Systems (MWSCAS), 2010 53rd IEEE International Midwest Symposium*, Seattle, WA, 2010.

[113] Aby Thomas and P.T. Vanathi, "An Enhanced Architecture of CMOS Phase Frequency Detector to Increase the Detection Range," *Semiconductor Technology and Science,* vol. 14, no. 2, pp. 198-201, 2014.

[114] G. B. Lee, P. K. Chan and L. Siek, "A CMOS phase frequency detector for charge pump phase-locked loop," in *Circuits and Systems, 1999. 42nd Midwest Symposium*, Las Cruces, NM, 1999.

[115] S. Li and M. Ismail, "A High-Performance Dynamic-Logic Phase-Frequency Detector," in *Trade-Offs in Analog Circuit Designn: the designer's companion*, Chris Toumazou, George Moschytz, Barrie Gilbert and Ganesh Kathiresan, Eds., The Netherlands: Kluwer Academic Publishers, pp. 821-842.

[116] Aby Thomas and P. T. Vanathi, "An Enhanced Architecture of CMOS Phase Frequency," *JOURNAL OF SEMICONDUCTOR TECHNOLOGY AND SCIENCE,* vol. 14, no. 2, pp. 198-201, 2014.

[117] D. Fischette, "Practical Phase-Locked Loop Design 2004 ISSCC Tutorial," [Online]. Available: http://www.delroy.com/PLL_dir/tutorial/PLL_tutorial_slides.pdf.

[118] Y. Moon, J. Choi, K. Lee, D. –K. Jeong and M. –K. Kim, "An All-analog Multiphase Delay locked Loop Using a Replica Delay Line for Wide Range and Low-jitter Performance," *IEEE Journal of Solid-State Circuits,* vol. 35, no. 3, p. 377–384, 2000.

[119] P. E. Su and S. Pamarti,, "Mismatch Shaping Techniques to Linearize Charge Pump Errors in Fractional- N PLLs," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 57, no. 6, pp. 1221-1230, 2010.

[120] L. Xue and Z. Zhan, "Differential charge pump circuit for high speed PLL application," in *Industrial Electronics & Applications, 2009. ISIEA 2009. IEEE Symposium*, Kuala Lumpur, 2009.

[121] I. A. Young, J. K. Greason and K. L. Wong, "A PLL Clock Generator with 5 to 1 lOMHz of Lock Range for Microprocessors," *IEEE Journal of Solid-State Circuirs,* vol. 27, pp. 1599-1607, 1992.

[122] W. Rhee, "Design of high-performance CMOS charge pumps in phase-locked loops," in *Circuits and Systems, 1999. ISCAS '99. Proceedings of the 1999 IEEE International Symposium*, rlando, FL, 1999.

[123] Jae-Shin Lee, Min-Sun Keel, Shin-Il Lim and Suki Kim, "Charge pump with perfect current matching characteristics in phase-locked loops," *Electronics Letters,* vol. 36, no. 23, pp. 1907-1908, 9 Nov 2000.

[124] E. Juarez-Hernandez and A. Diaz-Sanchez, "A Novel CMOS Charge-Pump Circuit with Positive Feedback for PLL Applications," in *Electronics, Circuits and Systems, 2001. ICECS 2001. The 8th IEEE International Conference*, 2001.

[125] Vassilis Kalenteridis, Konstantinos Papathanasiou and Stylianos Siskos , "Analysis and Design of Charge Pumps for Telecommunication Applications," in *VLSI-SoC: Design Methodologies for SoC and SiP,*, C. Piguet, R. Reis and D. Soudris, Eds., pp. 43-60.

[126] K. Moustakas and S. Siskos, "Low voltage CMOS charge pump with excellent current matching based on a rail-to-rail current conveyor," in *New Circuits and Systems Conference (NEWCAS) , 2015 IEEE 13th International*, Grenoble, 2015.

[127] Greason, J.K. and Wong, K.L., "A PLL Clock Generator with 5 to 110MHz of Lock Range for Microprocessors," *IEEE Journal of Solid-State Circuits,* vol. 27, no. 11, p. 1599–1607, 1992.

[128] J. Maneatis, " Low-Jitter and Process-Independent DLL and PLL Based on Self-Biased Techniques," *ISSCC Digest of Technical Papers,* 1996.

[129] M. Zlatanski, Integrated Streak Camera in Standard BiCMOS Technology. PHD thesis, University of Strasbourg, 2011.

[130] M. Abidi , K. Koua Calliste, M. Kanoun, S. Panier, L. Arpin, M. A. Tétraul , J. F. Pratte and R. Fontaine, "A Delay Locked Loop for fine time base generation in a positron emission

tomography scanner," in *5th International Conference on Design and Technology of Integrated Systems in Nanoscale Era (DTIS),*, Hammamet, 2010.

[131] P. Larsson, "A 2-1600-MHz CMOS clock recovery PLL with low-Vdd capability," *IEEE Journal of Solid-State Circuits,* vol. 34, no. 12, pp. 1951-1960, 1999.

[132] Hsiang-Hui Chang, Jyh-Woei Lin, Ching-Yuan Yang and Shen-Iuan Liu, "A wide-range delay-locked loop with a fixed latency of one clock cycle," *IEEE Journal of Solid-State Circuits,* vol. 37, no. 8, pp. 1021-1027, 2.

[133] M. Zlatanski, Integrated Streak Camera in Standard BiCMOS Technology. PHD thesis, University off Strasbourg, 2011.

[134] Villa, F., Lussana, R, Tamborini, D., Bronzi, D., Markovic, B., Tosi, A., Zappa, F. and Tisa, S., "CMOS single photon sensor with in-pixel TDC for Time-of-Flight applications," in *Time-to-Digital Converters (NoMe TDC), 2013 IEEE Nordic-Mediterranean Workshop*, Perugia, 2013.

[135] B. K. Swann, B.J. Blalock, L. G. Clonts, D. M. Binkley, J. M. Rochelle, E. Breeding, and K. M. Baldwin, "A 100-ps time-resolution CMOS time-to-digital converter for positron emission tomography imaging applications," *IEEE Journal of Solid-State Circuits,* vol. 39, no. 11, pp. 1839-1852, 2004.

[136] M. Kanoun, Y. Berube-Lauziere and R. Fontaine, "High precision time-to-amplitude converter for diffuse optical tomography applications," in *Design and Technology of Integrated Systems in Nanoscale Era, 2008. DTIS 2008. 3rd International Conference*, Tozeur, 2008.

[137] J. Léonard, N. Dumas, J. Caussé, S. Maillot, N. Giannakopoulou, S. Barre and W. Uhring, "High-throughput time-correlated single photon counting," *Lab on a Chip,* vol. 14, pp. 4338-4343, 2015.

[138] M. Gersbach, Y. Maruyama, R. Trimananda, M. W. Fishburn, D. Stoppa,, J. A. Richardson, R. Walker, R. Henderson and E. Charbon, "Time-Resolved, Low-Noise Single-Photon Image Sensor Fabricated Deep-Submicron CMOS Technology," *IEEE Journal of Solid-State Circuit,* vol. 47, pp. 1394-1407, 2012.

[139] D. Tamborini, F. Villa and A. Tosi, "16-Channel Module Based on a Monolithic Array of Single-Photon Detectors and 10-ps Time-to-Digital Converters," *IEEE Journal of Selected Topics in Quantum Electronics,* vol. 20, no. 6, p. 1, 2014.

[140] W. J. Stewart, Markov chains, queues, and simulation, P. U. Press, Ed., 2009.

[141] K. Sigman. [Online]. Available: http://www.columbia.edu/~ks20/stochastic-I/stochastic-I-MCII.pdf.

[142] A. Cuccato , S. Antonioli, M. Crotti, I. Labanca, A. Gulinatti, I. Rech and M. Ghioni, "Complete and compact 32-channel system for time-correlated single-photon counting measurements," *EEE Photonics Journal,* vol. 5, no. 5, pp. 6801514-6801514, 2013.

[143] C. Niclass, M. Sergio and E. Charbon, "A single photon avalanche diode array fabricated in deep-submicron CMOS technology,"," *Proc. Design, Autom. Test Eur.,* pp. 1-6, 2006.

[144] Maruyama Y. , J. Blacksberg and E. Charbon, "A 1024 × 8, 700-ps time-gated SPAD line sensor for planetary surface exploration with laser Raman spectroscopy and LIBS," *IEEE J. Solid-State Circuits,* vol. 49, no. 1, p. 179–189, 2014.

[145] L. Pancheri and D. Stoppa, "A SPAD-based pixel linear array for highspeed time-gated fluorescence lifetime imaging," *Proc. Eur. Solid State Circuits Conf. (ESSCIRC),,* p. 428–431, 2009.

[146] C. Scarcella, A. Tosi, F. Villa, S. Tisa and F. Zappa, ""Low-noise low-jitter 32-pixels CMOS single-photon avalanche diodes array for single photon counting from 300 nm to 900 nm," *Rev. Sci. Instrum,* vol. 84, no. 12, p. 776–779, 2013.

[147] J. G. Walker, "Iterative correction for `pile-up' in single-photon lifetime measurement," *Opt. Comm,* vol. 201, p. 271–277, 2002.

[148] F. Villa, R. Lussana, D. Tamborini, A. Tosi and F. Zappa, "High-Fill-Factor 60\times 1 SPAD Array With 60 Subnanosecond Integrated TDCs," *IEEE Photonics Technology Letters,* vol. 27, no. 12, pp. 261-1264, 2015.

List of Publications

I. Malass, W. Uhring, J. P. Le Normand, N. Dumas, V. Zint and F. Dadouche, "SiPM based smart pixel for photon counting integrated streak camera," *Design and Architectures for Signal and Image Processing (DASIP), 2013 Conference on*, Cagliari, 2013, pp. 135-140.

I. Malass, W. Uhring, J. P. L. Normand, N. Dumas and F. Dadouche, "10-ps Resolution hybrid time to digital converter in a 0.18 µm CMOS technology," *New Circuits and Systems Conference (NEWCAS), 2014 IEEE 12th International*, Trois-Rivieres, QC, 2014, pp. 105-108.

I. Malass, W. Uhring, J. P. Le Normand, N. Dumas and F. Dadouche, "Efficiency improvement of high rate integrated time correlated single photon counting systems by incorporating an embedded FIFO," *New Circuits and Systems Conference (NEWCAS), 2015 IEEE 13th International*, Grenoble, 2015, pp. 1-4.

I. Malass, W. Uhring, J. P. Le Normand, N. Dumas and F. Dadouche, "A single photon avalanche detector in a 180 nm standard CMOS technology," *New Circuits and Systems Conference (NEWCAS), 2015 IEEE 13th International*, Grenoble, 2015, pp. 1-4.

