Direct time-of-flight SPAD image sensors for light detection and ranging

Présentée le 15 janvier 2021

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

pour l'obtention du grade de Docteur ès Sciences

par

Preethi PADMANABHAN

Acceptée sur proposition du jury

Prof. D. Atienza Alonso, président du jury
Prof. E. Charbon, directeur de thèse
Dr D. Stoppa, rapporteur
Dr J. Hurwitz, rapporteur
Prof. A. P. Burg, rapporteur

■ École
polytechnique
fédérale
de Lausanne

2021

*Thinking should become your capital asset,
no matter whatever ups and downs
you come across in your life.*
— A. P. J. Abdul Kalam

Acknowledgements

It definitely feels like time flew in the blink of an eye as this journey towards finishing a PhD is coming to an end. This journey would not be as fruitful and rewarding if it was not for the many wonderful people I met. Their help and support cannot be quantified, however, I would like to express it here as my memento in writing.

I would like to express my gratitude to my advisor, Prof. Edoardo Charbon, without whom I would not have embarked on this journey. For giving me this excellent opportunity to work on challenging, yet exciting projects, I will always be grateful. I will be thankful for the freedom he gave to explore bold ideas which helped broaden my own creativity as a designer.

At this point, I would like to express my heartfelt gratitude to Dr. Claudio Bruschini. Although, he was not officially part of my PhD project, he still ensured to keep in touch with my research regularly. Next, I would like to thank Dr. Shouleh Nikzad and Dr. Bruce Hancock from NASA's JPL where I had also previously interned. I spent early months of my PhD characterizing a readout circuit I had designed for their GaN detectors. Many thanks go to Shouleh who ensured I had a thoroughly enriching experience throughout the project timeline. I would like to thank Bruce for his valuable insights on circuit design, from theory to practice, all of which helped me approach problems systematically.

Special mention and my huge thanks will go to those colleagues of mine with whom I spent a lot of time during my PhD. From my Master thesis up until 2 years of my PhD, I worked with Augusto Ximenes. I co-designed my first DTOF sensor along with him. I am grateful to him for helping me kick-start on that project, to show me what a good layout was and also for being available to hold interesting discussions on design. My next heartfelt gratitude goes to Chao Zhang, who had a significant contribution in my entire PhD. I have had the most interesting discussions with him on aspiring to design the next-best DTOF sensor. His calm demeanor even when time-crunched always inspired me. I cannot forget how he spotted a calibration error, just moments before submitting our work to a conference and also ended up fixing it without panicking. I learnt a lot of digital design from him. He has been involved in my first and last projects and provided me the best possible co-working experience. Next, I would like to thank Marco Cazzaniga for his timely help in the implementation of the TDC in my second sensor and as well, for his patience and calmness throughout our collaboration. I would like to thank Baris Can Efe for helping me out during the measurement of my last chip.

Augusto Carimatto, another colleague of mine, once told me, 'You don't expect your colleagues to become your friends. But if they do, its a bonus!' Well, I was glad to receive that bonus and Augusto Carimatto became my first friend in the lab. I am always motivated by his ever-positive attitude,

Acknowledgements

Abstract

Depth sensing is an increasingly important feature in many applications of consumer, automotive, augmented/virtual reality (AR-VR), space and bio-medical imaging. Long range, high depth resolution, high spatial resolution, and high frame rates are often conflicting requirements and difficult to be simultaneously achieved due to extreme operating conditions. Direct time-of-flight (DTOF) has evolved to becoming a powerful technique to perform light detection and ranging (LiDAR). Thanks to advances in low-jitter optical detectors, such as single-photon avalanche diodes (SPADs), and accurate chronometers like time-to-digital converters (TDCs), picosecond timing resolution is possible, thus enabling millimetric depth resolutions.

High ambient light is an inevitable challenge in LiDAR applications, whose levels may exceed up to 100 klux on a bright sunny day, making it particularly challenging to detect a target submerged within an overwhelming noise floor. High ambient light operation can be accommodated by means of optical filtering, a higher laser power or temporal filtering techniques. Optical filtering is often restricted to a narrow, 10-50 nm bandwidth, insufficient at high ambient light levels. Higher laser power is not always possible, due to eye safety regulations and power constraints. Temporal filtering such as time gating and coincidence detection can thus be powerful tools to cope with high ambient light.

This thesis focuses on the design of DTOF sensors for LiDAR. To that end, two SPAD-based DTOF sensors are designed. The first sensor is designed in a 3D-stacked 45/65 nm CMOS technology, thus, enabling a modular architecture where the module itself comprises of 8×16 pixels. With a 60 ps-resolution TDC at its core, the sensor provides centimetric accuracy up to 300 m range in free space. The second sensor, named *Jatayu*, advances the previous design by hosting 256×128 pixels, thereby, significantly improving on its spatial resolution. While retaining its modularity, *Jatayu* also enables multi-level coincidence detection and progressive time-gating to suppress background light. To the best of the author's knowledge, progressive gating has been implemented in a LiDAR for the first time in this thesis. Designed in a 3D-stacked 45/22 nm CMOS technology, the sensor achieves under 7 cm accuracy over 100 m ranging and 10 klux background light. With its capability of acquiring 128×128, 3D depth maps of high dynamic range scenes, *Jatayu* is highly suitable for a variety of imaging applications in many different scenarios.

Key words: LiDAR, depth sensing, time-of-flight (TOF), CMOS image sensors, single-photon avalanche diode (SPAD), background illumination reduction, coincidence detection, progressive gating

Résumé

La détection en profondeur s'avère importante dans différentes applications : de l'automotive, réalité augmentée et virtuelle (AR et VR), le spatial ainsi que l'imagerie biomédicale. Longue portée, haute définition spatiale ainsi que fréquences d'images élevées sont souvent difficiles à atteindre simultanément vu les conditions de fonctionnement extrêmes. Répondant aux exigences de la résolution de la portée et de la profondeur, l'imagerie à résolution temporelle et, en particulier, le temps de vol direct (DTOF) a évolué pour devenir une puissante technique de détection et de télémétrie de la lumière (LiDAR). Un capteur DTOF se compose d'une source d'éclairage pulsée telle qu'un laser, éclairant une cible d'intérêt et les photons réfléchis rebondissant sur la cible sont ensuite horodatés par des détecteurs optiques. Grâce aux avancées en matière de détecteurs optiques à haute précision, tels que les diodes d'avalanche mono-photon (SPADs), et les chronomètres précis comme les convertisseurs temps-numérique (TDCs), la résolution de synchronisation picoseconde est possible, permettant ainsi des résolutions de profondeurs millimétriques. La lumière ambiante élevée est un défi inévitable dans les applications LiDAR, dont les niveaux peuvent dépasser 100 klux pendant une journée ensoleillée, ce qui rend particulièrement difficile la détection d'une cible immergée dans un plancher sonore accablant. La détection en présence de lumière de fond élevée peut être atténuée par des moyens de filtrage optique, une puissance laser plus élevée ou des techniques de filtrage temporel. Le filtrage optique est souvent limité à une bande passante étroite de 10 à 50 nm, insuffisante à des niveaux de lumière ambiante élevés. Une puissance laser plus élevée n'est pas toujours possible, en raison de la réglementation sur la sécurité oculaire et contraintes de puissance. Le filtrage temporel comme la synchronisation temporelle et la détection de coïncidences peuvent être des outils puissants pour faire face à la lumière ambiante élevée.

Cette thèse se concentre sur la conception de capteurs DTOF qui maintiennent les exigences LiDAR à l'avant-garde tout en faisant progresser l'état de l'art. À cette fin, deux capteurs DTOF basés sur SPAD sont conçus. Le premier capteur est conçu dans une technologie CMOS 3D empilée de 45/65 nm, permettant ainsi une architecture modulaire où le module lui-même se composait de 8×16 pixels. Avec une résolution de 60 ps TDC à son centre, le capteur fournit une résolution de profondeur centimétrique tout en démontrant jusqu'à 300 m de télémétrie dans l'espace libre. Le second capteur, nommé *Jatayu*, fait progresser la conception précédente en implementant 256×128 pixels, améliorant considérablement sa résolution spatiale. Tout en conservant sa modularité, Jatayu permet également la détection de coïncidences multi-niveaux et la synchronisation temporelle progressive pour supprimer de l'illumination de fond. À la connaissance de l'auteur, la mise en œuvre progressive de la synchronisation est la première à être effectuée pour un scénario LiDAR au sein de cette thèse. Conçu dans une technologie CMOS empilée en 3D de 45/22 nm, le capteur

Résumé

atteint une précision inférieure à 7 cm sur une portée de 100 m et une lumière de fond de 10 klux. Avec sa capacité à acquérir 128×128, des cartes de profondeur 3D de scènes à plage dynamique élevée, *Jatayu* est parfaitement adapté à une variété d'applications d'imagerie dans de nombreux scénarios différents.

Mots clés : LIDAR, Détection de profondeur, temps de vol (TOF), capteurs d'image CMOS, diode avalanche mono-photonique (SPAD), réduction du bruit de fond, détection des coïncidences, synchronisation progressive.

Contents

Contents

List of Figures

List of Tables

1 Introduction

This chapter introduces time-resolved imaging and in particular, time-of-flight (TOF) for light detection and ranging (LiDAR) applications. Taxonomy of depth sensing techniques is presented where the primary discussion is held on direct-TOF sensors which is the focus of this thesis. The reader is acquainted with various challenges in a LiDAR application, which thereby set the premise for this thesis. Finally, significant contributions of this work are described, which along with the thesis organization, provide the necessary context for this work.

1.1 Time-resolved imaging

Time-resolved imaging is a composition of techniques which exploits the temporal information of photons to infer a scene of interest. This information itself can be of various types depending on the application. With its vast scope, time-resolved imaging can be applied in a number of areas including but not limited to consumer, automotive, computer vision, gaming, augmented (and virtual) reality (AR-VR), space and bio-medical applications [1, 2, 3, 4, 5]. High timing resolution on the order of picoseconds is attainable in today's time-resolved image sensors, thereby, opening new pathways in 3D vision and sensing in consumer applications. Time-of-flight (TOF) is a key depth-sensing technique where the travel time of photons is used to estimate the distance of targets and reconstruct them over a scene of interest. In both, consumer and automotive applications, TOF can be used to extract information such as distance and in turn, reconstruct a 3D point cloud of a scene. Practical examples can be found in today's smart phones, in advanced driver assistance systems (ADAS) and autonomous driving, while in biomedical applications, time-resolved imaging may be used to extract the lifetime of fluorophores in fluorescence lifetime imaging (FLIM) and time of arrival of gamma photons in positron emission tomography (PET).

1.2 Time-of-flight (TOF) for LiDAR

Light detection and ranging (LiDAR) is a method for measuring distances using light. Various depth sensing techniques enable this measurement, a broad classification of which is shown in Figure 1.1,

where the focus of this thesis is highlighted (DTOF). Optical depth sensing techniques are broadly classified in their active and passive forms. Advantages of passive systems are that they do not need any active illumination. For example, a typical stereo vision system functions with just two cameras separated by a known distance where it simulates a human binocular vision and uses sophisticated computer algorithms to reconstruct 3D images. Naturally, the downfall is its heavy dependence on computationally intensive processing. Depth-from-focus calculates distances by modeling the quality of images by optimally choosing the camera focal setting for every point in the scene to capture the best possible image [6]. The only major advantage is the presence of a single camera when compared to a stereo vision system employing two cameras. The major drawback of both methods however is inability to detect in scenes of poor contrast, i.e., it does not work on blank walls or in the dark. In another passive technique, called, light-field, a main lens is chosen to select the desired field-of-view (FOV) and create an intermediate image in front of a micro camera array where each micro camera in the image sensor sees a slightly different perspective of the target. The images generated by the camera in this setup are processed using software algorithms to calculate the depth of the scene. An important concern however is the dependence of depth resolution on the depth-of-field, which in turn depends on the focal length of the main lens[7, 6].