I. Malass, W. Uhring, J. P. Le Normand, N. Dumas and F. Dadouche, "Evaluation of size influence on performance figures of a Single Photon Avalanche Diode fabricated in a 180 nm standard CMOS technology" in *Analog Integrated Circuits and Signal Processing, Springer (Accepted , to be published)*

I. Malass, W. Uhring, J. P. Le Normand, N. Dumas and F. Dadouche, "A hybrid Time to Digital Converter based on Digital Delay Line Loop and Analog Time to Amplitude Converter" *New Circuits and Systems Conference (NEWCAS), 2016 IEEE 14th International*, Vancouver, 2016 *(Accepted , to be published)*

UNIVERSITE DE STRASBOURG

ECOLE DOCTORALE MATHEMATIQUES SCIENCES DE L'INFORMATION ET DE L'INGENIEUR

RESUME DE LA THESE DE DOCTORAT

Discipline : Électronique, Électrotechnique et Automatique

Spécialité (facultative) : Micro et Nanoélectronique

Présentée par : MALASS Imane
(*Nom Prénom du candidat)*

Titre: Integrated Streak Camera based on a Time Correlated Single Photon Counting System

Unité de Recherche : ICube (UMR 7357)
(N° et Nom de l'Unité)

Directeur de Thèse : UHRING Wilfried
(Nom Prénom – Grade)

Co-Directeur de Thèse (s'il y a lieu) : LE NORMAND Jean-Pierre
(Nom Prénom – Grade)

Localisation : 23, rue du Loess BP 20 F-67037 Strasbourg Cedex 2

Thèse confidentielle : ■ NON ☐ OUI

8.1 Introduction

Les systèmes d'imagerie ultra rapide sont conventionnellement basés sur la technologie des tubes à vide comme les tubes photomultiplicateurs (PMT pour Photomultiplier Tubes), les tubes imageurs à balayage (ST pour Streak Tubes) et les tubes intensificateurs (IIT pour Image Intensifier Tubes). Ces technologies offrent des résolutions temporelles de l'ordre de quelques centaines de ps dans le cas des systèmes à PMT et de l'ordre de quelques dizaines de ps dans le cas des imageurs 2D utilisant les IIT. La caméra de balayage de fente (CBF) est l'instrument de détection directe de la lumière le plus rapide au monde, le taux d'échantillonnage de ces appareils peut atteindre 1 Péta S/s. En sacrifiant une dimension spatiale ces caméras sont capables d'offrir des résolutions temporelles de l'ordre d'une centaine de fs dans le cas du fonctionnement en mode simple tir et de l'ordre du ps dans le cas du fonctionnement en mode accumulation. Toutefois, les systèmes d'imagerie utilisant les tubes à vide sont encombrants, fragiles et couteux.

Au cours des 10 dernières années des nombreux travaux de recherche ont été réalisés dans le but de trouver une meilleure alternative. Des imageurs 2D capables de capturer plusieurs méga-images/secondes ont été réalisées en technologie à transfert de charge (CCD pour Charge Coupled Device), la résolution temporelle de ces dispositifs est de l'ordre de 1 µs avec un taux d'échantillonnage de l'ordre de 1 TS/s. Une meilleure résolution a été obtenue par l'imagerie à balayage de fente et plusieurs CBF ont été réalisées en technologie CMOS et BiCMOS, ces dispositifs offrant une résolution temporelle de 1 ns ont été conçus spécialement pour le fonctionnement en mode simple tir et leur sensibilité est inadéquate pour la mesure des phénomènes optiques de faible intensité. Ce problème peut être résolu dans le cas des signaux récurrents en adoptant la technique de comptage de photon unique résolu dans le temps ou TCSPC (Time Correlated Single Photon Counting).

La TCSPC est une technique qui consiste à utiliser un photomultiplicateur à gain élevé pour détecter l'arrivée d'un photon, le moment de la détection est mesuré à l'aide d'un convertisseur de temps numérique (TDC pour Time to Digital Converter) ou un convertisseur temps vers amplitude (TAC pour Time to Amplitude Converter) et la forme du signal est reconstruit après répétition du processus pour obtenir un nombre suffisant de mesure. Suite à l'apparition des photomultiplicateurs à silicium (SiPM) basées sur des photodiodes à avalanche polarisée en mode Geiger ou SPAD (Single Photon Avalanche Detector), plusieurs systèmes TCSPC à 2 dimensions ont été réalisées et la première CBF intégrée utilisant des SPADs a été réalisée en 2007. Le travail réalisé durant cette thèse avait pour but la conception et la réalisation d'une caméra à balayage de fente (CBF) intégrée basée sur l'architecture d'un système TCSPC (TCSPC-SC for TCSPC Streak Camera).

Le TCSPC-SC tire avantage du mode d'imagerie linéaire « streak » pour surmonter la limitation de surface inhérente aux systèmes TCSPC bidimensionnelles. Cette solution

permet l'intégration de fonctionnalités électroniques additionnelles dans chaque pixel sans l'inconvénient d'un taux de remplissage faible qui résulte en une faible efficacité de détection. Le TCSPC-SC en technologie 180 nm CMOS standard comprend deux parties principales: le système de photodetection basé sur Une SPAD et le système de mesure de temps. La structure de SPAD a été sélectionnée parmi 6 structures différentes suite à un processus de caractérisation détaillé et complet afin de mesurer leurs performances. Le système de mesure de temps utilisé dans le TCSPC-SC est un convertisseur de temps (TDC) hybride qui combine le concept des TDC numériques utilisant des DLL et des compteurs et celui du TAC afin d'obtenir une bonne résolution temporelle et une large dynamique de mesure temporelle. De plus le TDC hybride a été spécialement conçu pour être intégrés dans un système TCSPC qui intègre un vecteur de TDCs ce qui a nécessité une conception attentive pour limiter la consommation et la surface d'occupation pour réaliser une architecture flexible et facilement extensible. En utilisant ces deux blocs, une structure de test englobant 8 unités ou Macropixels a été réaliser afin de démontrer la faisabilité et tester le TCSPC-SC toutefois notre but final est de réaliser prochainement un système TCSPC-SC complet englobant plus de Macropixels.

8.2 Architecture globale du circuit

Contrairement aux autres dispositifs d'imagerie, une caméra à balayage de fente offre une image spatiotemporelle I=f(x,t), pourtant en sacrifiant une dimension spatiale elle est capable d'offrir des résolutions temporelles comprises entre 1 ps et 100 fs. L'architecture de la CBF présentée dans ce document est basée sur un système TCSPC intégré, un tel système englobe deux parties principales : des SPADs avec leur circuit électronique associé dite circuit d'extinction et un convertisseur de temps numérique. Une SPAD est une photodiode à avalanche polarisée en mode Geiger, c'est-à-dire au-delà de sa tension d'avalanche. Dans ce mode, la génération d'un électron par effet photoélectrique est suffisante pour déclencher une avalanche auto-maintenue, ainsi chaque détection d'un photon va générer une avalanche qui doit être arrêtée afin de permettre à la SPAD de réaliser une nouvelle détection, cette tâche est réalisée par un circuit dit « circuit d'extinction » qui va arrêter l'avalanche en faisant diminuer la tension de polarisation de façon à ce que l'avalanche ne peut plus se maintenir d'elle-même (phase d'extinction) puis, une fois l'avalanche arrêter , ramène la SPAD à sa polarisation initiale (phase de reset). Les blocs fonctionnels de notre circuit sont illustrés dans la Figure 163, ce dernier consiste en 8 lignes identiques partageant une unité de mémoire, une boucle à verrouillage de retard (DLL pour Delay locked Loop) et un compteur 12bits. Le système peut être divisé en deux parties principales : les Macropixels effectuant la détection des photons et les unités de mesure de temps.

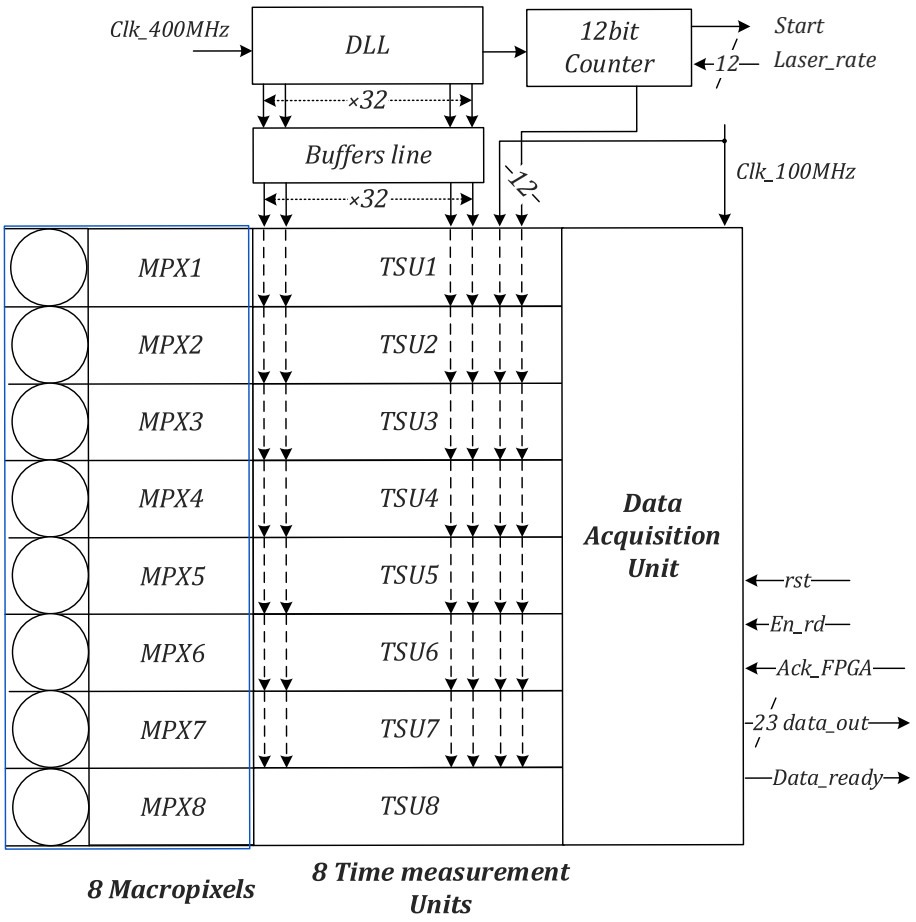


Figure 163 - Schéma fonctionnel de la structure de test TCSPC-SC réalisé

8.2.1 Photodiode à avalanche

Les photodiodes à avalanche polarisée en mode Geiger ou SPADs (pour l'anglais Single Photon avalanche detector) sont des jonctions PN polarisées au-delà de leur tension d'avalanche. Un champ électrique intense règne alors dans la zone de charge d'espace de sorte que l'apparition d'un seul porteur de charge est suffisante pour déclencher une avalanche auto-maintenue qui se traduit par une augmentation très rapide d'un courant d'avalanche. En supposant que le porteur de charge a été généré par effet photoélectrique, le début d'avalanche marque alors l'arrivée d'un photon à quelque picosecondes de gigue près. Au cours des dernières années, un grand travail a été réalisé pour optimiser les SPADs et les intégrer dans des applications telles que l'imagerie par mesure de temps de vie de fluorescence ou FLIM (pour l'anglais Fluorescence Lifetime Imaging), la tomographie par émission de positrons ou PET (Positron Emission Tomography) , et les mesures de temps de vol ou TOF (pour l'anglais Time of Flight measurements). Un grand nombre des SPADs a ainsi été implémenté en technologies standard et dédiée, et alors que les SPADs fabriquées en technologies dédiées ont été capable d'atteindre des très bonnes performances en termes de bruit et d'efficacité, les SPADs standard restent moins chers, consomme moins d'énergie permet le plus grand niveau d'intégration et de miniaturisation puisqu'il est possible d'intégrer la SPAD avec l'électronique associée sur un même chip. une SPAD réalisée en technologie 180 nm CMOS Image Sensor (CIS) standard et les résultats des mesures de caractérisation en termes de bruit, d'efficacité et de résolution temporelle.

8.2.1.1 Structure de la SPAD

Afin d'être utilisée comme une SPAD, la jonction pn doit avoir une structure qui permet d'avoir une bonne probabilité de détection et un taux d'obscurité ou DCR (pour l'anglais Dark Count Rate) suffisamment faible , en particulier il est nécessaire d'avoir un champ électrique uniforme sur la totalité de la région et de minimiser la concentrations des impuretés et des centres de génération et de recombinaison. La coupe transversale la structure réalisée est présentée dans la Figure 2, la région active est une jonction P+/Nwell confinée dans un caison Deep Nwell moins dopé pour l'isoler du substrat. Un double de garde Pwell/STI (Shallow Transisolation) est utilisé pour limiter les effets de bord, l'anneau en Pwell a été inséré pour séparer la zone en STI de la région active pour limiter le nombre des fausses détections dues aux injections de charge provenant de l'interface STI vers la zone de multiplication.

8.2.1.2 Circuit de caractérisation

Le circuit de caractérisation consiste de 8 SPADs ayant la structure présentée avec leurs circuits d'extinctions assemblées dans la forme de 8 pixels (Figure 165). Les 8 SPADs ont des surfaces différentes avec des diamètres de région active compris entre 5 µm et 40µm. L'électronique d'extinction associée à chaque SPAD permet de l'activer ou la désactiver

individuellement et la tester complètement sans interférence des autres structures. Il incluse un PMOS controlé en tension assurant l'extinction passive, un NMOS pour l'extinction passive et un PMOS pour le reset actif. Ces 2 derniers sont commandés par le block de "Quench Control" qui permet aussi de choisir le mode d'extinction (Passif, Actif, Mixte, Off) et de contrôler les durées des phases d'extinctions et de reset. Les 8 SPADs ont des surfaces différentes avec des diamètres de région active comprises entre 5 µm et 40µm toutefois dû à l'utilisation d'un anneau de garde en Pwell qui impose des limites sur le diamètre minimal de la zone active, la SPAD de 5 µm de diamètre n'était pas opérante. Par contre les autres SPADs ont subi une avalanche autour d'une tension de polarisation inverse de 12.2V (Figure 3) de plus le test de photoluminescence a montré que l'avalanche était uniformément répartie sur la totalité de la surface active (Figure 4).

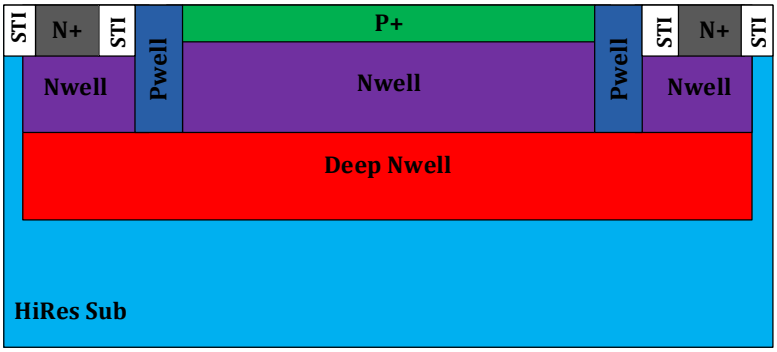


Figure 164 - représentation transversale de la structure de SPAD utilisée dans le macropixel

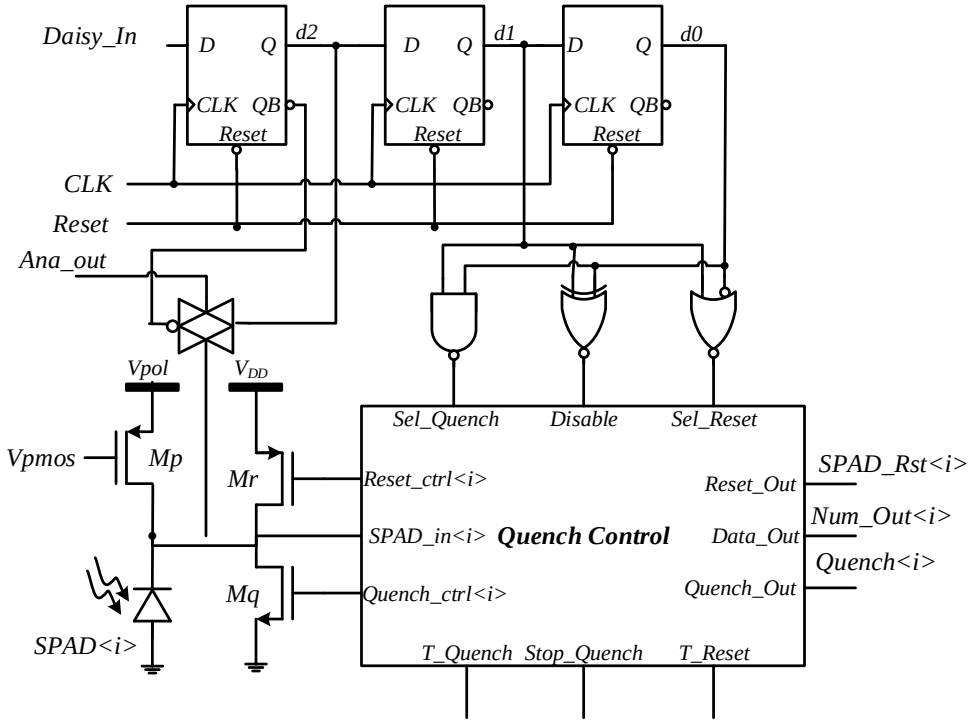


Figure 165 - Présentation schématique du pixel de test incluant la SPAD avec le circuit d'extinction associé.

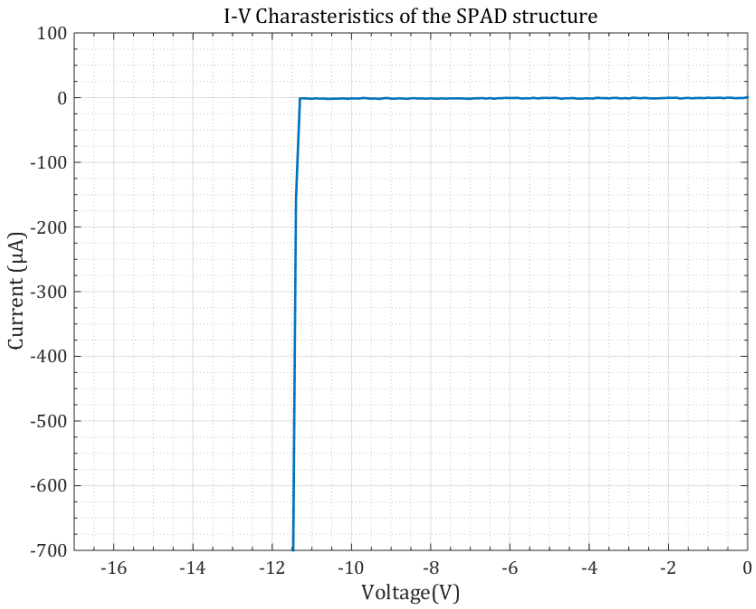


Figure 166 – Caractéristique IV de la SPAD en polarisation inverse

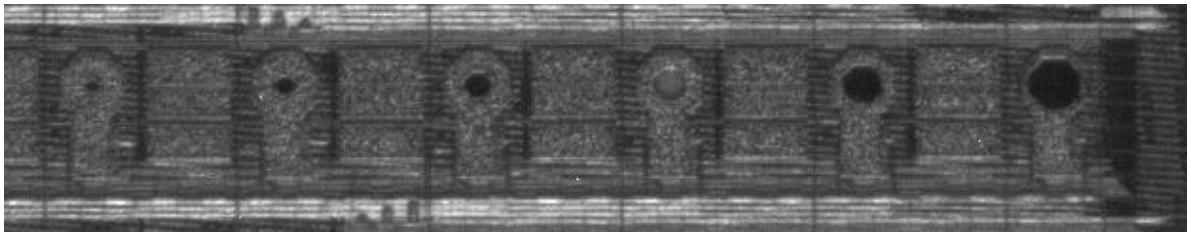


Figure 167 - Test de photoluminescence réalisée sur la SPAD ayant un diamètre de zone active égale à 25 µm pour une tension d'excès de 300 mV

8.2.1.3 Résultats de caractérisation

a. Taux d'obscurité et post-déclenchement

Le taux d'obscurité ou DCR (Pour l'anglais Dark Count Rate) représente la valeur moyenne des fausses détections observées en noir. Le DCR est le résultat d'un ou plusieurs phénomènes de générations (effet thermique, effet tunnel, charges piégées), il est influencé par plusieurs facteurs: le procédé de fabrication, la température, la tension d'excès, la surface et la structure de la SPAD. Le DCR est arbitraire mais les SPADs présentent parfois des avalanches secondaires statistiquement corrélées aux avalanches primaires, ce phénomène est dit post-déclenchement ou Afterpulsing. Le post-déclenchement est causé par des impuretés qui captent des porteurs de charges lors de l'avalanche puis les libèrent après que la SPAD est repolarisée au-delà de sa tension d'avalanche déclenchant alors une fausse détection. Cet effet s'amplifie lorsque la température diminue ce qui augmente le temps nécessaire pour la libération des charges piégées, il est possible de le minimiser en augmentant le temps mort suffisamment pour que toutes les charges soient libérées. pour

mesurer le DCR on a commencé par la mesure du post-déclenchements en mesurant les intervalles de temps séparant deux avalanches consécutives pour un temps mort de 30 ns qui est la plus petite valeur possible. La Figure 6 montre les résultats obtenus, une déviation de la distribution exponentielle attendue est causée par la présence du post-déclenchement et résulte en une probabilité de post-déclenchement égale à 0.2%, les mesures permettent aussi de déduire que le post-déclenchement peut être éliminé à partir d'un temps mort de 1µs. Le DCR a été mesuré pour plusieurs tensions d'excès et des températures variant entre -30°C et 60°C, Figure 7 montre la variation du DCR des 7 SPADs en fonction de la température et de la surface active pour une tension d'excès de 100 mV et de 400 mV. Pour une température de 15°C, Le DCR mesurée est inférieur à 10 KHz pour les SPADs ayant des diamètres de 40 µm et de 30 µm, il est inférieur à 5KHz pour les autres SPADs.