Active depth sensing techniques, as the name suggests, employ an active illumination source to perform 3D imaging. Interferometry is an active technique which provides depth resolution on the order of nanometers, where the interference fringe of the backscattered laser beam is measured with respect to a reference beam . The technique is however limited in the achievable range which is on the order of millimeters. Furthermore, the dependence of laser wavelength on the environmental conditions introduces additional burden on calibration [8, 9].
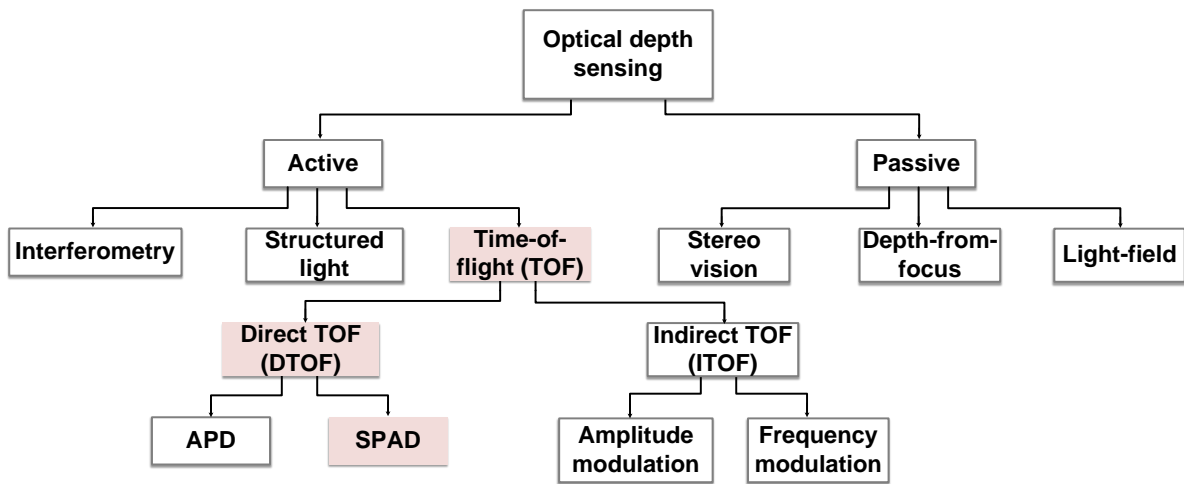
Figure 1.1 – Classification of optical depth sensing techniques.

Structured light is another active technique where a pattern (of dots or stripes) is projected onto a target and its deformation is used to reconstruct the depth and the target shape. A recent example of this technique is the Face ID feature on Apple's iPhone [10]. While guaranteeing high accuracy over short distances, the disadvantage of structured light is for medium-long-range measurements where the deformation of the projected pattern may not be perceived easily. Furthermore, high

background noise from ambient light may also interfere with the projected pattern, additionally making it challenging to acquire data from the target. As a result of this, structured light technique is popularly used for distances under 1–2 m.

Another active method, time-of-flight (TOF), as briefly introduced before, relies on the travel time of light (pulses or waves) bouncing off a target to measure distances. TOF sensors are basically categorized into their indirect (ITOF) and direct (DTOF) forms.

In the (amplitude-)modulated (AM) ITOF operation, the emitted signal is a continuous-wave signal modulated in time, which is usually a sinusoidal signal. The phase difference between the emitted and received light signals, is used to measure the distance traveled by the light, from the sensor to the target and back again as seen in Figure 1.2a. Consequently, the distance, D_{mod}, can be expressed as,

$$D_{mod} = \frac{c\Delta\Phi}{4\pi f_{mod}},$$

(1.1)

where, f_{mod} is the modulated frequency and c is the speed of light. Pulsed method is one of the other ways to perform an ITOF measurement- a concept of which is shown in Figure 1.2b. An example case with 4 windows is shown where the measurement is based on the gated integration of the optical pulses over time [11], where multiple windows allow measurement of the signal (A in Figure 1.2b) and background noise level (B in Figure 1.2b) separately. All the measurements have a duration equal to that of the emitted pulse and 4 windows are used to determine the phase shift. Like in the modulated method, the distance, D, can be calculated as follows,

$$D = \frac{cT_{oF}}{2}.$$

(1.2)
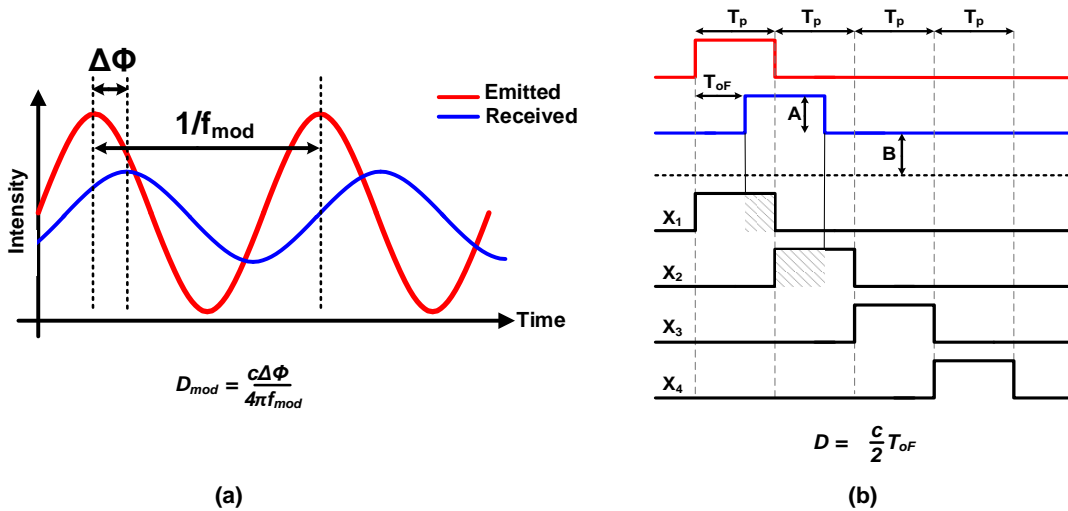
Figure 1.2 – Conceptual representation of ITOF sensors- (a) Modulated ITOF and (b) Pulsed ITOF.

In a frequency-modulated (FM) ITOF system, the optical frequency is modulated in time where,

a frequency difference between the emitted and reflected signal is processed to determine the distance of a target [12, 13]. In addition to measuring distances, this technique can also provide a velocity measurement using Doppler effect, which can be particularly useful in automotive LiDAR applications. Unlike AM technique, FM-based method is able to cope with multi-path reflections by resolving the multiple beat tones in the frequency domain.

ITOF sensors have been implemented in a number of consumer applications for ranging and depth mapping. However, continuously growing demand for higher spatial resolutions (VGA – MP) and operation over wide FOV ($50°$ – $120°$) have limited their applications to short ranges [14, 15, 16, 17]. A class of ITOF sensors based on short-pulse modulation and multi-tap lock-in pixels is becoming an attractive candidate due to higher achievable range resolution [18], however, it is currently limited to distances under 10 m [19]. Another drawback of ITOF sensors is their limited ability to distinguish two nearby objects (multi-path interference) [20].

DTOF sensors, on the other hand, are able to mitigate these challenges with detection ranges reaching up to several hundred meters [21, 22], principally determined by the available optical power and their innate ability to discriminate multiple echoes easily [23]. The main focus of this thesis is on the analysis and design of DTOF sensors for LiDAR applications and therefore, the subsequent sections will be dedicated to depth sensing based on DTOF only. The high-level block diagram of a DTOF sensing system is shown in Figure 1.3.
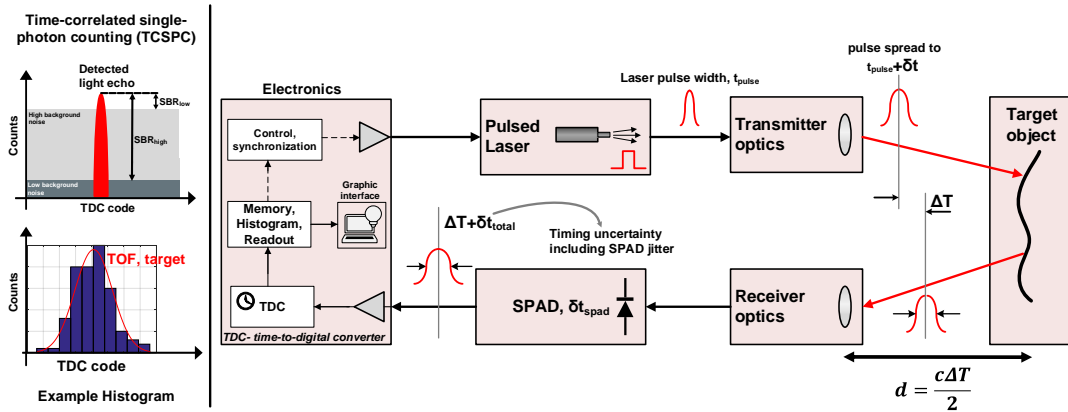
Figure 1.3 – High-level block diagram of a DTOF sensing system

A DTOF sensor consists of a pulsed illumination source which may be a LED, a laser diode or a VCSEL (vertical cavity surface-emitting laser) array, operating at the desired repetition rate which is usually dictated by the maximum distance required to be measured. This source is illuminated on a scene of interest with the target and the reflected photons from the target are then detected by an appropriate photodetector, typically an avalanche photodiode (APD) or a single-photon avalanche diode (SPAD). The time-of-arrival of these photons is then measured by a time-stamping electronic circuit. In DTOF sensors, time-to-digital converters (TDCs) are typically used for this purpose [24, 25, 22, 26, 27, 28]. On combining timing information with the speed of light, c, the distance of the target from the sensor is determined as shown in Figure 1.3. A typical example is a commercial range-finder where single-point timing information is utilized to measure ranges up several tens of

meters. When such a single-point timing measurement is spatially extended over more number of points over a given field-of-view (FOV), a 3D image can then be reconstructed of the scene within the FOV.

1.3 LiDAR application challenges

A LiDAR system, depending on whether it is employed outdoor or indoor, requires a sensor capable of measuring ranges between 10 – 100 m. A close-in LiDAR may require accuracies down to a few millimeters while long-distance LiDARs can as well work with accuracies on the order of a few centimeters. [14, 21, 29]. Nonetheless, accuracy and range requirements must be met over a wide range of operating conditions. The following sections briefly discuss some of them.

1.3.1 Background noise suppression

Time-correlated single-photon counting (TCSPC) is a common method utilized in DTOF systems to acquire large number of detections where the detected signal is represented as a histogram corresponding to the time-of-arrival of individual photons incident on the photodetector [30, 24, 23, 25, 29]. Under an ideal condition where the background noise is low, the target peak can be easily distinguished on the measured histogram. However, most often, high background noise from ambient sunlight is a primary challenge in LiDAR applications. Depending on whether the system is indoor or outdoor, the background noise may range around 1 klux in a well-lit indoor environment while reaching up to 100 klux on a bright sunny day outdoor [31]. As a result of this, often, the returning target peak in a DTOF system is submerged under an overwhelming noise floor (see Figure 1.3), thus making it extremely difficult to detect the signal. Another implication of high background noise is distortion in the measured histogram due to pile-up [30], which results in large depth errors. Therefore, it is paramount to address the high background noise challenge to maintain a signal-to-background noise-ratio (SBR) high enough to allow quality depth measurement. Incorporating noise-filtering techniques, both, optically, by using optical bandpass filters and electrically, by smart sensor design are both valuable in this regard.

1.3.2 Optical power budget and safety regulations

An ideal and convenient solution to coping with high background noise is to increase the power of the illumination source. However, the maximum permissible optical power of the laser will be dictated by eye-safety regulations which are in turn governed by various other system parameters such as the laser beam size and divergence, the illumination wavelength, exposure time, the FOV, optical filter, lens etc. In pulsed laser systems, as is the case in this thesis, pulse energy, pulse repetition rate, beam size and divergence are all accounted for while defining the allowable optical power from a laser source. Most often, the constraints become stringent for visible and near-infrared (IR) wavelengths, where the human eye is more sensitive to and consequently, is prone to more harm at higher optical powers. Hence, wavelengths beyond near-IR (> 1500 nm) or in the UV

region (200 - 350 nm) can permit higher optical powers. Maximum permissible exposure (MPE) is a term commonly used to indicate the highest power or energy density (in W/cm^2 or J/cm^2) of a light source that is considered safe or has a negligible probability for creating damage. A calculation of MPE is important to account for eye-safety. IR wavelengths beyond 1500 nm are absorbed by the transparent parts of the eye before reaching the retina. As a result of this, MPE for these wavelengths can be higher than for visible light, thus, making them favorable for LiDAR applications. Furthermore, with an average solar radiation on the earth's atmosphere of around 1361 W/m^2, the solar irradiance is higher at visible wavelengths while starting to significantly drop from 900 nm ($\geq$ near-IR wavelengths) [32]. This factor further motivates the choice of near-IR to IR lasers for LiDARs as they help improve the overall system SBR. Nevertheless, depending on the exposure time and the wavelength, MPE can be used to calculate the permissible optical power on the laser. While it is not in the scope of this thesis to discuss this further, however, it is noteworthy to regard MPE as an important factor while implementing LiDAR systems. The reader is directed to [33] for more information on laser safety regulations published by American National Standards Institute (ANSI).