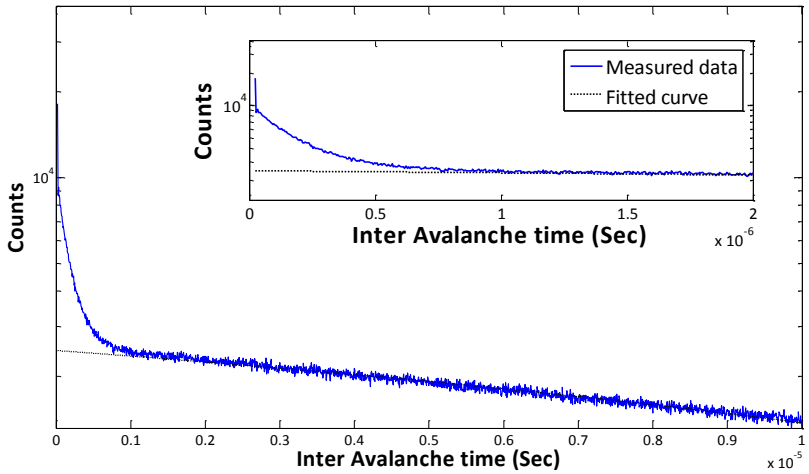


Figure 168 - Mesure de post-déclenchement pour un temps mort de 30 ns et une tension d'excès de 300mV

b. Probabilité de photodétection

La probabilité de photodétection ou PDP (pour l'anglais photodetection probability) est définit comme le rapport des photons détectés par la SPAD sur le nombre des photons incidents, elle est le produit du taux de remplissage, la probabilité de déclenchement et la probabilité d'absorption. La PDP augmente avec la surface de la zone active et la tension d'excès, elle dépend de la technologie et des niveaux des dopages utilisés aussi bien que la longueur d'onde de la source utilisée. La PDP des 7 SPAD ont été mesurées à température ambiante pour des longueurs d'onde allant de 350 nm à 1000 nm avec plusieurs tensions d'excès et un temps mort de 30 ns, Un PDP maximum de 20% a été mesuré autour de 430nm pour des tension d'excès inférieur à 500mV, ce qui est assez bon vue la tension d'excès limitée à cause de l'électronique utilisée. Les résultats des mesures obtenus pour une tension d'excès de 300 mV sont illustrés dans les Figure 8.

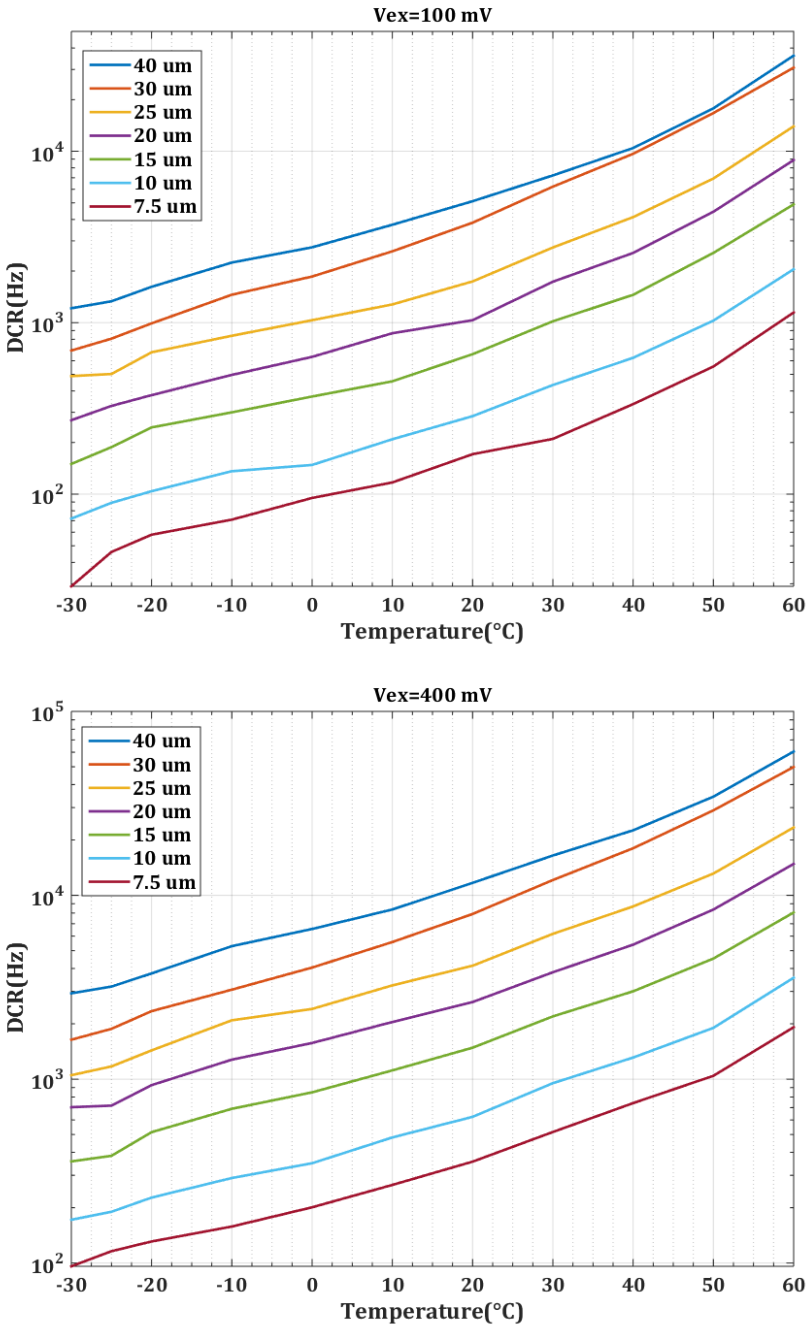


Figure 169- Variation du DCR des 7 SPADs en fonction de la température pour une tension d'excès de 100 mV et de 400 mV

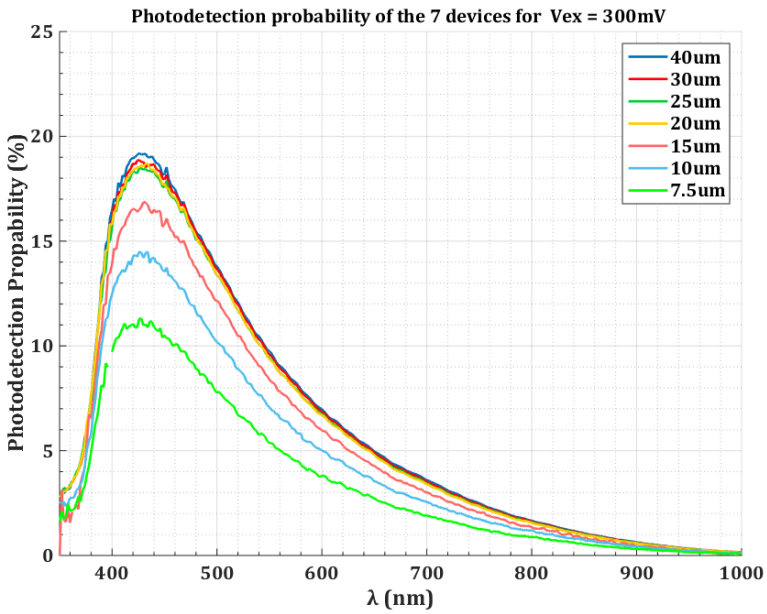


Figure 170 - La probabilité de photodétection des 7 SPADs à température ambiante pour une tension d'excès de 300 mV

c. Résolution temporelle

La résolution temporelle ou Jitter est définie comme la gigue temporelle entre le moment de l'arrivée d'un photon dans la zone de charge et le début de l'avalanche. Elle augmente avec la surface active et diminue avec la tension d'excès La gigue temporelle des 7 SPAD a été caractérisée à l'aide d'une source Laser à 20 MHz de fréquence répétition en diminuant son intensité suffisamment pour se placer dans une configuration de TCSPC, avec une tension d'excès de 300 nm et un temps mort de 30 ns. La réponse temporelle du système a été caractérisée pour une longueur d'onde de 405 nm (Figure 8) et de 830 nm (Figure 9). Les distributions obtenues suivent une loi gaussienne pour la partie centrée et une loi exponentielle pour la queue. La partie centrée est due aux photons absorbés dans la zone de charge d'espace (ZCE) et la queue est due aux porteurs de charges minoritaires qui arrivent à la ZCE par diffusion ou les photons qui pénètrent profondément dans le substrat. Pour caractériser le jitter, on utilise la largeur à mi-hauteur (FWHM pour Full Width Half Maximum), elle est égale à 100 ps pour λ=405 nm et 120 ps pour λ=830 nm et on voit d'après les graphes que la queue est plus marquée pour λ=830 nm.

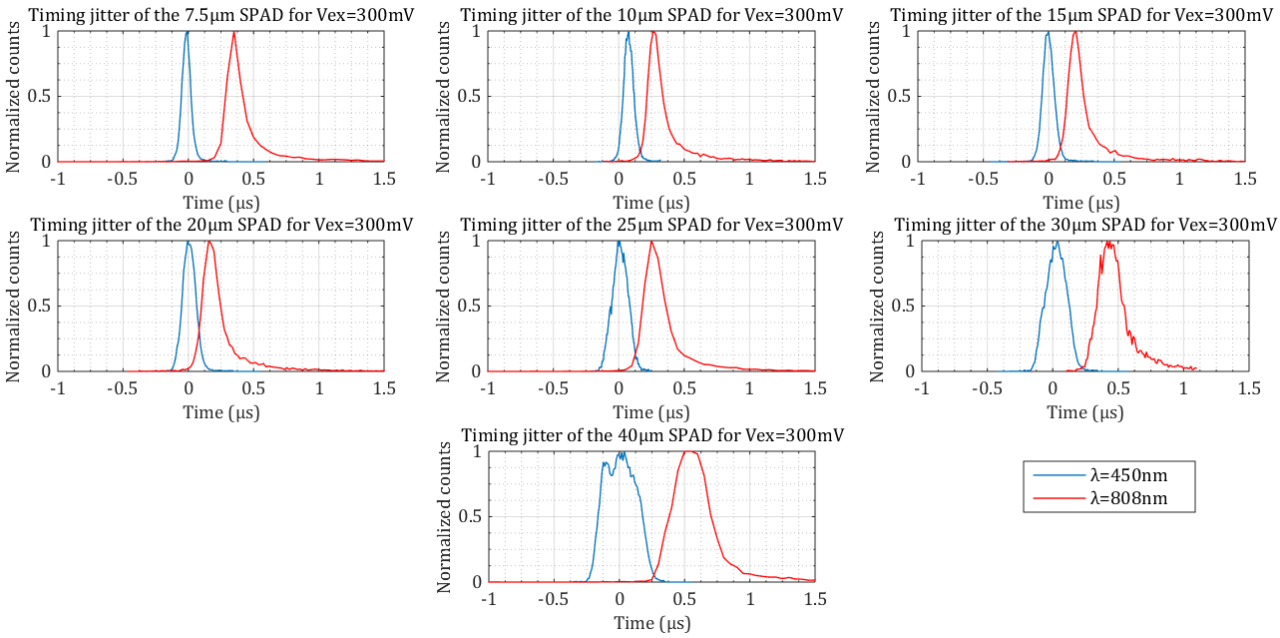


Figure 171 – distributions Normalisées de la réponse temporelle des 7 SPADs pour une tension d'excès de 300mV en utilisant des sources laser de 450 nm et de 808 nm

8.2.2 Architecture d'un Macropixel

La détection des photons est réalisée dans le circuit par un bloc de 8 Macropixels identiques, chaque Macropixel (Figure 163Figure 172) est constitué d'une SPAD avec son circuit d'extinction associé qui forment ensemble un Macropixel autonome (Figure 172). le PMOS commandé en tension « *Mp* » assure l'extinction passive permettant ainsi une flexibilité dans le choix de la valeur de la résistance d'extinction, l'extinction active est réalisée à travers le NMOS « *Mq* » et le reset active de la SPAD se fait à l'aide du PMOS « *Mr* ». Le circuit d'extinction dans chaque Macropixel est contrôlé par le bloc « Quench_Control » , de plus les durées des 2 phases d'extinctions et de reset peuvent être ajustées à l'aide des deux entrées analogiques externes T_Quench et T_Reset ce qui permet de contrôler le temps mort lié au processus d'extinction. Afin de limiter le nombre des entrées et sorties externes, un registre à décalage de 8 bits est utilisé pour fixer les niveaux des entrées En<i> permettant d'activer ou de désactiver individuellement

La structure de la SPAD intégrée dans le Macropixel est celle représentée en Figure 164, elle a été conçu avec un diamètre actif de 15 µm comme celle-ci représentée un bon compromis

entre le DCR et le PDP avec un DCR inférieur à 2KHz et un PDP de 20% pour λ=450nm, Vex=570 mV (Figure 173) et un taux de remplissage de 69%.

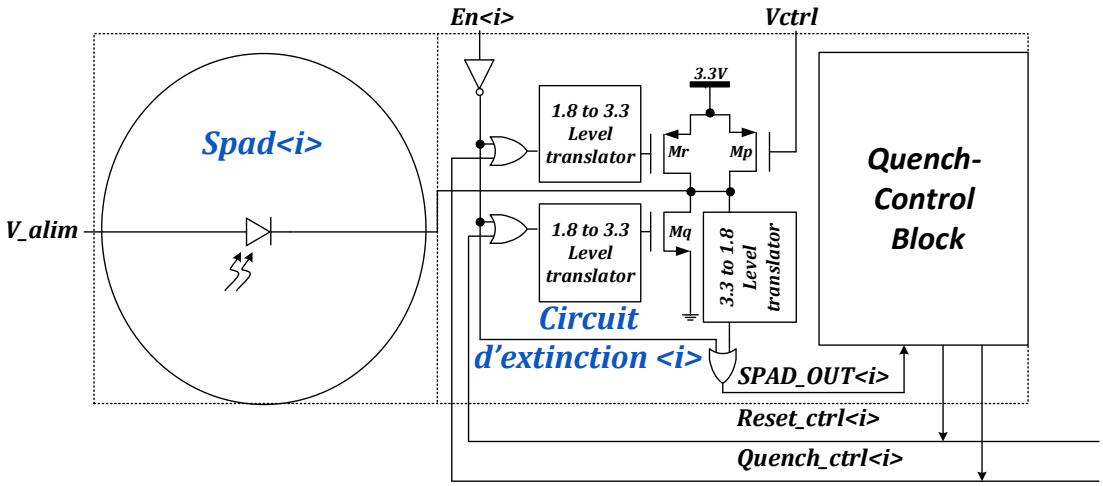


Figure 172 - Schéma fonctionnel d'un Macropixel de photodétection

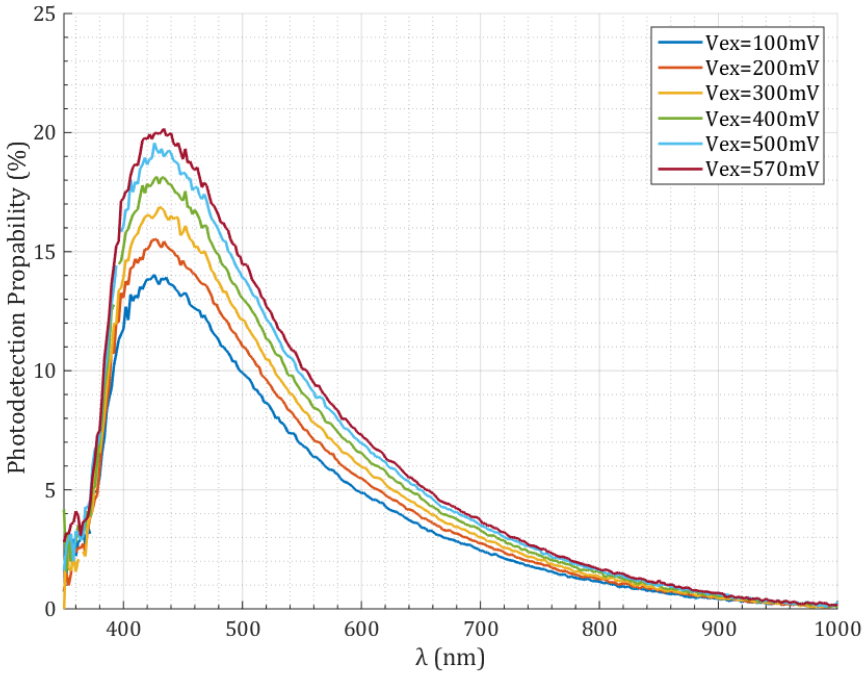


Figure 173 - La probabilité de photodétection de la SPAD ayant 20 µm de diamètre active à température ambiante pour une tension d'excès entre 100 mV et 570 mV

8.2.3 Le Convertisseur de temps hybride

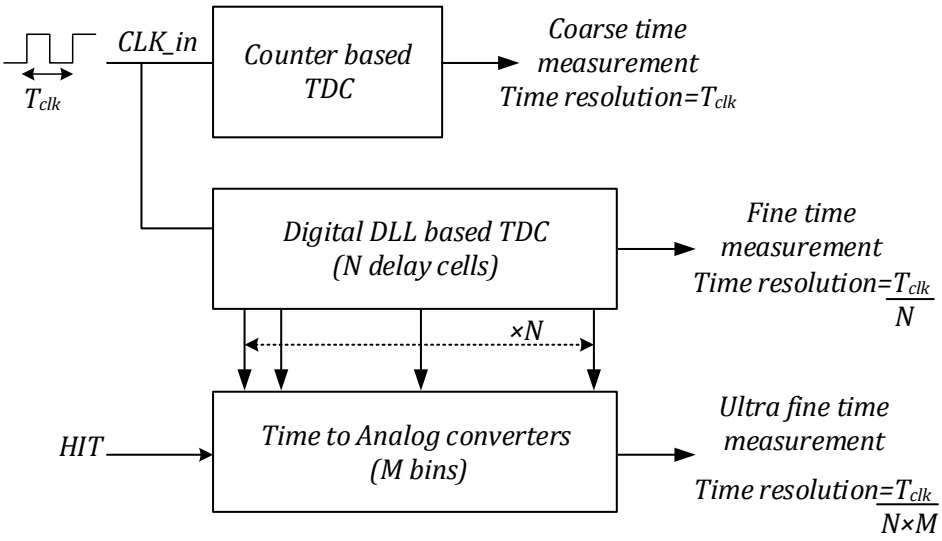


Figure 174 - Principe de fonctionnement du TDC hybride

Lorsqu'une avalanche est déclenchée dans une SPAD la sortie Out<i> du Macropixel<i> contenant la SPAD passe de 0 à1 signalant au convertisseur de temps qu'il y a eu détection d'un photon. Afin de mesurer le moment d'arrivée d'un photon un convertisseur de temps convertit en mot binaire l'intervalle de temps séparant l'arrivée des deux signaux : le signal START généré périodiquement par le compteur et le signal HIT<i> relié à la sortie du Macropixel. Les convertisseurs de temps sont généralement divisés en deux types: Les convertisseurs de temps analogiques basés sur un TAC suivi d'un convertisseur numérique (ADC) et les convertisseurs de temps numériques. Les TACs utilisent la technique d'intégration du courant, ils offrent de très bonnes résolutions temporelles avec une consommation d'énergie raisonnable mais souffrent d'une dynamique de mesure limitée. Cependant les TDCs généralement basées sur les lignes à retard offrent de bonnes dynamiques de mesure mais leur résolution temporelle est limitée par le retard élémentaire dans la technologie utilisée, une meilleure résolution peut être obtenue mais ceci est aux prix d'un circuit complexe et une consommation d'énergie élevée. Le convertisseur de temps utilisé dans notre circuit combine les concepts des TACs et des TDCs pour obtenir une résolution temporelle de 10ps et une dynamique de mesure allant jusqu'à 5.21µs.

L'unité de mesure de temps (Figure 163) dans notre circuit est constituée de 8 convertisseurs de temps partageant une DLL de 32 étages, un compteur 12 bit et une unité de mémoire enregistrant les résultats des mesures. Le TDC hybride peut être présenté comme un convertisseur de temps à 3 étages (Figure 174), le premier étage est un TDC à compteur offrant une mesure grossière du temps avec une large dynamique de mesure, Le deuxième étage est un TDC à DLL offrant une résolution temporelle fine de Tclk/N où et *N* est le nombre d'éléments de la ligne de retard dans la DLL. Le dernier étage désigné par unité

d'échantillonnage de temps (TSU pour Time Sampling Unit) est un TAC où les sorties intermédiaire de la ligne à retard commandée en tension (VCDL pour Voltage Controlled Delay Line) sont utilisées pour charger une ligne de capacités. Lorsque le signal HIT<i> passe à 1, les niveaux des sorties intermédiaires de la ligne à retard sont enregistrés. La pente du signal numérique étant quantifié en M niveaux, la tension stockée dans la cellule où il y a eu une transition de'1' vers '0' est échantillonnée ce qui donne une résolution temporelle ultrafine de Tclk/M×N