1.3.3 Laser interference

A pulsed-LiDAR system relies on the reflected laser pulse to estimate the time-of-arrival of photons from the target. There is often a possibility of blinding the sensor with undesired photons which return to the sensor. Also, since most TOF sensors are designed to detect the first return, blinding effect could easily prevent any further legitimate detection. Further, multiple LiDAR systems can coexist while appearing as interferences to each other. A pictorial representation of such a scenario is shown in Figure 1.4, where, in the worst case, two laser sources from two different LiDAR systems could have the same repetition rate (and/or synchronized) leading to depth errors. It is important to enable the sensor to deal with such scenarios so that the target return can still be correctly estimated. Prior solutions include techniques based on code-division multiple-access techniques (CDMA) [34] and using pseudo-random sequences of the illumination to improve robustness in a multi-camera environment [35]. This thesis introduces a simpler method based on digital pulse-position modulation, the details of which will be covered in Chapter 3.

1.3.4 High-dynamic range scenes

In any DTOF-based LiDAR system, the timestamp of the photons bouncing off a target is used to reconstruct a depth map. A given FOV may contain multiple targets with various (and wide range) surface reflectivity. Due to the nature of detection, a target with higher reflectivity is bound to return more number of photons when compared to target with lower reflectivity, thereby favoring the detection of the former over a given measurement window. However, in a scenario such as this, it is important to capture all the objects in the scene without causing depth errors. An outdoor example of such a scenario is seen in an automotive LiDAR system where the presence of retro-reflectors and traffic signs must be detected along with other objects (such as pedestrian walking or a lamp post) within the FOV. Therefore, a DTOF system should be equipped to handle the dynamic range
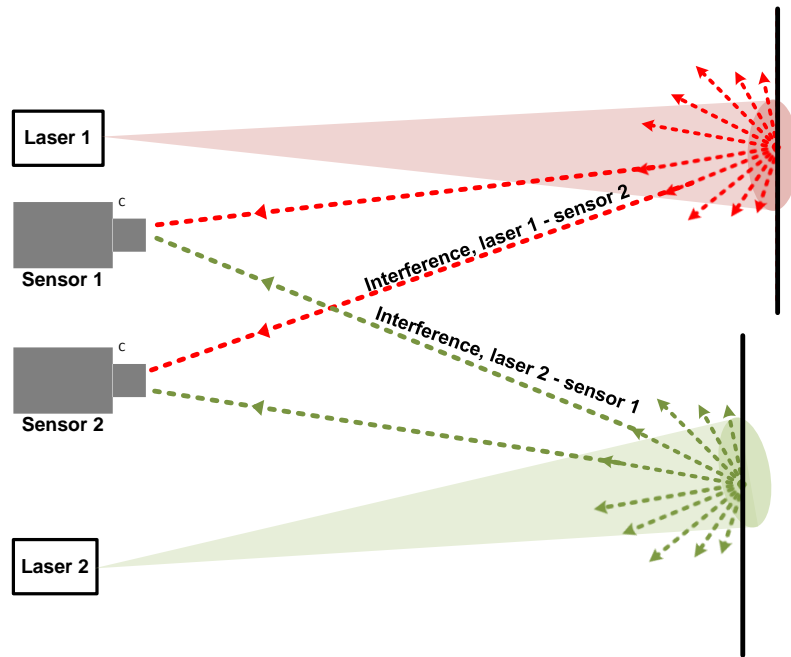
Figure 1.4 – Laser interference scenario- Pictorial representation with two LiDAR systems with Lambertian targets as example.

of a given scene in the presence of ambient light over different operating distances.

1.3.5 Adverse weather phenomena

In addition to imaging under bright sunlight, other weather phenomena such as the presence of rain, fog or cloud, also adds to the detection challenge due to their scattering and absorption properties. In particular, interaction of light with suspended particles can manifest itself differently from the background noise which appears more as a uniform distribution on the histogram. Therefore, scattering effects need to be modeled and appropriate sensor solutions need to be developed. The non-uniform temporal distribution [36, 37] of particles in fog or cloud may reflect as distinct peaks in the acquired histogram, causing depth errors sometimes. Gated imaging is one way to selectively eliminate unwanted peaks occluded by suspended matter in light propagation path [38]. Furthermore, gating can enable range-selective detection while retaining high SBR around the target of interest. Exploiting the benefits of gating, this thesis also proposes a new method based on progressive gating as a way to improve SBR; more details will be discussed in Chapter 4.

1.3.6 Detector sensitivity

Photodetectors capable of measuring single photons have existed for several decades in time-resolved systems implemented out of photomultiplier tubes (PMTs) and micro-channel plates (MCPs) [39]. However, owing to their bulky nature, vacuum-based operation, complexities due to their high

voltage requirements (several kVs) and high costs, their applications have become limited over the years. Today, time-resolved systems have evolved into their solid-state forms and to that front, Geiger-mode single-photon avalanche diodes (SPADs) have emerged as promising photodetectors in such systems [40, 41]. High speed and picosecond timing resolution achievable in SPADs make them popular candidates for DTOF sensors which require accurate timing measurements.

The performance of the DTOF sensor is primarily dictated by the SPAD performance. Therefore, with respect to sensitivity, the most important SPAD parameter is the photon detection probability (PDP), which represents the probability of producing an avalanche in the device in response to the photon absorption at a given wavelength. In CMOS SPADs, the PDP usually peaks in the visible region reaching up to 70% for single devices [42, 43]. However, as discussed before, for a LiDAR application, a near-IR/IR sensitivity is favored due to lower constraints on the permissible optical power of the illuminator and also, potentially higher SBR.

Moving from individual SPADs to image sensors introduces other parameters to be considered to enhance sensitivity. Fill factor is one such factor, which represents the ratio of photosensitive area to the total pixel area. The SPAD's fill factor directly affects the overall photon sensitivity, given that it is multiplied by the PDP to give the overall photon detection efficiency (PDE). Fill-factor effects are more pronounced in a monolithic (2D) implementation of sensors, where the active area is also shared with pixel-circuitry thus reducing the achievable fill-factor, as pictorially seen in Figure 1.5a,b.

(a)   **(b)**   **(c)**
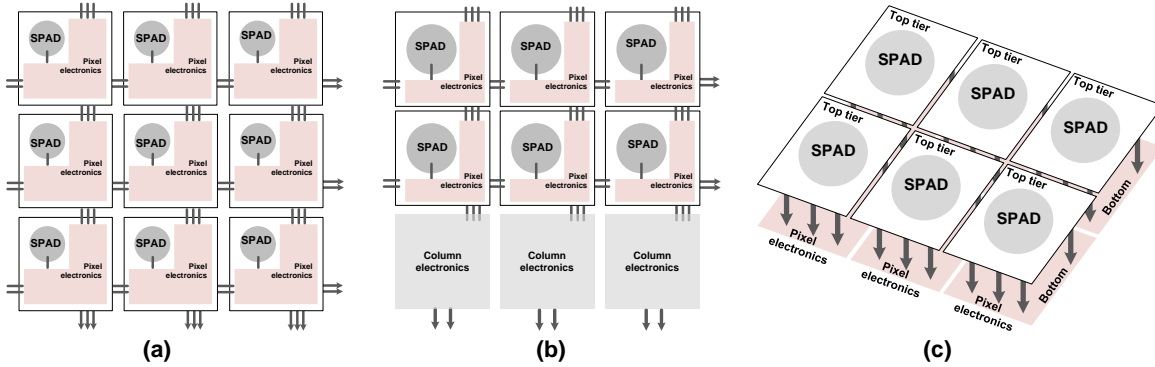
Figure 1.5 – Comparison of imager architectures- (a) 2D array, per-pixel architecture, (b) column-parallel architecture and (c) 3D-stacked architecture.

The advent of 3D-stacked technology has however, helped circumvent this issue to a large extent. In particular, today's 3D-stacked sensors enable much higher fill-factor (and therefore, PDE), by dedicating separate tiers optimized for SPADs and electronics independently, see Figure 1.5c. Consequently, a more advanced sensor functionality is possible without compromise on the fill-factor. Furthermore, a much better near-IR sensitivity, reaching between 10–15 % around 850 nm, makes them very suitable for LiDAR applications [44, 45, 28]. Figure 1.6 shows the PDP of some state-of-the-art 3D stacked SPADs to given an overview of the achievable PDP spectra in BSI SPADs.

Another LiDAR-relevant SPAD parameter is the dark count rate (DCR) which denotes the uncorre-

lated avalanche events recorded in the absence of light, usually represented in counts/second (cps). As long as the DCR is lower than the background noise rate incident on the SPAD, it is usually not a major issue. Most often the background noise events per pixel is orders of magnitude (at least 2–3 orders of magnitude) higher than the DCR.
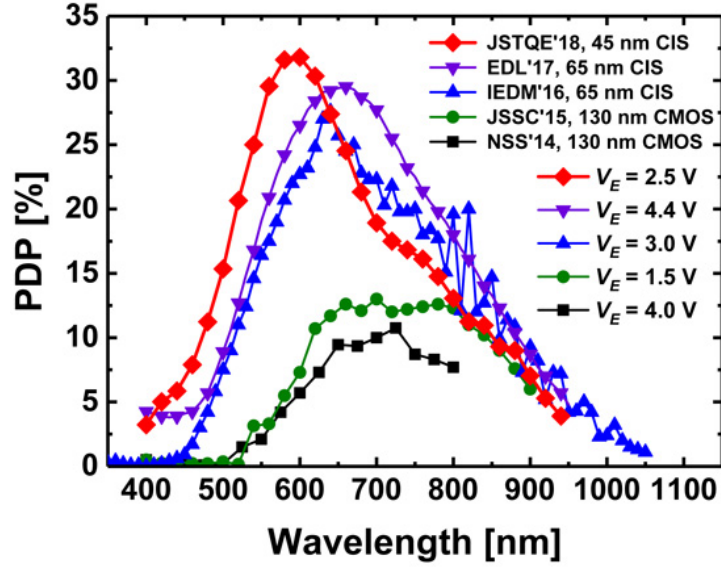
Figure 1.6 – PDP comparison of state-of-the-art back-illuminated SPADs [45].

Furthermore, it is always desirable to have high dynamic range, particularly required in high background noise scenario seen in LiDAR. Consequently, low SPAD dead times are favored to allow more number of detections. However, achieving lower dead times can be aimed for, so long as it does not significantly increase the afterpulsing effect, which refers to the triggering of avalanches due to release of trapper carriers at a later time, which results in false correlations in the measurement [46].

In summary, all the relevant SPAD characteristics should already be optimized for during the design phase based on the application.

1.3.7 Improving timing statistics

As mentioned earlier, SPADs exhibit a timing response characterized by a low timing jitter, usually expressed as full width at half maximum (FWHM). State-of-the-art 3D-stacked SPADs achieve FWHM on the order of 100 ps [47, 44], thus, enabling them for millimetric precision required in LiDAR applications. In addition to the aforementioned SPAD jitter, DTOF systems usually face, multiple other timing uncertainties (see Figure 1.3). A chronometer, like a TDC, used in a DTOF system timestamps the detected photons with a timing uncertainty determined by its root-mean-square (RMS) quantization error ($\sigma_{TDC} = TDC_{res}/\sqrt{12}$). The computed histogram after TCSPC is representative of all the sources of timing uncertainties (shown as offsets or delays, $\Delta T + \delta t_{total}$ in Figure 1.3) arising from the laser pulse, the SPAD and any electrical circuit through the propagation of the detected event. The total timing uncertainty , σ_{total}, is then given by the summation of the

RMS values of the individual contributors assuming that they are all statistically independent.

$$\sigma_{total} = \sqrt{\sigma_{laser}^2 + \sigma_{SPAD}^2 + \sigma_{TDC}^2 + \sigma_{other}^2}, \tag{1.3}$$

where the component, σ_{other} accounts for jitter from any additional electronic circuitry through the propagation of the photon-event such as a combination tree which may combine events from multiple pixels and the surrounding logic. Typically, the contribution of σ_{laser}, σ_{TDC} and σ_{other} can be considered negligible compared to the SPAD jitter, σ_{SPAD} and as a result of this, most often the SPAD jitter is the dominant contributor to the achievable timing performance. Therefore, it is desirable to minimize this parameter during the design phase in order improve the timing resolution of the whole DTOF system to provide better depth accuracy required in LiDARs.