8.2.3.1 Architecture globale du TDC hybride

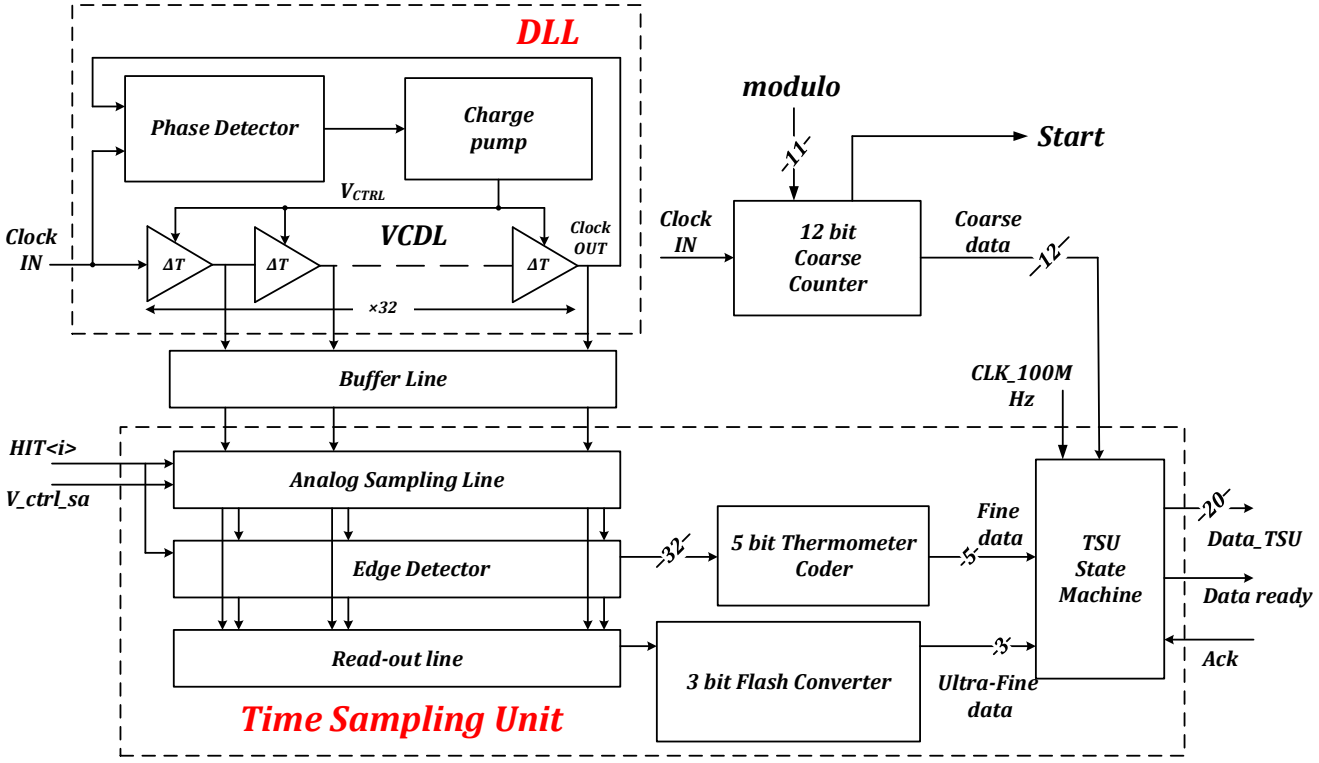


Figure 175 - Schéma fonctionnel du TDC hybride

Les blocs essentiels du TDC hybride sont illustrés dans la Figure 175, la DLL englobe une VCDL de 32 éléments à retard connecté à un détecteur de phase et une pompe de charge. Les 32 sorties intermédiaires de la VCDL sont reliées à une ligne d'échantillonnage analogique (ASL pour Ananlog Sampling Line) à 32 éléments à travers une ligne de buffers. Chaque cellule d'échantillonnage analogique (Figure 176) comprend un buffer dégénéré contrôlé par la tension externe V_ctrl_sa, ce buffer sert à limiter les effets de la variation du mode de fonctionnement du NMOS M_{NS}, sur la forme du signal à échantillonner, il permet aussi de contrôler le temps de descente du signal afin de régler la résolution du TDC. Les cellules de la ASL suivent la variation de de la VCDL, lorsque le signal numérique HIT<i>passe à 1 M_{NS} devient bloquant et les sorties *Vsa<k>* arrêtent d'évoluer gardant leur dernière valeur.

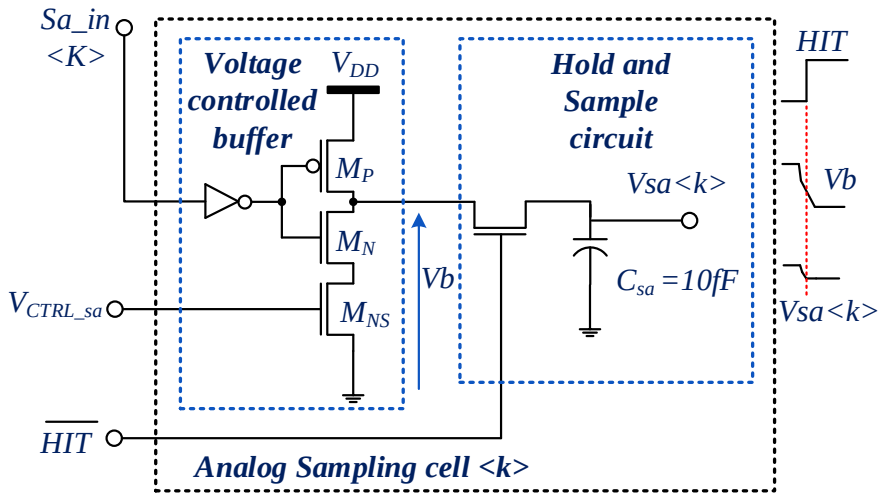


Figure 176 - Schéma d'une cellule d'échantillonnage analogique

Le détecteur du front (Edge Detector contient 32 éléments identique (EDC pour Edge Detector Cell) (Figure 177) activés lorsque HIT<i> passe à '1'et sert à détecter la cellule où le signal échantillonné passe de '1' à '0' La sortie D<k> parallèle à cette cellule passe alors à '1' tandis que les autres restent à '0'. Les EDCs ont été dimensionnés de façon à ce que le seuil de détection du niveau '0' soit égale à celle de la détection du niveau '1' pour éviter le cas où aucune transition n'est détectée à cause du décalage. De plus une mémoire RS est utilisée pour éviter la fluctuation de sortie lorsque le signal varie autour du seuil de basculement celle-ci est commandé par un signal WR qui correspond au signal HIT inversé et décaler de 20 ns.

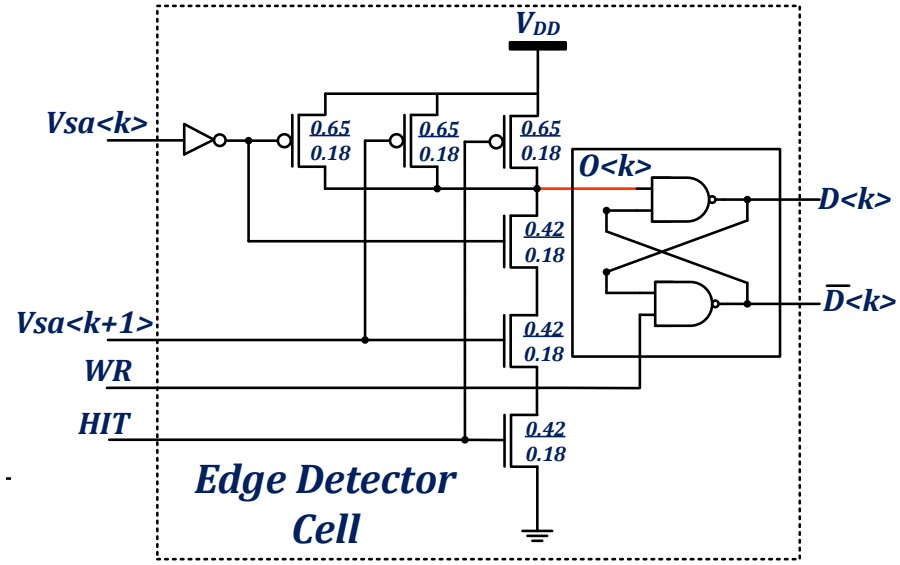


Figure 177 - Schéma d'une cellule EDC.

Les bits des sorties D<k> sont converties en un mot de 5 bits représentant les résultats de la mesure fine de temps à travers un encodeur thermométrique (5 bit Thermometer Coder), elles sont aussi utilisés pour commander la ligne de lecture (ROL pour ReadOut Line) (Figure

178) contenant 32 PMOS à source suiveur. Seule La sortie de la cellule qui a reçu un D<k>='1' sera transmise à l'entrée *Vread* du convertisseur flash à 8 niveaux qui convertit la tension analogique reçue en un mot de 3 bits. Finalement une machine d'état (TSU state machine) s'occupe de la transmission du résultat final combinant les résultats des 3 mesures grossière, fine et ultrafine sous forme d'un mot de 20 bits en total vers l'unité de mémoire.

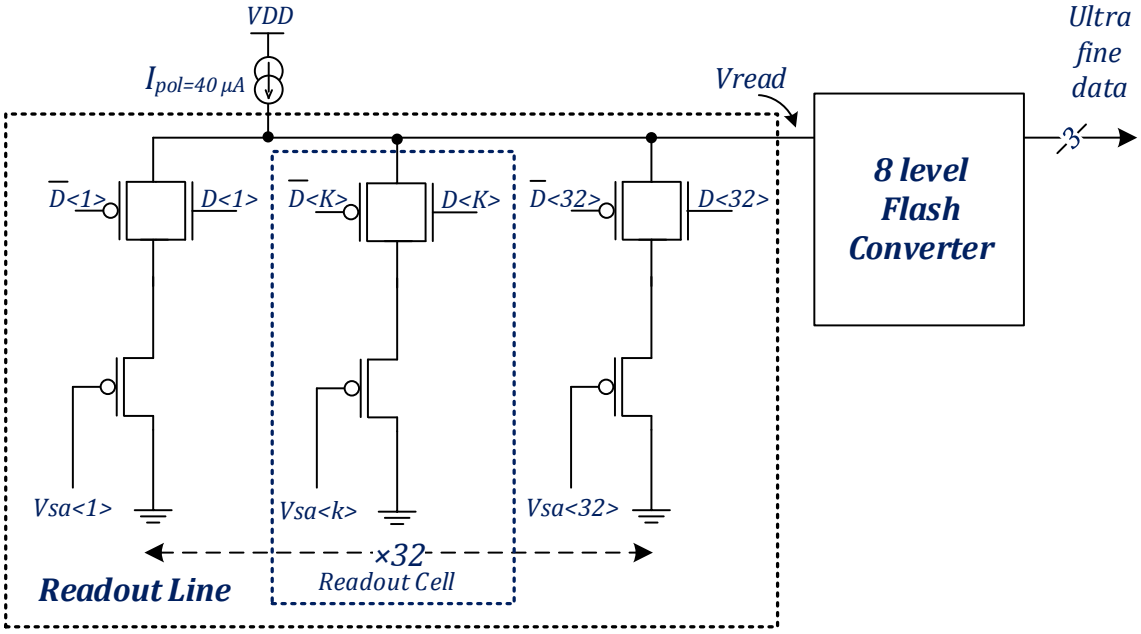


Figure 178 - Schéma de la ligne ROL à 32 éléments et les convertisseurs à sa sortie

8.2.4 Simulations

Le TDC hybride a été réalisé en technologie 180 nm CMOS et son fonctionnement a été validé par simulation sous Cadence pour une d'entrée de DLL de 2,5 ns, Vctrl-sa = 607 mV et en faisant varier le temps d'arrivée du 'HIT' d'un pas de 1 ps afin de mieux évaluer la méthode de conversion à 3 étages réalisée par l'UFC en termes de précision et de linéarité. La Figure 17 montre les résultats de la simulation des conversions fines et ultrafines sur un intervalle de 90 ps, et la Figure 18 montre les résultats de conversion de temps obtenus à la fin du processus de conversion. En combinant les résultats des conversions fines et ultrafines, la résolution temporelle initiale de 80 ps a été augmentée jusqu'à 10 ps, sans avoir à utiliser une fréquence d'horloge plus élevée. Les simulations ont également montré un niveau de non-linéarité estimée à ~2 ps résultant de la non-linéarité de la pente du signal de sortie de la cellule d'échantillonnage analogique. En fait, la faible divergence entre les pentes réelles et idéales (Figure 19) conduit à des variations dans les valeurs des intervalles de temps équivalents à chaque intervalle de tension. Toutefois, cette divergence peut être corrigée en utilisant un convertisseur non linéaire assurant ainsi que chaque sous-intervalle de tension sera exactement équivalent à un intervalle de temps de 10 ps.