1.3.8 Data rate

Another major challenge in DTOF image sensors is the large volume of data being generated which directly scales with the size of the sensor array. Due to limited readout (and I/O) bandwidth, on-chip processing is usually required to maintain a reasonable frame rate [23, 28, 27]. For example, integrated histogramming implemented in [27] achieved up to 14.9:1 data compression ratio using partial histogramming techniques. Nevertheless, there is a continuous need for resource-optimal ways of implementing histogramming on chip.

1.4 LiDAR implementation- scanning vs. flash

LiDAR systems can be implemented in scanning or flash modes of operation. The pictorial differentiation of the two forms in shown in Figure 1.7. Scanning LiDAR typically consists of at least a laser and a detector, which are mounted on a rotating or vibrating scanner [25, 48]. Scanning LiDARs usually benefit from an increased signal-to-background noise ratio (SBR) due to higher achievable optical power while scanning over only sections of the target FOV, see Figure 1.7b, where the horizontal FOV is scanned section-wise. However, the presence of moving mechanical parts in conventional scanning systems introduces additional compexity and also raises long-term reliability concerns. Recent advancement in solid-state scanning LiDARS with MEMS-based mirrors are achieving smaller form factors with better design and fewer optical components. A very recent example is another product from Apple's IPad Pro with a LiDAR scanner, released in 2020 [49].
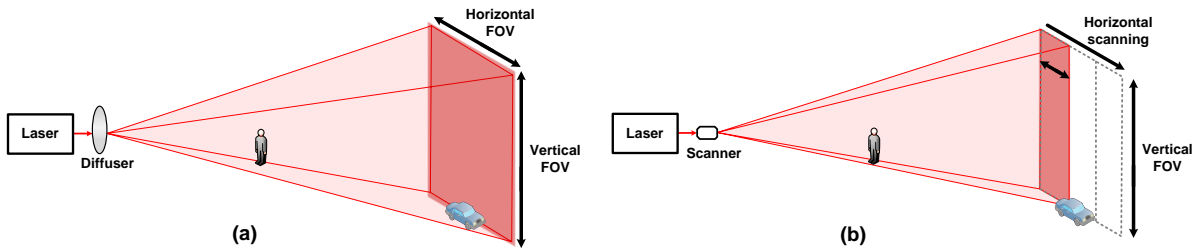
Figure 1.7 – Flash vs. scanning LiDAR conceptual representation.

Now, flash LiDARs on the other hand, benefit from a much simpler system by illuminating the entire FOV simultaneously (see Figure 1.7a), the drawback however being lower achievable SBR at longer ranges (> 20 m). Recent developments in illumination and optics have offered innovative solutions through VCSEL (Vertical Cavity Surface-Emitting Laser) array technology and laser diode arrays which help circumvent the low-SBR issue to some extent.

Nonetheless, for a quality LiDAR system, a trade-off between both methods have to be made. For long range measurements > 20–30m, a smaller spatial FOV is often used allowing concentration of higher energy density (of the laser) over the smaller area, and therefore an improved SBR. Additionally, a number of interdependent parameters have to be traded-off while framing target specifications, while ensuring that there is a continuous feedback between the DTOF sensor and the illuminator/optics system in order to provide the optimal condition required for signal detection.

1.5 Thesis contributions

This thesis is an attempt towards advancing the state-of-the-art DTOF sensors required for LiDAR applications. Multiple application challenges have been addressed where primary focus has been on **background noise suppression**. Two DTOF sensors have been designed and implemented in this thesis. The design and characterization of the first sensor is a collaborative work with an equal division of labor between the author and Augusto Ximenes. The author was responsible for the shared TDC design, modeling and analysis of injection locking including top-level assembly of the chip. The modular architecture using decision-tree with source preservation was developed by Augusto Ximenes. The second sensor was a collaborative effort between the author, Chao Zhang and Marco Cazzaniga. The author developed the entire coincidence-based architecture from modeling up to chip design. The chip was also entirely characterized by the author including firmware design for testing. The TDC was implemented by Marco Cazzaniga while the digital readout block was implemented by Chao Zhang.

Following is a summary of all the contributions made in this thesis.

The **first contribution** includes the implementation of a **GaN-based sensor** where the author designed a **CMOS front-end circuit with capacitive transimpedance amplifiers** to read out picoampere range photocurrents. In particular, the designed sensor mitigated the high reverse bias challenge, > 80 V, required in the GaN APDs used in this work. Chapter 2 elaborates on this work.

While different detector technologies are evaluated, the majority of this thesis, however, focuses on SPAD-based sensors. 3D-stacked back-illuminated SPADs are presented along with relevant characteristics for LiDAR application. It should be noted that focus of this thesis is not on the SPAD design itself, but on everything within the sensor design starting from pixel circuitry. Chapter 2 presents various SPAD front-end circuits designed for quenching and recharge.

The **second contribution** is on **resource-sharing** which is becoming almost inevitable as image sensors scale in size. To this end, modeling and analysis is presented to provide a power-efficient

solution to pixel sharing in Chapter 3. Based on the established concept, a **modular 8×16 DTOF sensor prototype** is designed in a 3D stacked 45 nm / 65 nm CMOS technology. A combination tree, called decision tree is shared between multiple pixels which acts as an arbiter, while propagating events based on a first-come-win-all policy. A shared TDC facilitates long-distance ranging and imaging required in LiDAR applications. Furthermore, laser interference challenge is addressed by proposing a technique based on pulse-position modulation where up to 18.6 dB interference suppression is achieved.

The **third contribution** of this thesis is towards a **noise-resilient DTOF architecture**. The previous sensor was limited to low-light operating conditions due to the absence of any noise-filtering techniques on chip. The author contributed in the development of a second DTOF sensor, named-*Jatayu* which is adapted to function under high ambient light due to new concepts implemented based on **multi-level coincidence detection and progressive gating.** The first step towards the design of *Jatayu* has been on **an analytical model** developed on MATLAB to thoroughly evaluate the new architecture. The findings from the simulation of the model are used to design a CMOS implementation of the sensor. Chapter 4 elaborates on this model along with the simulated results.

The **fourth contribution** includes the **design of *Jatayu*** based on the modeled concepts. *Jatayu* is a DTOF sensor with 256×128 pixels, implemented in a 3D stacked 45 nm / 22 nm CMOS technology. Resource sharing is similar to the previous sensor while a coincidence-based tree replaces the decision tree to manage multiple events whilst providing background noise suppression. The sensor demonstrates up to **100 m ranging under 10 klux background illumination**. Up to 7-level coincidence detection with tunable coincidence windows is implemented to provide activity-dependent imaging. Progressive gating is proposed to provide target-selective ranging while also electrically improving the SBR. Measurement results show up to 31 dB SBR improvement with this technique. Flash LiDAR images spanning over short-medium range are successfully acquired over wide-dynamic range targets.

Furthermore, a robust timing solution is proposed based on injection locking which significantly improves the timing jitter (and phase noise). The **fifth contribution** is towards developing a **phase macro-model** to thoroughly analyze the **injection locking concept** in ring oscillators to further aid in the design process. Chapter 5 describes this concept as well as the new sensor, *Jatayu*, which incorporates them.

1.6   Thesis organization

Chapter 2 presents two different detector technologies where a CMOS readout circuit is implemented for a GaN-based APD. The second half of this chapter introduces SPADs in 3D-stacked technology along with their characterization results. Chapter 3 proposes the concept of resource sharing to provide a power-efficient sensor solution. Following this, a modular DTOF sensor based on resource-sharing is presented. Chapter 4 describes an analytical model for an alternative DTOF sensor design addressing ambient light suppression. Multiple concepts based on coincidence

detection and gating are introduced where simulations on an example scenario are made. Chapter 5 builds on the aforementioned model, where a second DTOF sensor with 256×128 pixels is presented. In addition, this chapter also presents a robust timing solution for shared-TDC, DTOF architectures. Chapter 6 finally concludes this thesis along with recommendations for future work.

REFERENCES

[1] G. Yahav, G. J. Iddan, and D. Mandelboum, "3d imaging camera for gaming application," in *2007 Digest of Technical Papers International Conference on Consumer Electronics*, pp. 1–2, IEEE, 2007.

[2] E. Bastug, M. Bennis, M. Médard, and M. Debbah, "Toward interconnected virtual reality: Opportunities, challenges, and enablers," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 110–117, 2017.

[3] A. C. Ulku, C. Bruschini, I. M. Antolović, Y. Kuo, R. Ankri, S. Weiss, X. Michalet, and E. Charbon, "A 512× 512 spad image sensor with integrated gating for widefield flim," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1–12, 2018.

[4] I. Gyongy, N. Calder, A. Davies, N. A. Dutton, R. R. Duncan, C. Rickman, P. Dalgarno, and R. K. Henderson, "A 256×256, 100-kfps, 61% fill-factor spad image sensor for time-resolved microscopy applications," *IEEE Transactions on Electron Devices*, vol. 65, no. 2, pp. 547–554, 2017.

[5] C. Bruschini, H. Homulle, and E. Charbon, "Ten years of biophotonics single-photon spad imager applications: retrospective and outlook," in *Multiphoton Microscopy in the Biomedical Sciences XVII*, vol. 10069, p. 100691S, International Society for Optics and Photonics, 2017.

[6] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 5, pp. 972–986, 2011.

[7] C. Perra, F. Murgia, and D. Giusto, "An analysis of 3d point cloud reconstruction from light field images," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, IEEE, 2016.

[8] R. Dändliker, Y. Salvadé, and E. Zimmermann, "Distance measurement by multiple-wavelength interferometry mesure de distance par interférométrie à plusieurs longueurs d'onde," *Journal of Optics*, vol. 29, no. 3, p. 105, 1998.

[9] F. Li, J. Yablon, A. Velten, M. Gupta, and O. Cossairt, "High-depth-resolution range imaging with multiple-wavelength superheterodyne interferometry using 1550-nm lasers," *Applied optics*, vol. 56, no. 31, pp. H51–H56, 2017.

[10] Apple 2018. https://www.apple.com/iphone-xr/specs/.

[11] J. Illade-Quinteiro, V. M. Brea, P. López, D. Cabello, and G. Doménech-Asensi, "Distance measurement error in time-of-flight sensors due to shot noise," *Sensors*, vol. 15, no. 3, pp. 4624–4642, 2015.

[12] B. Behroozpour, P. A. Sandborn, N. Quack, T.-J. Seok, Y. Matsui, M. C. Wu, and B. E. Boser, "Electronic-photonic integrated circuit for 3d microimaging," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 161–172, 2016.

[13] C. Slinger and M. Harris, "Introduction to continuous-wave doppler lidar," *Summer School in Remote Sensing for Wind Energy, Boulder, USA*, vol. 11, 2012.

[14] C. S. Bamji, S. Mehta, B. Thompson, T. Elkhatib, S. Wurster, O. Akkaya, A. Payne, J. Godbaz, M. Fenton, V. Rajasekaran, *et al.*, "Impixel 65nm bsi 320mhz demodulated tof image sensor with 3μm global shutter pixels and analog binning," in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*, pp. 94–96, IEEE, 2018.

[15] C. Niclass, C. Favi, T. Kluter, F. Monnier, and E. Charbon, "Single-photon synchronous detection," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 7, pp. 1977–1989, 2009.

[16] 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.

[17] M.-C. Amann, T. M. Bosch, M. Lescure, R. A. Myllylae, and M. Rioux, "Laser ranging: a critical review of unusual techniques for distance measurement," *OptEn*, vol. 40, pp. 10–19, 2001.

[18] K. Yasutomi, Y. Okura, K. Kagawa, and S. Kawahito, "A sub-100μ m-range-resolution time-of-flight range image sensor with three-tap lock-in pixels, non-overlapping gate clock, and reference plane sampling," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 8, pp. 2291–2303, 2019.

[19] K. Yamada, K. Akihito, T. Takasawa, K. Yasutomi, K. Kagawa, and S. Kawahito, "A distance measurement method using a time-of-flight cmos range image sensor with 4-tap output pixels and multiple time-windows," *Electronic Imaging*, vol. 2018, no. 11, pp. 326–1, 2018.

[20] F. Remondino and D. Stoppa, *TOF range-imaging cameras*, vol. 68121. Springer, 2013.

[21] M. Perenzoni, D. Perenzoni, and D. Stoppa, "A 64 × 64-pixels digital silicon photomultiplier direct tof sensor with 100-mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6 km for spacecraft navigation and landing," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 151–160, 2016.