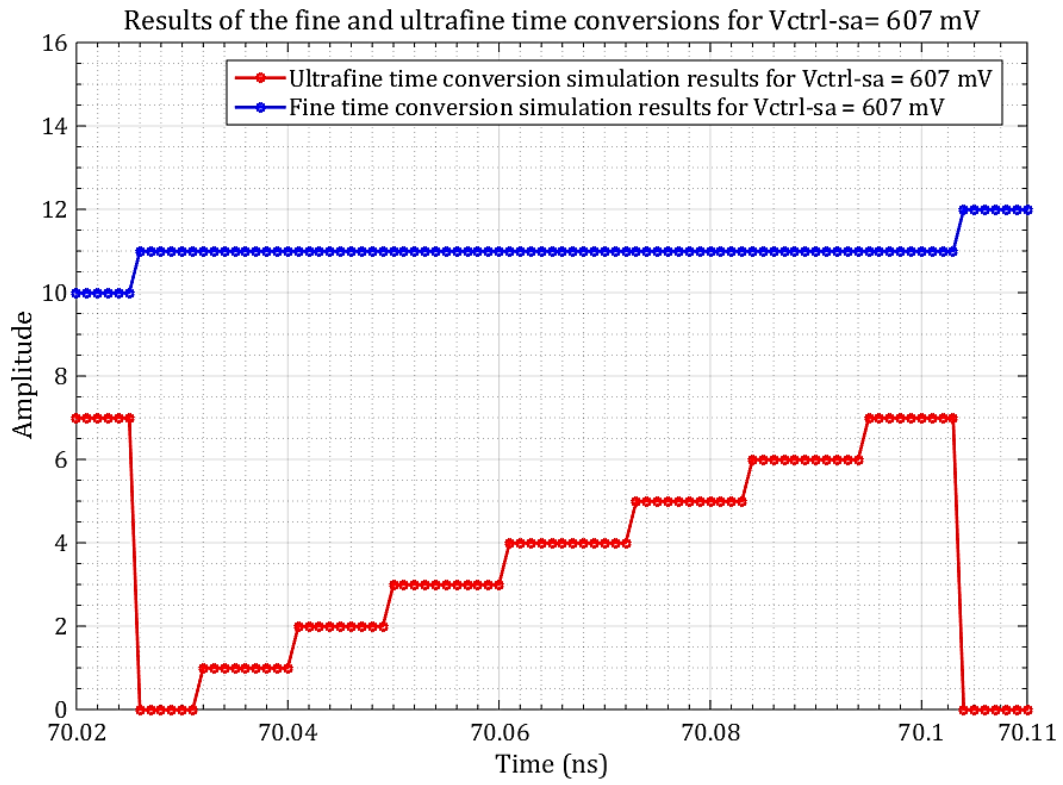


Figure 179 - Résultats de la simulation des conversions fines et ultrafines sur un intervalle de 90 ps

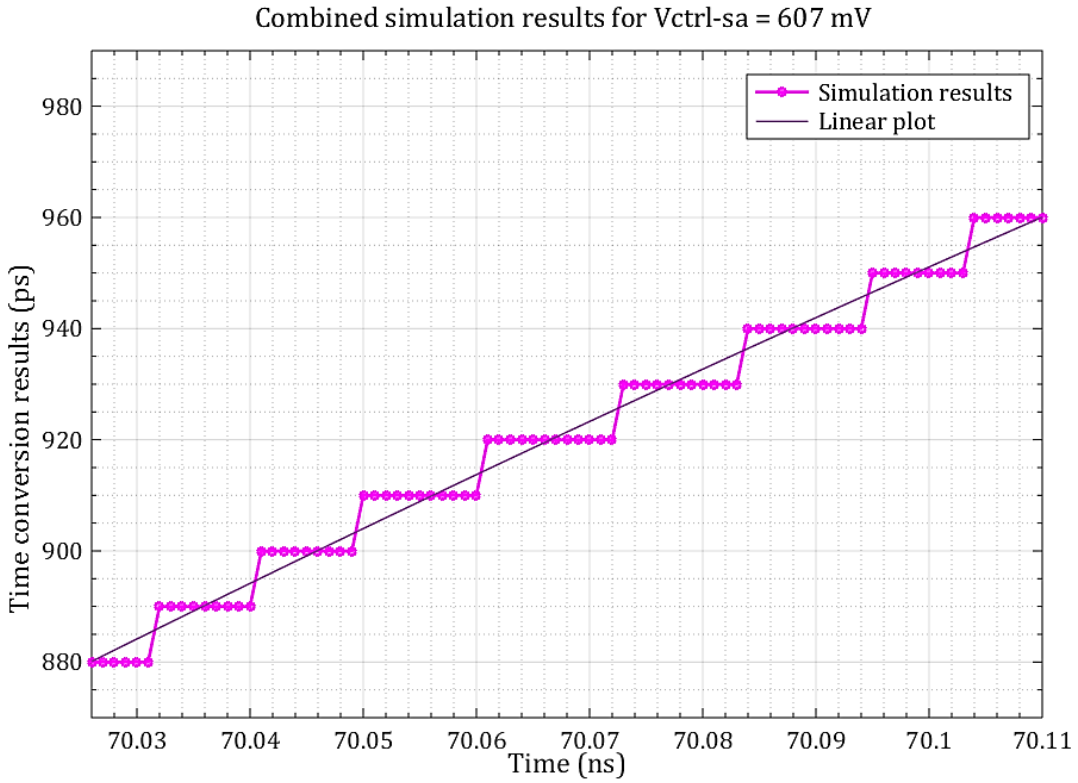


Figure 180 - Résultats de la simulation de la conversion totale sur un intervalle de 90 ps

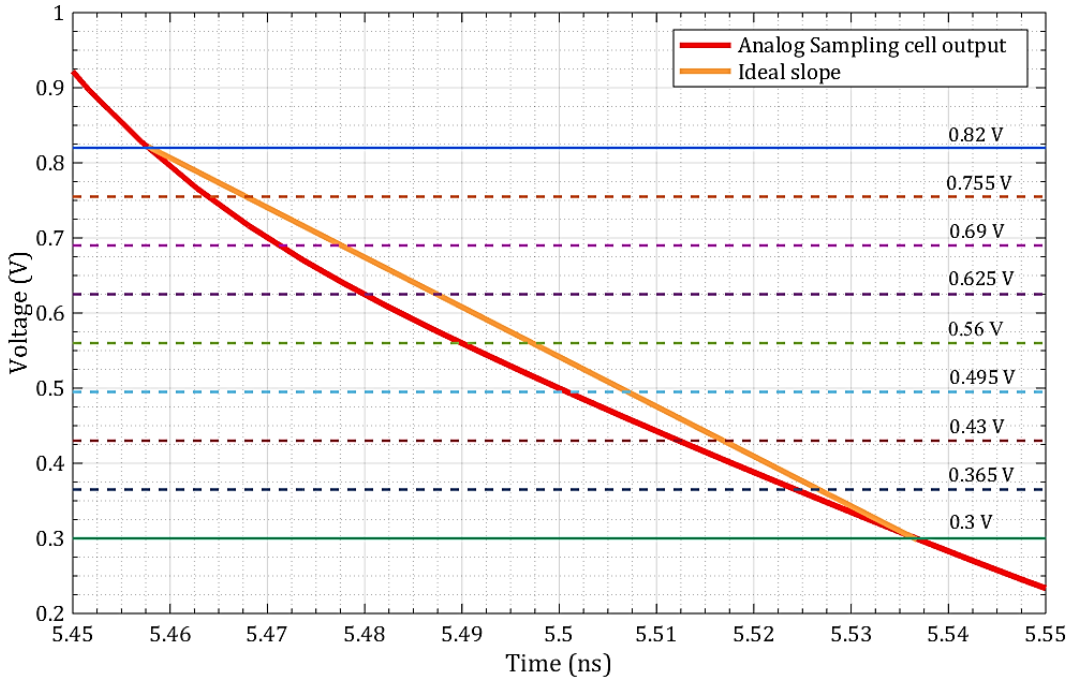


Figure 181 - Écart entre la pente descendante de la cellule d'échantillonnage analogique et une pente idéale avec les 8 sous-intervalles des tensions délimitant les 8 sous-intervalles de temps

8.2.5 Mesures

Le TDC hybride a été fabriqué dans une technologie 180 nm CMOS standard et des mesures ont été réalisées pour évaluer les opérations de conversion temps fines et ultrafines accomplies par le bloc TSU du TDC. Un générateur «Stanford research system DG645" a été utilisé pour effectuer les mesures. Les résultats moyens pour 130 mesures des sorties de la conversion temporelle fine, de la conversion de temps ultrafine et des sorties totales sont illustrés pour un intervalle de mesure de 2800 ps dans la Figure 20. L'amélioration de la résolution temporelle obtenue après l'ajout des résultats de la conversion de temps ultrafine est visible malgré les fluctuations dans les sorties de l'UFC (Figure 21). ces perturbations qui sont surtout plus marquées pour les étages supérieurs, sont très probablement une conséquence du « layout » du circuit et ce problème peut donc être résolu par un « layout » plus adapté pour limiter les chutes de potentiel et en ajoutant des condensateurs de découplage pour limiter les fluctuations parasites des tension d'alimentation. Une vue approchée des mesures sur un intervalle de 500 ps pour les 3 conversions de temps fine, ultrafine et totale est représenté sur la Figure 22, l'amélioration obtenue avec la conversion à 3 étages est tout à fait remarquable. Cependant les résultats ont été lissés en raison de la gigue de 7 ps rms (40 ps pic-pic) associée aux mesures. De plus, l'INL (Figure 23) a été améliorée de 21 ps rms jusqu'à ~ 5.6 rms ps après l'ajout des résultats de la conversion ultrafine.

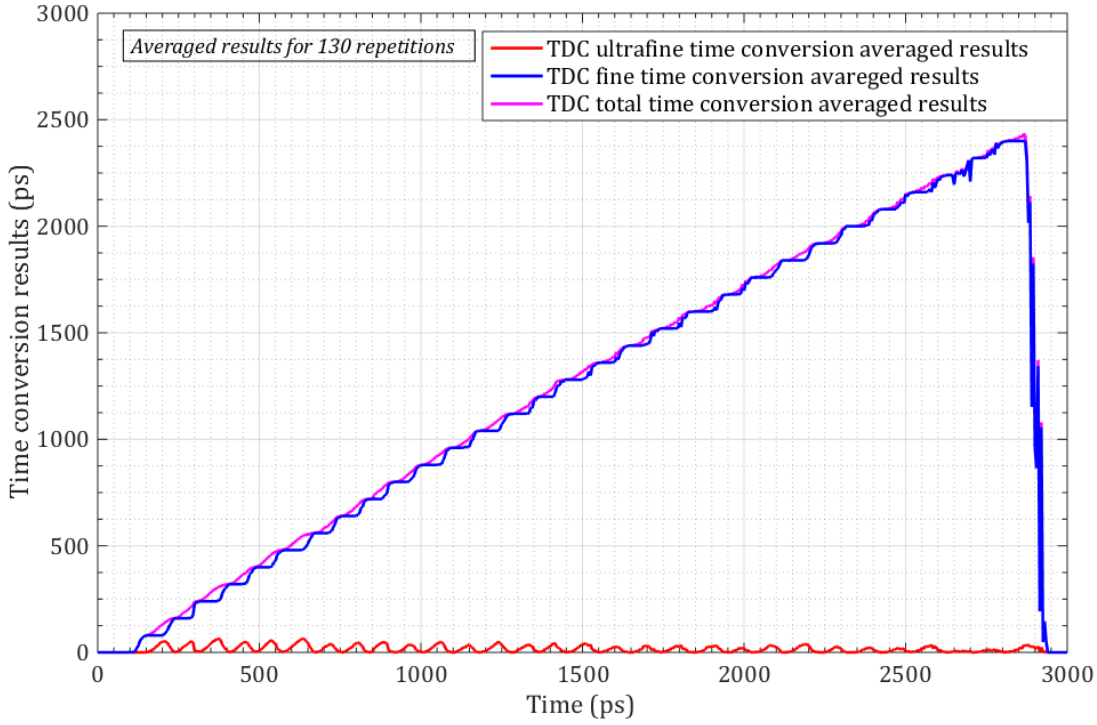


Figure 182 – valeurs moyennées des résultats des mesures sur une intervalle de 2800 ps pour les conversions de temps fine, ultrafine et totale