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

[23] N. A. Dutton, S. Gnecchi, L. Parmesan, A. J. Holmes, B. Rae, L. A. Grant, and R. K. Henderson, "11.5 a time-correlated single-photon-counting sensor with 14gs/s histogramming time-to-digital converter," in *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*, pp. 1–3, IEEE, 2015.

[24] C. Veerappan, J. Richardson, R. Walker, D.-U. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, *et al.*, "A 160× 128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter," in *2011 IEEE International Solid-State Circuits Conference*, pp. 312–314, IEEE, 2011.

REFERENCES

[25] C. Niclass, M. Soga, H. Matsubara, M. Ogawa, and M. Kagami, "A 0.18-μ m cmos soc for a 100-m-range 10-frame/s 200×96-pixel time-of-flight depth sensor," *IEEE Journal of solid-state circuits*, vol. 49, no. 1, pp. 315–330, 2013.

[26] D. Portaluppi, E. Conca, and F. Villa, "32× 32 cmos spad imager for gated imaging, photon timing, and photon coincidence," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 2, pp. 1–6, 2017.

[27] C. Zhang, S. Lindner, I. M. Antolović, J. M. Pavia, M. Wolf, and E. Charbon, "A 30-frames/s, 252×144 spad flash lidar with 1728 dual-clock 48.8-ps tdcs, and pixel-wise integrated histogramming," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 1137–1151, 2018.

[28] R. K. Henderson, N. Johnston, S. W. Hutchings, I. Gyongy, T. Al Abbas, N. Dutton, M. Tyler, S. Chan, and J. Leach, "5.7 a 256 × 256 40nm/90nm cmos 3d-stacked 120db dynamic-range reconfigurable time-resolved spad imager," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, pp. 106–108, IEEE, 2019.

[29] A. R. Ximenes, P. Padmanabhan, M.-J. Lee, Y. Yamashita, D.-N. Yaung, and E. Charbon, "A modular, direct time-of-flight depth sensor in 45/65-nm 3-d-stacked cmos technology," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 3203–3214, 2019.

[30] W. Becker, *Advanced time-correlated single photon counting techniques*, vol. 81. Springer Science & Business Media, 2005.

[31] W. R. McCluney, *Introduction to radiometry and photometry*. Artech House, 2014.

[32] O. Coddington, J. Lean, P. Pilewskie, M. Snow, and D. Lindholm, "A solar irradiance climate data record," *Bulletin of the American Meteorological Society*, vol. 97, no. 7, pp. 1265–1282, 2016.

[33] I. White and H. Dederich, "American national standard for safe use of lasers, ansi z 136.1–2007," *Laser Institute of America: Orlando*, 2007.

[34] T. Fersch, R. Weigel, and A. Koelpin, "A cdma modulation technique for automotive time-of-flight lidar systems," *IEEE Sensors Journal*, vol. 17, no. 11, pp. 3507–3516, 2017.

[35] B. Buttgen and P. Seitz, "Robust optical time-of-flight range imaging based on smart pixel structures," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 6, pp. 1512–1525, 2008.

[36] G. Satat, M. Tancik, and R. Raskar, "Towards photography through realistic fog," in *2018 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–10, IEEE, 2018.

[37] T. G. Phillips, N. Guenther, and P. R. McAree, "When the dust settles: the four behaviors of lidar in the presence of fine airborne particulates," *Journal of Field Robotics*, vol. 34, no. 5, pp. 985–1009, 2017.

[38] T. Gruber, F. Julca-Aguilar, M. Bijelic, and F. Heide, "Gated2depth: Real-time dense lidar from gated images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1506–1516, 2019.

[39] D. R. Schaart, E. Charbon, T. Frach, and V. Schulz, "Advances in digital sipms and their application in biomedical imaging," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 809, pp. 31–52, 2016.

[40] A. Rochas, M. Gosch, A. Serov, 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.

[41] C. Niclass, A. Rochas, P.-A. Besse, and E. Charbon, "Design and characterization of a cmos 3-d image sensor based on single photon avalanche diodes," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 9, pp. 1847–1854, 2005.

[42] E. A. 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.

[43] C. Bruschini, H. Homulle, I. M. Antolovic, S. Burri, and E. Charbon, "Single-photon avalanche diode imagers in biophotonics: review and outlook," *Light: Science & Applications*, vol. 8, no. 1, pp. 1–28, 2019.

[44] M.-J. Lee, A. R. Ximenes, P. Padmanabhan, T.-J. Wang, K.-C. Huang, Y. Yamashita, D.-N. Yaung, and E. Charbon, "High-performance back-illuminated three-dimensional stacked single-photon avalanche diode implemented in 45-nm cmos technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, pp. 1–9, 2018.

[45] M.-J. Lee and E. Charbon, "Progress in single-photon avalanche diode image sensors in standard cmos: From two-dimensional monolithic to three-dimensional-stacked technology," *Japanese Journal of Applied Physics*, vol. 57, no. 10, p. 1002A3, 2018.

[46] M. Anti, A. Tosi, F. Acerbi, and F. Zappa, "Modeling of afterpulsing in single-photon avalanche diodes," in *Physics and Simulation of Optoelectronic Devices XIX*, vol. 7933, p. 79331R, International Society for Optics and Photonics, 2011.

[47] S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf, and E. Charbon, "A high-pde, backside-illuminated spad in 65/40-nm 3d ic cmos pixel with cascoded passive quenching and active recharge," *IEEE Electron Device Letters*, vol. 38, no. 11, pp. 1547–1550, 2017.

[48] R. Halterman and M. Bruch, "Velodyne hdl-64e lidar for unmanned surface vehicle obstacle detection," in *Unmanned Systems Technology XII*, vol. 7692, p. 76920D, International Society for Optics and Photonics, 2010.

REFERENCES

[49] Apple 2020. https://www.apple.com/newsroom/2020/03/apple-unveils-new-ipad-pro-with-lidar-scanner-and-trackpad-support-in-ipados/.

2 Detector technologies

Detectors play a substantial role in LiDAR sensors. Out-of-visible spectral sensitivity such as near-infrared (NIR) or ultraviolet (UV) wavelengths are commonly preferred due to inherent solar band rejection favoring high signal-to-background noise ratio (SBR). Sensitivity in the NIR or UV region is achieved indigenously in detectors implemented from the III-V compound family due to their wide bandgap. However, integrating electronics with such detectors becomes challenging due to substrate incompatibility with silicon used in mass-produced CMOS technology. Consequently, Si-based single-photon avalanche diodes (SPADs) are more and more preferred due to their ease of integration with CMOS and relatively, mature process flow and manufacturability. This chapter explores a III-V based detector technology as well as Si-based SPAD technology. A CMOS-based circuit is also presented as a hybrid readout solution for a III-V family GaN device. The work presented on the III-V work is written based on the work published in [1] and Si-based SPAD work is based on [2].

2.1   III-Nitride semiconductor detector technology

The ultraviolet (UV) spectrum has been of special interest in space exploration, planetary studies, as well as biomedical applications [3, 4]. Heterostructure devices from the III-Nitride material family such as gallium nitride (GaN) and its alloys are capable of photon-counting necessary for faint object detection. Due to their wide bandgap spanning in the range from 3.4-6.2 eV, they provide inherent out-of-band rejection in the visible wavelengths, thus achieving solar-blind UV sensitivity. Addressing the aforementioned challenges, Jet Propulsion Laboratory (JPL) in California, developed GaN-based avalanche photodiodes (APDs) for imaging applications [5].

Reverse-biased typically at high voltages ($\approx$80 V), these APDs generate currents on the order of hundreds of picoamperes in the proportional-mode, while avalanching to more than a few microamperes when biased beyond the breakdown voltage in the Geiger mode. In Figure 2.1, one can see an example device geometry of a GaN APD – a p-i-n structure, with an atomic layer deposited (ALD) Al_2O_3 and SiO_2 sidewall isolation layer used in this work [5].
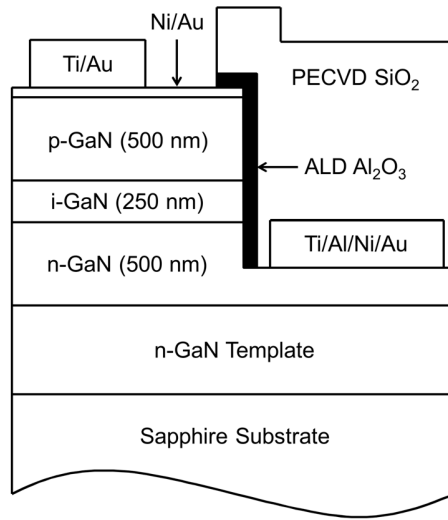
Figure 2.1 – Device geometry of a GaN APD used [5];

Figure 2.2a shows the avalanche operation demonstrated in GaN p-i-n APDs with low dark current and large avalanche gain, shown in Figure 2.2b. At 360 nm, they have an external quantum efficiency of about 60 % and four orders of magnitude out-of-band rejection ratio. The avalanche gain in these devices reaches 10^5, as shown in Figure 2.2b, thus competing with the state-of-the-art GaN devices [5]. In the proportional-mode of operation, at a reverse bias of ~70 V, these APDs generate equivalent currents on the order of 100–200 pA. Beyond 70 V, which is typically their breakdown voltage, the diodes enter into the Geiger mode of operation.
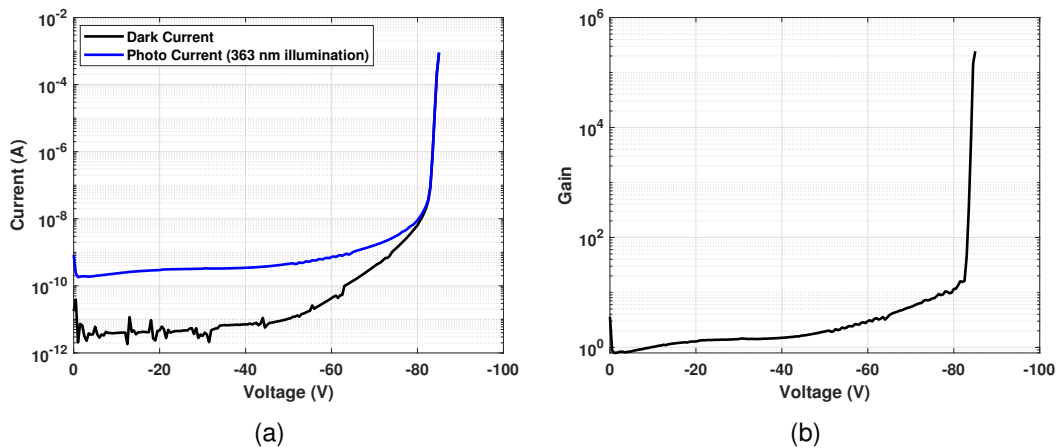
Figure 2.2 – GaN APDs developed at JPL : typical I–V characteristics (a) and (b) avalanche gain (redrawn from [5]).

For in-depth detail on the device and process-level description of these detectors, the reader is directed to [5].

2.2 CMOS interface circuit design

CMOS-based readout circuits are commonly chosen due to their scalability, easy integration, and reliability [6, 7]. However, due to the heterostructural nature of III-Nitride APDs, immediate integration may not become possible. The growing role of such detectors has resulted in more and more solutions around a hybrid sensor design. Evolving from 2D integration, hybridization is gathering more attention with the advent of 3D-integrated technology [8, 9, 10, 11].

This section elaborates on our proposed hybrid approach where a CMOS readout circuit is custom-designed for the GaN APDs developed at JPL. The GaN APDs are operated in the proportional mode and the readout circuit is custom-designed in a 0.35μm HV CMOS technology. The two main challenges (also, functionalities) addressed in this work are the readout's ability to handle high avalanche voltages (up to 80 V) typical of these devices (as seen in Figure 2.2) and quenching of any possible avalanche in the APDs.

A transimpedance amplifier (TIA) circuit became an apparent readout choice in order to amplify the picoampere-scale photodiode currents of the GaN APDs. The block diagram of the two basic TIA architectures are shown in Figure 2.3 along with a photodiode equivalent circuit at the input with its capacitance, C_{pd} and current, I_{pd}.