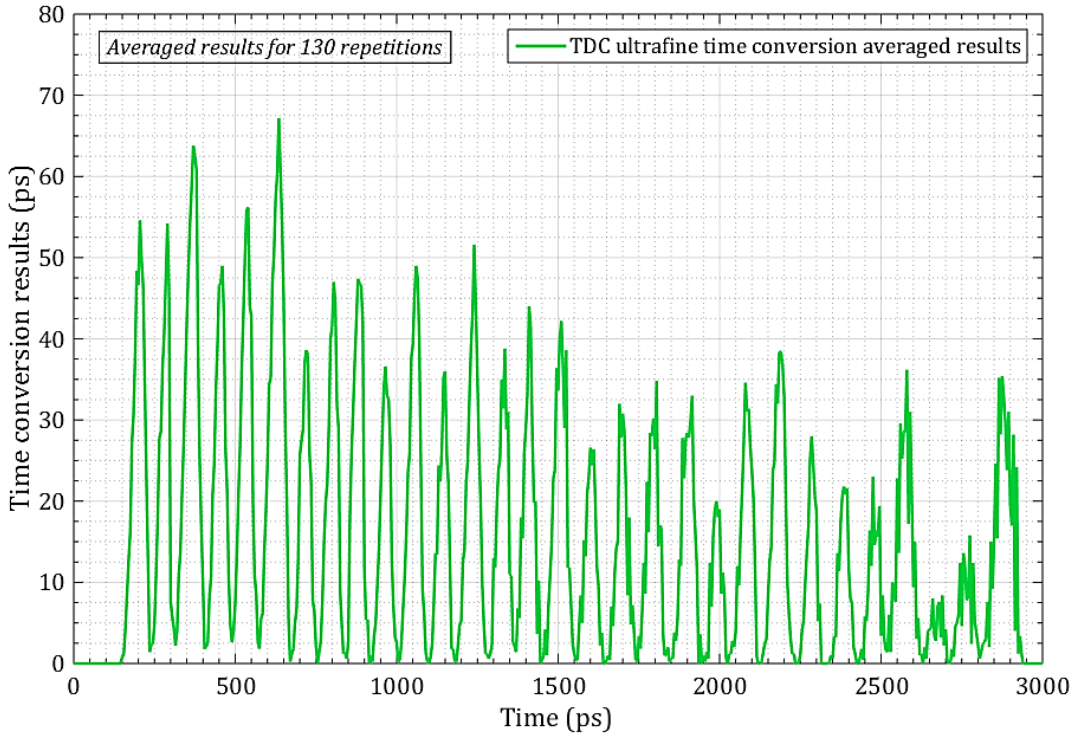


Figure 183 - valeurs moyennées des résultats des mesures sur une intervalle de 2800 ps pour les conversions de temps, ultrafine

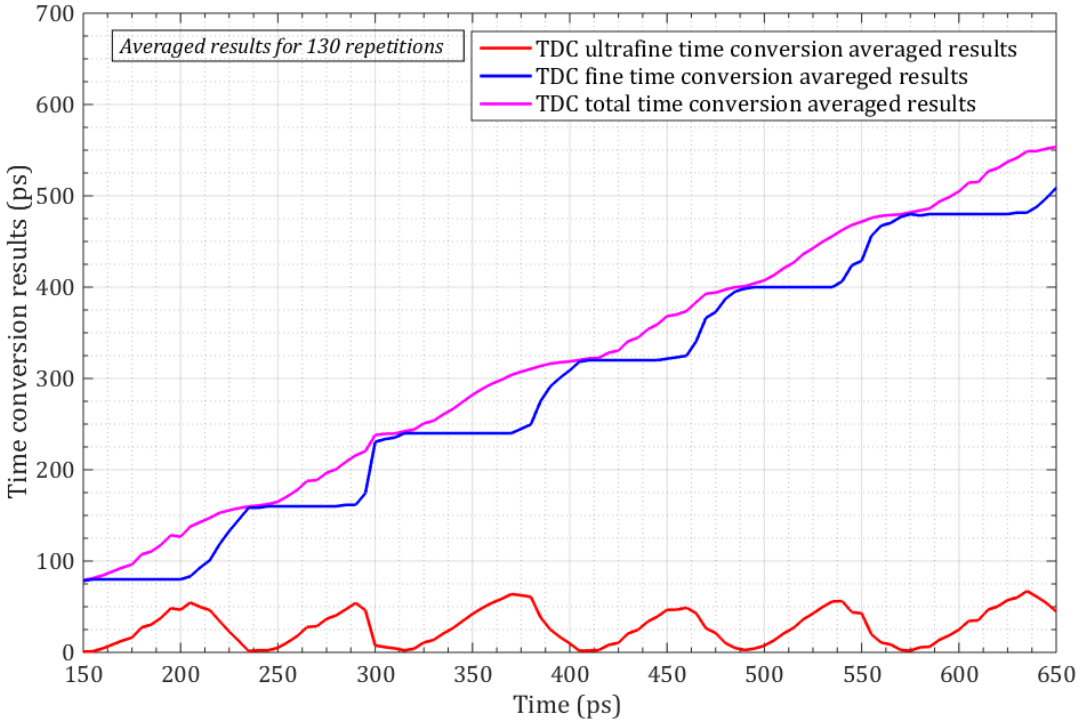


Figure 184 – valeurs moyennées des résultats des mesures sur une intervalle de 500 ps pour les conversions de temps fine, ultrafine et totale

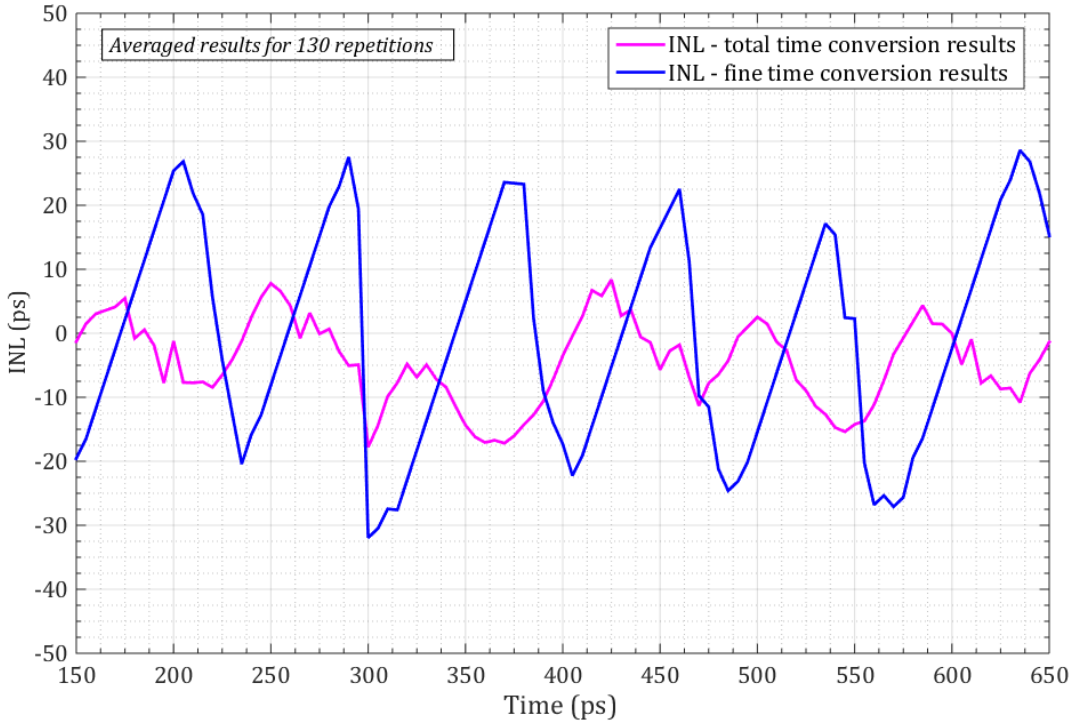


Figure 185 – la non linéarité intégrale (INL) estimée pour la conversion fine et la conversion totale sur une intervalle de 500 ps.

8.3 Conclusion

Cette thèse visait à réaliser une structure de test d'un CBF intégré basée sur un système TCSPC (TCSPC-SC). La première étape dans la réalisation de la TCSPC-SC consistait à concevoir une SPAD en technologie CMOS standard pour être utiliser comme détecteur de photon unique dans le système TCSPC. 6 Structures de SPAD avec 8 variations de diamètre ont été mises en œuvre dans une technologie 180 nm CMOS standard. Après un processus de caractérisation préliminaire, la structure SPAD la plus performante a été sélectionnée et entièrement caractérisée en termes de bruit, résolution temporelle et sensibilité. Les mesures ont montré une DCR inférieure à 5 kHz à 15 ° C avec une probabilité de post-déclenchement de ~ 0,2%, un PDP maximum de 20% pour λ=430 nm et une gigue temporelle inférieure à 100 ps FWHM pour λ = 450 nm et Vex= 500 mV. Suite aux résultats de caractérisation, une SPAD avec un diamètre de surface active de 15 µm, un taux de remplissage de 69%, un DCR inférieure à 2 kHz et un PDP de 20% @ 450 nm pour une polarisation supérieure à 600 mV a été choisie pour être utilisée comme détecteur de photon unique dans le TCSPC-SC. Ensuite nous avons démontré une architecture TDC hybride spécialement conçu pour être utilisé dans le TCSPC-SC grâce à une conception visant à limiter la consommation d'énergie et la surface d'occupation. Ceci a été réalisé grâce à une architecture qui combine le TAC analogique et les architectures numériques utilisant DLL et compteurs numérique. Le TDC hybride a été conçu pour fonctionner avec des fréquences d'entrée variables, l'opération à vitesse maximale de 400 MHz a résulté en une résolution temporelle de 10 ps LSB, un INL estimé à ~ 5.6 rms ps, une dynamique de 10 µs et un temps mort de 15 ns. La consommation électrique maximale estimée était de 44 mW pour un taux de conversion de 20 MHz avec 35 mW consommé par le DLL global et en 9 mW consommé par chaque TSU. Suite à la conception des deux blocs principaux, une structure de test TCSPC-SC a été réalisée dans une technologie 180 nm CMOS standard. Le circuit final intègre une seule ligne de Macropixels intégrant une SPAD comme un détecteur de photons unique dans chaque unité avec un taux remplissage de 69% atteint grace à l'approche « Streak ».