A TIA with a resistive feedback (RTIA) shown in Figure 2.3a generates an output voltage directly proportional to the feedback resistor according to the equation: $V_{out} = I_{pd} \times R_{fb}$. The main advantage of a RTIA is that the achievable gain is proportional to the feedback resistor, R_{fb} under first-order approximation. However, this advantage is limited by the area occupied by the resistor itself which in turn also limits the achievable transimpedance. Therefore, a RTIA-based circuit was not pursued further and a TIA with capacitive feedback (shown in Figure 2.3b) was chosen instead. A CTIA-based topology provides gain which is inversely proportional to the feedback capacitance C_{fb}. This relationship also reduces the area constraints on C_{fb} when compared to the resistive feedback in RTIA-based circuit.
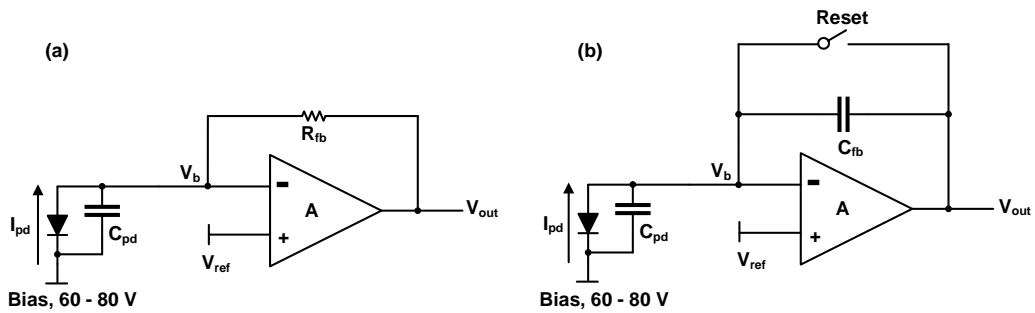
Figure 2.3 – (a) Resistive transimpedance amplifier (RTIA); and (b) capacitive transimpedance amplifier (CTIA).

21

2.2.1 Capacitive Transimpedance Amplifier (CTIA)

A photodiode circuit model derived from the GaN APD characteristics, comprises its capacitance, $C_{pd} \approx$ 1–3 pF and equivalent photo-current, I_{pd}, which is on the order of 100–200 pA at a reverse bias of about 80 V. A conventional CTIA with an open loop gain, A, with a parallel reset using a PMOS transistor (MP$_{reset}$) was designed, as shown in Figure 2.4.
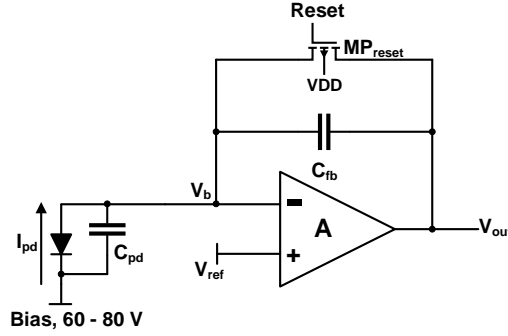
Figure 2.4 – CTIA block diagram.

The CTIA integrates the incoming photodiode current, I_{pd}, on the feedback capacitor, C_{fb}, in order to generate an equivalent observable voltage, V_{out}, at the output.

The low-frequency gain is set by the ratio of photodiode capacitance (C_{pd}) and the feedback capacitance (C_{fb}). The CTIA transfer function can be written as,

$$\frac{I_{pd}}{s}(s) \cdot (1 - e^{-st_{int}}) = V_{out}(s)\left(C_{fb} + \frac{C_{pd} + C_{fb}}{A(s)}\right), \tag{2.1}$$

where t_{int} is the integration time.

When an assumption of A $\gg \frac{C_{pd}}{C_{fb}}$ is made and a DC input signal is considered as in this work, the transient behavior of the CTIA is approximated to the following equation,

$$V_{out} = \frac{1}{C_{fb}} \int I_{pd} dt. \tag{2.2}$$

The -3-dB bandwidth of the CTIA under this assumption is approximated by $1/t_{int}$.

The readout operation begins with resetting of the feedback capacitor C_{fb} by switching the voltage at the gate of the reset transistor MP$_{reset}$ before every integration (Reset → 0). This action sets the DC operating points of the transistors in the CTIA circuit. At the time of reset, $V_{out} = V_{ref}$. Immediately after the release of the reset switch (as Reset → VDD), the photodiode current starts to flow in, integrating on C_{fb} and the output, V_{out}, of the amplifier starts to drop from the initial value V_{ref} set by the reset transistor (when Reset → 0). The negative feedback of the amplifier maintains the input node, V_b, at virtual ground under infinite gain. However, due to the finite gain (A) of the CTIA, there is also a small rise in the voltage level of V_b.

The typical waveforms of the Reset signal and the CTIA output voltage, V_{out}, are shown below for one integration cycle.
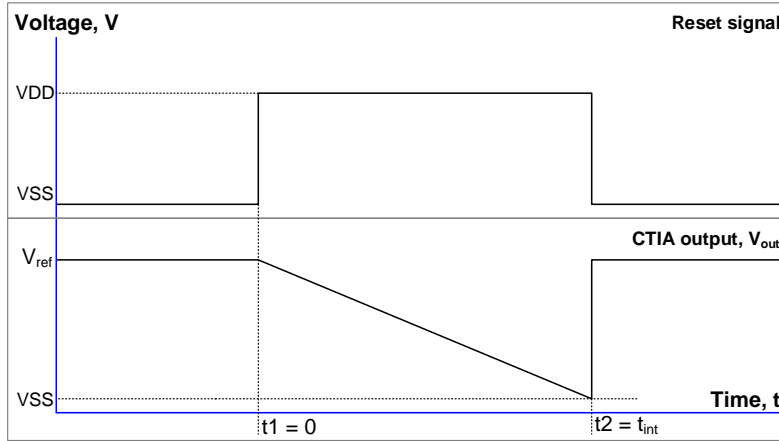
Figure 2.5 – CTIA typical waveforms

2.2.2 Design challenge- high bias voltage and quenching

One of the main challenges while designing a readout for these GaN devices is the requirement to accommodate high reverse bias voltages applied on them. This implies providing a means to isolate the low-voltage CMOS circuitry (operating, typically up to 3.3 V in this process) from the high bias voltages (up to 80 V). The large current flow during avalanche separates charges, creating a dipole and collapsing the voltage across the APD. This leaves the high voltage directly across the low-voltage readout circuit resulting in its potential damage.

Furthermore, process-level variations in the fabrication of GaN devices also result in breakdown voltage variability through different devices. Another challenge is hence, posed by any potential avalanche breakdown which could occur under a high (and, varying) bias due to the presence of high electric fields in the device. Under such conditions, carrier multiplication must also be quenched to avoid damage to the APD as well as enabling successive photon detection.

The aforementioned challenges were addressed by providing a protection circuit using a high voltage (HV) NMOS transistor at the input of the CTIA, shown in blue color in Figure 2.6.

Initially, the gate voltage of the HV NMOS is set such that it biases the transistor to operate in its ohmic region. In this region, there is very little voltage drop across the HV NMOS; thus, the detector receives most of the applied bias voltage. When the CTIA saturates after completing the integration process, the input voltage, V_b , starts rising as the incoming photodiode current can no longer integrate on the feedback capacitance. During an avalanche breakdown when the CTIA saturation will occur rapidly, this rising input voltage can reach damaging levels if it is neglected.

(a) CTIA- Block diagram

(b) CTIA- Transistor-level schematic
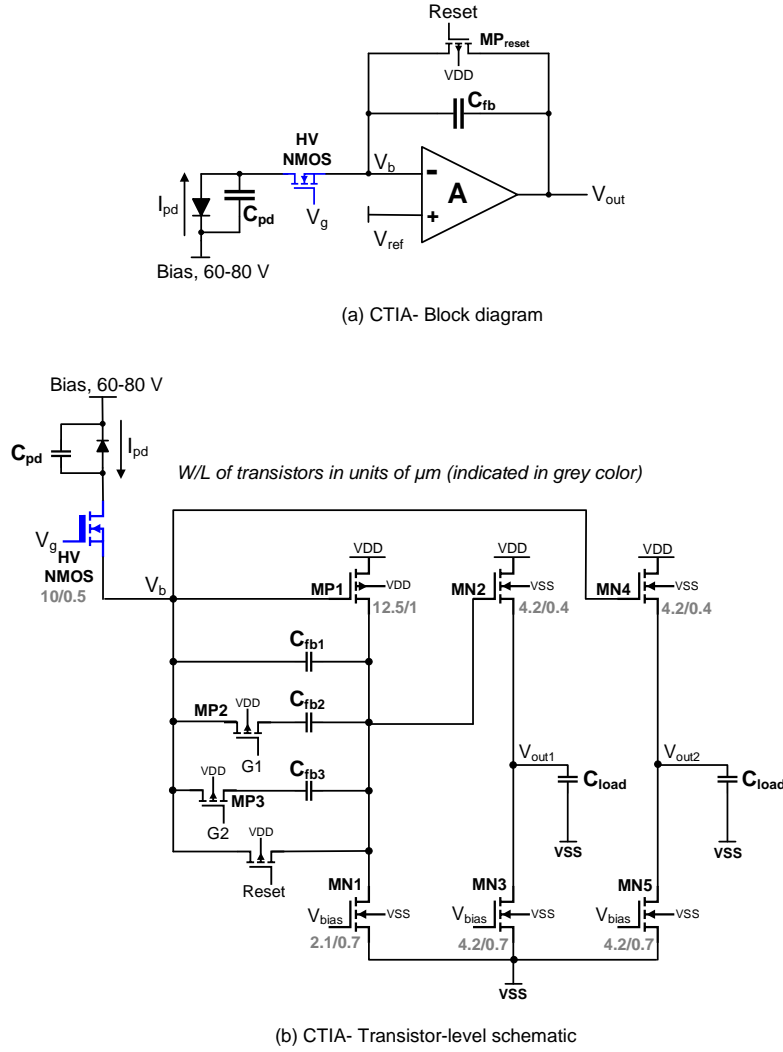
Figure 2.6 – (a) CTIA block diagram with HV NMOS transistor; and (b) CTIA transistor-level schematic.

Introducing the HV NMOS in the input path reduces the gate-to-source voltage V_{gs} of the HV NMOS over rising levels of the node V_b. The HV NMOS thus shuts off when its V_{gs} becomes lower than its threshold voltage, thus isolating the low-voltage CMOS readout circuit from the APD stage. Following this, the photodiode current will end up discharging the photodiode capacitance C_{pd} and no longer flow into the CTIA, eventually also reducing the bias across the APD, which is specifically useful in the quenching mechanism whenever there is avalanche current surge.

The transistor-level schematic of the CTIA is shown in Figure 2.6b. The core of the CTIA is a common-source amplifier (transistors MP1 and MN1). Given that the photodiode current is

generated out of a p-on-n type APD device, a PMOS input (MP1) is used as the gain transistor of the common-source stage. This stage is followed by a NMOS source follower (transistors MN2 and MN3) which acts as an output buffer to avoid any voltage degradation at the output, capable of driving low impedance loads. There is another source follower stage (transistors MN4 and MN5) connected to the input node V_b to allow observation of rising voltage levels. The readout provides variable gain by featuring four effective feedback capacitances (C1–C4) configured by switching the transistors, MP2 and MP3 in Figure 2.6b, obtained such that C1 = C_{fb1}, C2 = C_{fb1} + C_{fb2}, C3 = C_{fb1} + C_{fb3}, C4 = C_{fb1} + C_{fb2} + C_{fb3}. The relative capacitance values were designed such that C2 − C1 = 100 fF, C3 − C2 = 200 fF, C4 − C3 = 100 fF. The bias voltages V_{bias} for the load transistors is generated using a simple current mirror (not shown in the figure).
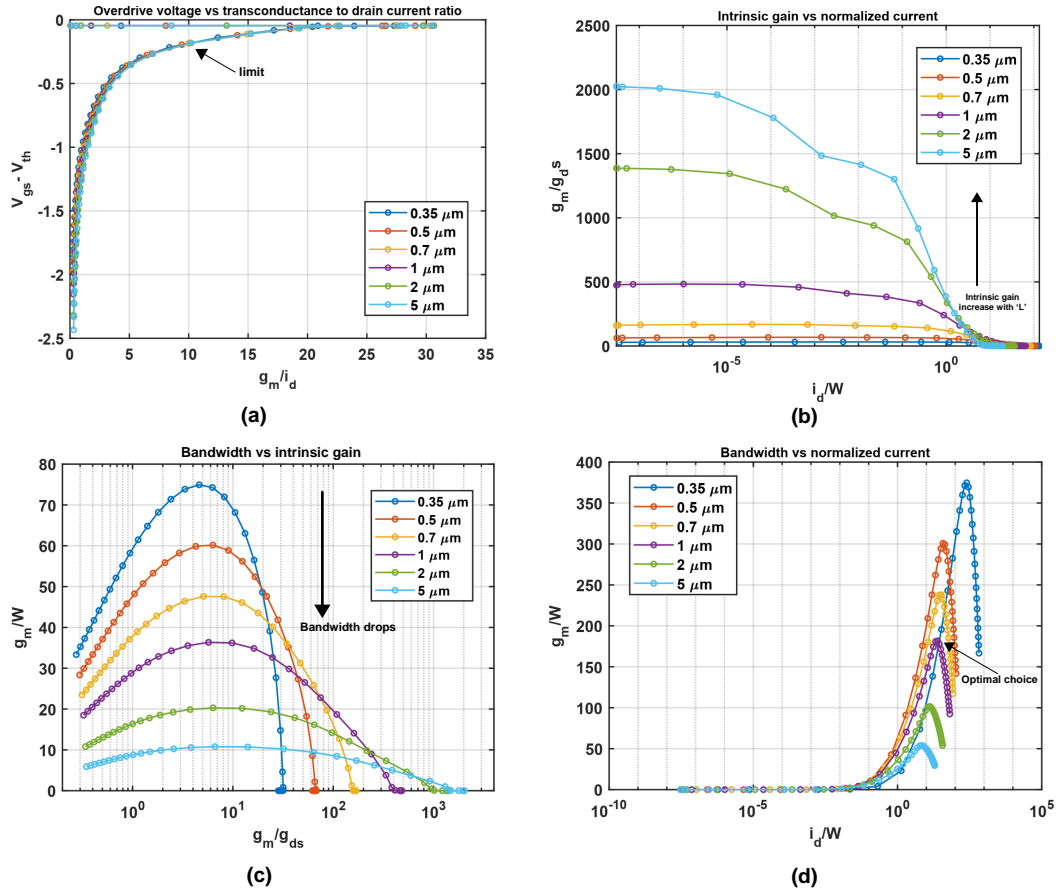
Figure 2.7 – Simulation results– PMOS transistor characterization over increasing length, L; used for appropriate sizing of transistors in Figure 2.6b.

The device sizes were chosen by simulating the PMOS and NMOS transistors in this technology after which various device characteristics were derived. Transistor-level parameters such as the intrinsic gain of the transistor g_m/g_{ds}, the transconductance-to-drain current ratio g_m/i_d, the overdrive voltage, $V_{gs}-V_{th}$ and their relationships were obtained to size the transistors with appropriate width (W) and length (L). A g_m of 0.18 mS and $g_m/i_d \approx 15$ were chosen for the input PMOS transistor, MP1, of the common-source stage. This translated into an equivalent drain current, i_d = 12 µA.

Combining the above estimation with PMOS characterization plots simulated (in Figure 2.7) resulted in transistor sizes, as indicated in Figure 2.6b (annotated in grey color). To summarize, the following course of action will take place in the readout circuit under an avalanche breakdown: integration of the photodiode current; saturation of the CTIA; shutting-off of the HV NMOS; reduction in the diode bias and eventually, quenching. The high voltage bias applied on the APD directly appears at the drain of the HV NMOS. However, the HV CMOS technology used in this circuit allows us to exploit the HV NMOS for this purpose. Finally, after the HV NMOS shuts off, the CTIA is reset again which sets the appropriate DC bias conditions for all the transistors in the CTIA to begin the next integration cycle.

2.2.3   Noise analysis

The current design is a lead-in step towards developing robust readout circuits for large-array GaN APDs in future. Although the current design is not noise-optimized, a preliminary analysis is made, identifying several noise sources in the sensor and later, compared with the measurement results. This study will also help set precise specifications for the custom-design of future readout circuits.

Temporal noise and fixed pattern noise are the major sources of noise in conventional image sensors [12, 13]. In the presented GaN-based CMOS sensor, an array of 1×8 dedicated CTIAs is implemented for a linear array of 8 GaN devices. In the readout circuit, every channel in the 1×8 array is read out independently without any particular technique to minimize fixed pattern noise. This is because, any spatial variation observed in the readout array is mainly dominated by the differences in the detector performance, resulting from the variability in the 8 GaN devices. Therefore, noise analysis in this thesis is focused only on the temporal noise from the implemented CMOS readout. The temporal noise sources include the read noise (including amplifier noise), reset noise arising from the reset action on the CTIA feedback and, shot noise. Read noise which also comprises of the CTIA thermal noise is primarily dictated by the input PMOS transistor MP1 in Figure 2.6. The small-signal model of the input PMOS in the common source amplifier shown below in Figure 2.8.

The input referred noise voltage per unit bandwidth of this transistor is given as follows [14].

$$\overline{V_{n,in}^2} = 4kT\frac{\gamma}{g_m}, \tag{2.3}$$

where k is the Boltzmann constant, T is the absolute temperature and γ is the noise excess factor, a constant which is assumed to be 2 for the 0.35 µm process used in this work. From Figure 2.8, one can deduce the output referred noise density as follows.

$$\overline{V_{n,o}^2} = \int_{-\infty}^{+\infty} \overline{V_{n,in}^2} \mid H(f) \mid^2 df, \tag{2.4}$$

where H(f) = V_{out}/V_{in} is the transfer function of the CTIA. The small-signal model shown in Figure 2.8 can be described as

$$sC_{pd}(V1 - V_{in}) + g_m V1 + V_{out}(\frac{1}{r_{out}} + sC_{load}) = 0. \tag{2.5}$$
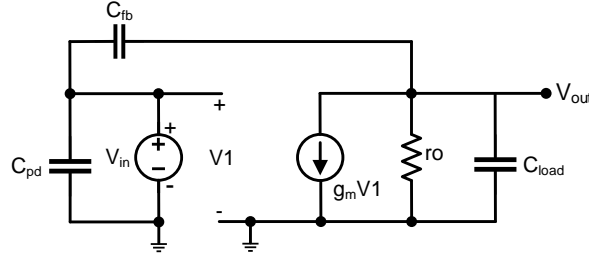
Figure 2.8 – CTIA small-signal model- PMOS input transistor

Given that,

$$V1 - V_{in} = V_{out}\frac{C_{fb}}{C_{fb} + C_{pd}}, \tag{2.6}$$

and assuming that open loop gain $g_m$ro » C_{pd}/C_{fb}, Equation (2.4) can be expressed as,

$$\overline{V_{n,o}^2} = \frac{8kT}{(\frac{C_{fb}}{C_{fb}+C_{pd}})(C_{load} + \frac{C_{fb}C_{pd}}{C_{fb}+C_{pd}})}. \tag{2.7}$$

Considering one of the possible feedback capacitances (to be able to compare directly with measurement results which will follow) and assuming C_{fb} = 400 fF, C_{pd} = 2 pF and C_{load} = 20 pF, an output referred noise voltage of about 90 µV is obtained.

The CTIA also contributes to the reset noise arising from the release of reset switch at the start of every integration cycle. For a given C_{fb}, the reset noise voltage is estimated as, $\sqrt{kT/C_{fb}}$ [15]. Thus, for C_{fb} = 400 fF, $\sqrt{kT/C_{fb}} \approx$ 100 µV was theoretically estimated.

2.3 Measurement results

The readout chip was designed and fabricated in a 0.35 µm HV CMOS technology. The chip consists of 8 units of the CTIA cell shown in Figure 2.6. The photomicrograph of the chip is shown below in Figure 2.9 where the 8 units can be identified. At a supply voltage of 3.3 V, every unit consumes about 198 µW, thus resulting in a total power consumption of about 1.5 mW for the 1 × 8 CTIA array. As seen in Section 2.2.2, the high voltage reverse bias on the GaN APD is presented directly at the inputs of the readout circuit if the breakdown mechanism collapses the voltage across the APD. In order to account for such a possibility, the input pads are laid on one side, providing physical isolation to the low voltage output pads which are on the opposite side. The pitch is 400 µm

in accordance with the pitch of the detector array.

A dedicated electrical test setup was used in order to characterize the fabricated readout chip independently of the detector while a second optical setup, to measure the I–V characteristics of the GaN devices using the readout IC. In the first setup, a constant current was obtained using a voltage source and a series resistance ≈20 MΩ; the voltage source was swept to obtain varying input current conditions. From the resulting measurement data, the gain parameters, effective feedback capacitances and slew rate of the CTIA were extracted. The voltage limiting functionality provided by the HV NMOS was also verified. In the second test setup, the GaN APDs were connected to the input pads of the readout chip and various measurements were performed.
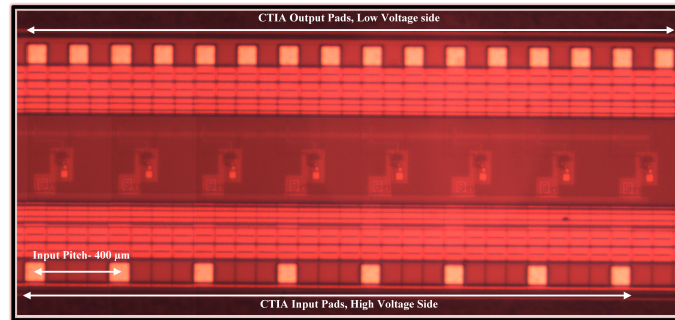
Figure 2.9 – Chip photomicrograph: Eight CTIA unit cells can be identified with their input pads on the bottom side and output pads on the top.

2.3.1 CTIA characterization results
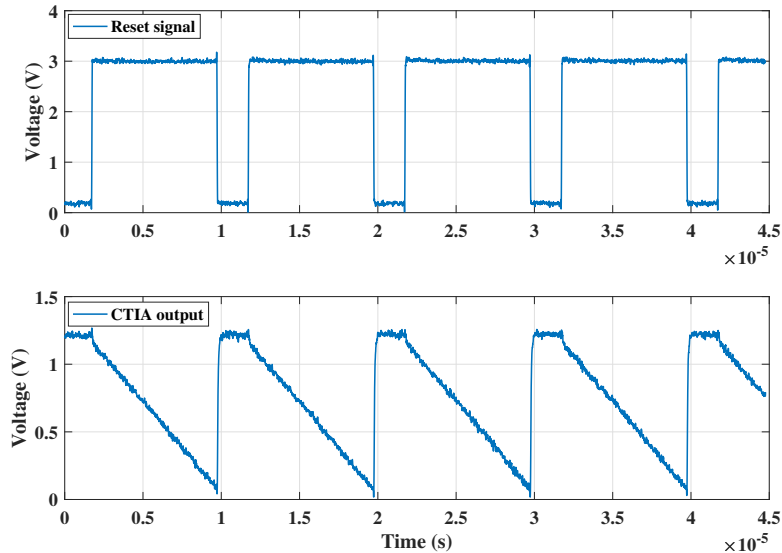
Figure 2.10 – Typical CTIA operation: Reset signal and CTIA output voltage waveforms.

The transient measurement results showing the typical working of the CTIA (as explained in Section 2.2.2) are shown below in Figure 2.10 over four integration cycles. As can be seen, when the reset

voltage at the gate of the PMOS transitor MP_{reset} is 0 V (denoted in Figures 2.4 and 2.6), the output voltage V_out_1 is equal to V_b, which is set according to the DC bias condition applied, $\approx$ 1.2 V. When the reset voltage is pulled up to 3.3 V (supply voltage, VDD), the feedback path of the CTIA opens and the integration of the incoming current on the feedback capacitor begins as expected. This results in a negative ramp (because of a positive current flowing into the CTIA) at the output of the CTIA.

CTIA transient behavior– The voltage source, V_s, connected to a series resistance, R_{in}, of approximately 20 MΩ was swept from 2.8–30 V, providing equivalent input currents to the CTIA, equal to $\frac{V_s - V_b}{R_{in}}$, ranging from 20 nA–0.8 µA. The CTIA output waveform data (at node V_out_1 in Figure 2.6) was sampled on an oscilloscope for varying input conditions which was then used to extract various amplifier characteristics.

Differentiating Equation (2.2) with respect to time results in the following.

$$\frac{dV_{out}}{dt} = \frac{I_{pd}}{C_{fb}}.$$

(2.8)

The term $\frac{dV_{out}}{dt}$ indicates the slope of the CTIA output voltage waveform captured on the oscilloscope.

The slopes were obtained for increasing values of V_s (i.e., voltage source connected to the series resistance, R_{in}) and thus, equivalently, increasing input currents. This slope was extracted with the 4 possible feedback capacitor combination; the result of which is shown in Figure 2.11a. As expected, feedback capacitances C3 and C4 result in lower slope values due to their higher integration time compared to higher slopes for lower feedback capacitances, C1 and C2. According to Equation (2.8), the obtained slope increases for increasing values of the voltage, V_s, (and input currents). Similarly, the direct proportionality of the slope with increasing $\frac{1}{C_{fb}}$ is also evident in Figure 2.11b.
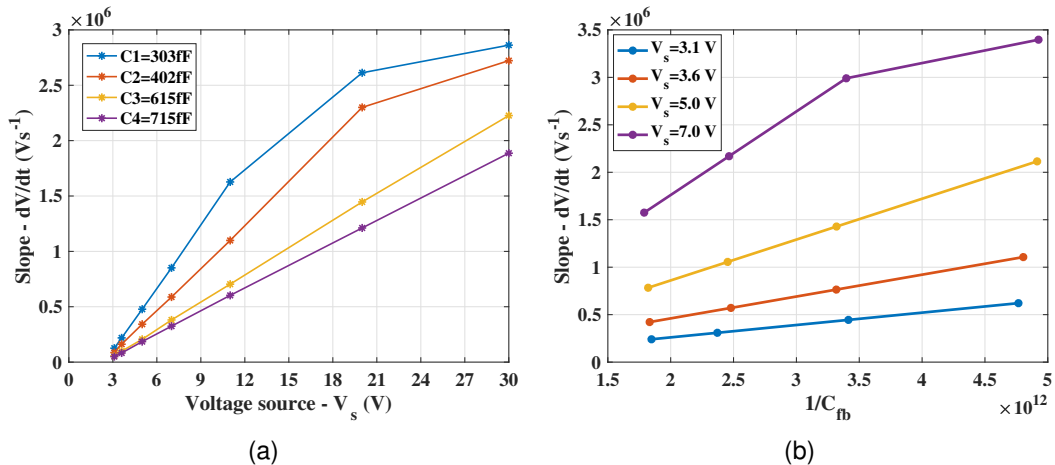
Figure 2.11 – CTIA transient behavior : (a) slope versus voltage source, V_s; and (b) slope versus $1/C_{fb}$.

In Figure 2.11a, it can be observed that, for lower feedback capacitances C1 and C2, the slope of

the output from the source follower (at node V_{out1} in Figure 2.6) saturates for higher input voltage, V_s (equivalently, higher input currents), while the slope is linear for the entire sweep range in case of C3 and C4. Similarly, the inverse trend in the slope approaches saturation for higher input voltage (7 V) in Figure 2.11b.

Slew rate– Figure 2.12 shows the measured slope, $\frac{dV_{out}}{dt}$, for input currents up to 7 µA calculated from CTIA output waveform. It was observed that for input currents approximately above 1 µA, the slopes obtained with all four feedback capacitances, C1–C4, saturated to about ≈2.8 V/µs. This value indicates the slew rate of the CTIA and therefore, on revisiting Figure 2.11, it can be noticed that the slope indeed saturates as it approaches the slew rate point. Independently, this saturation at the output occurred mainly because of the source follower which cannot sink input currents beyond 1.5 µA.
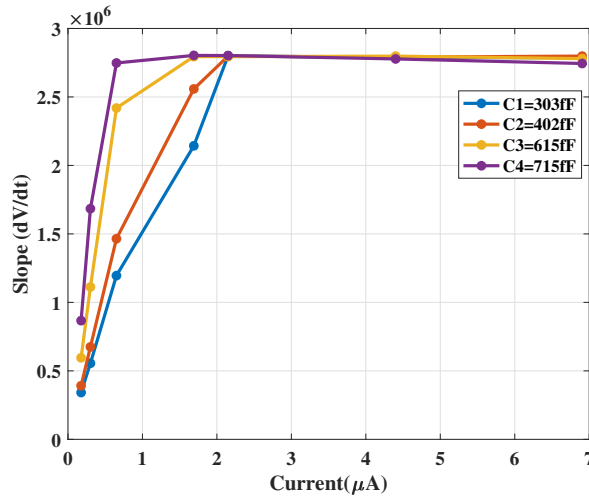
Figure 2.12 – CTIA slope under higher input currents

Geiger mode is an important region of operation in APDs to assess single-photon sensitivity in photodiodes. GaN APDs require to be biased well above voltages >80 V. In this region of operation, the avalanche current increases substantially (to several microamperes). The readout circuit needs to accommodate these higher input currents. Slew rate is thus, an important parameter which indicates the upper limit on the pulldown current of the output source follower in the readout circuit. As seen in Figure 2.12, it is evident that the current design limits the input current to about 1.5 µA. Future versions of the readout chip will therefore require improvement to overcome this limitation to examine Geiger mode in these APDs.

Effective feedback capacitances– The output slope values extracted through results shown in Figure 2.11a were utilized along with corresponding input currents to calculate the effective feedback capacitances based on Equation (2.8). The nominal values of the feedback capacitances from the design are C1 = 50 fF, C2 = 150 fF, C3 = 350 fF, C4 = 450 fF. However, gain measurement results show that the real capacitance values (shown in Figure 2.13) are higher compared to the design values. A reason for this deviation between nominal and measured C_{fb} values is the underestimation

of parasitics during post-layout simulations. However, the relative differences between the measured feedback capacitances (C2 − C1 = 100 fF, C3 − C2 = 200 fF, C4 − C3 = 100 fF) align with the design specifications described in Section 2.2.2. The bar chart in Figure 2.13 further shows that the measured values result in a standard deviation of about 10.6 fF for the smallest feedback capacitance (≈300 fF), indicating that there is only minimal variation in the extracted values from the measurement.
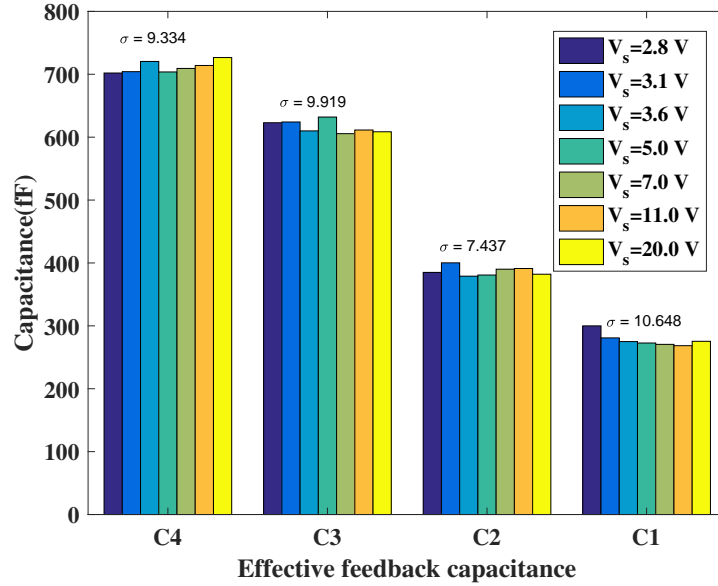
Figure 2.13 – Variation in measured feedback capacitances (labels as mentioned in Section 2.2.2)

Charge-to-voltage conversion factor– Charge-to-voltage conversion factor (CVF) (also referred to as conversion gain) is a common characterization parameter for CTIA-based circuits. It is basically the ratio between the CTIA output voltage and amount of electrons being transferred over the CTIA feedback capacitance to the output, usually expressed in µV/e⁻. In a CTIA circuit, the feedback capacitor sets this figure, for a constant bias condition on the GaN APD.

$$\Delta V_{out} = \frac{\Delta Q_{in}}{C_{fb}}. \tag{2.9}$$

From the results obtained in Section 2.3.1 for the four feedback capacitances, a CVF of 0.39 µV/e⁻ was obtained at a mean feedback capacitance value of 402 fF.

Voltage limiter functionality–The HV NMOS introduced at the input of the CTIA acts like a voltage limiter as explained in Section 2.2.2. This functionality was verified for increasing input currents by monitoring the node V_b (as denoted in Figure 2.6) after CTIA saturation. Figure 2.14a shows the input node (V_b) signal obtained from the oscilloscope measurement. HV NMOS is biased with 4.5 V DC source such that its V_{gs} lies approximately 1 V above its threshold voltage ($V_{gs} − V_{th} \approx 1$ V). Upon completion of integration as the CTIA saturates, V_b starts rising; this is seen in Figure 2.14a.

This rise continues until a point when $V_{gs} < V_{th}$ and the HV NMOS ceases to conduct. Thereafter, V_b saturates; in this case, the saturation occurs at ≈ 3.8 V. It will be seen in subsequent sections that this functionality successfully allowed the possibility of diode-bias voltages as high as 80V without damaging the low-voltage CMOS circuit.

(a)   (b)
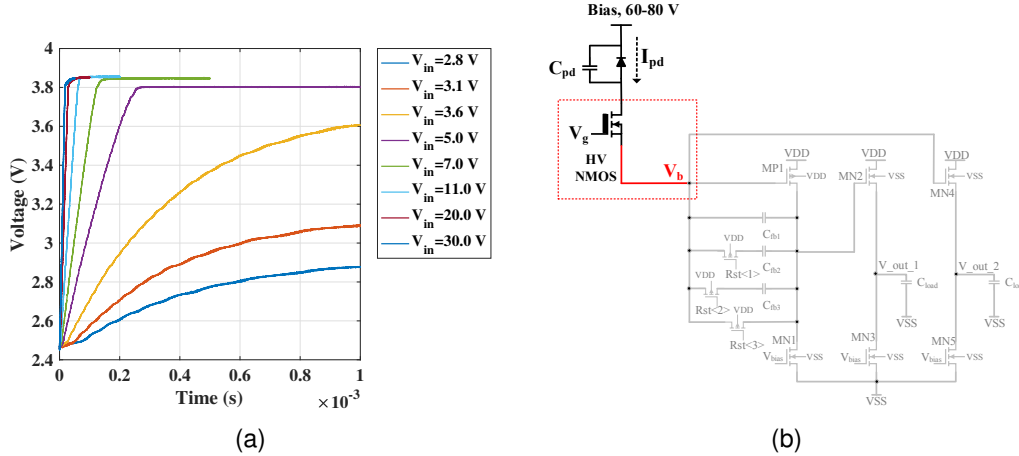
Figure 2.14 – Voltage limiting functionality: (a) rise in the CTIA input node V_b; and (b) CTIA schematic- input node, V_b, highlighted.

2.3.2 GaN + CMOS measurement results– demonstration of UV sensitivity

The standalone characterization of the CTIA circuit was followed by measurements in combination with GaN APDs connected at the input of the CTIA. The CTIA output waveform was used along with the feedback capacitance to extract the effective APD currents. The APD was also illuminated using an available UV LED source and the reverse bias was swept up to 90 V. The I–V characteristics were obtained as shown in Figure 2.15a.

(a)   (b)
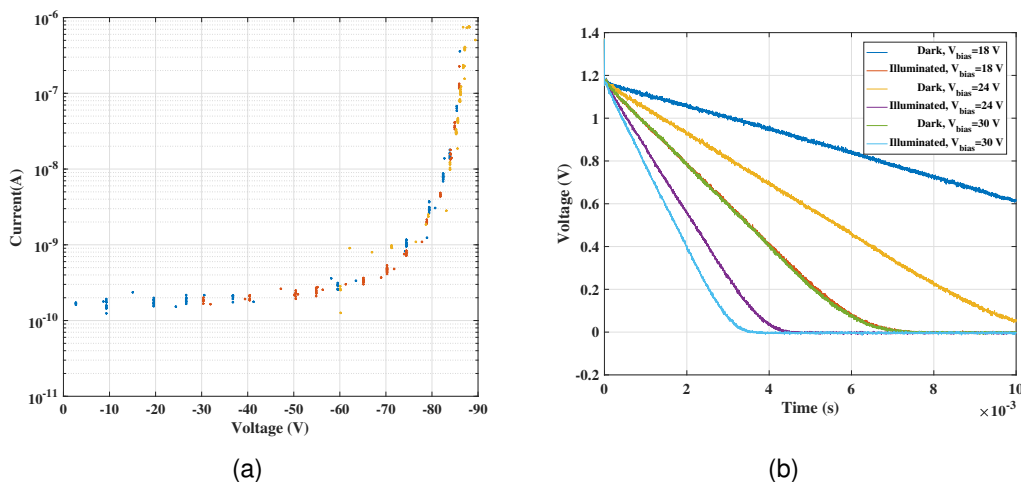
Figure 2.15 – Characteristics of GaN sensor obtained using the CMOS readout circuit : (a) extracted I–V curve of a GaN APD; and (b) CTIA oscilloscope waveforms.
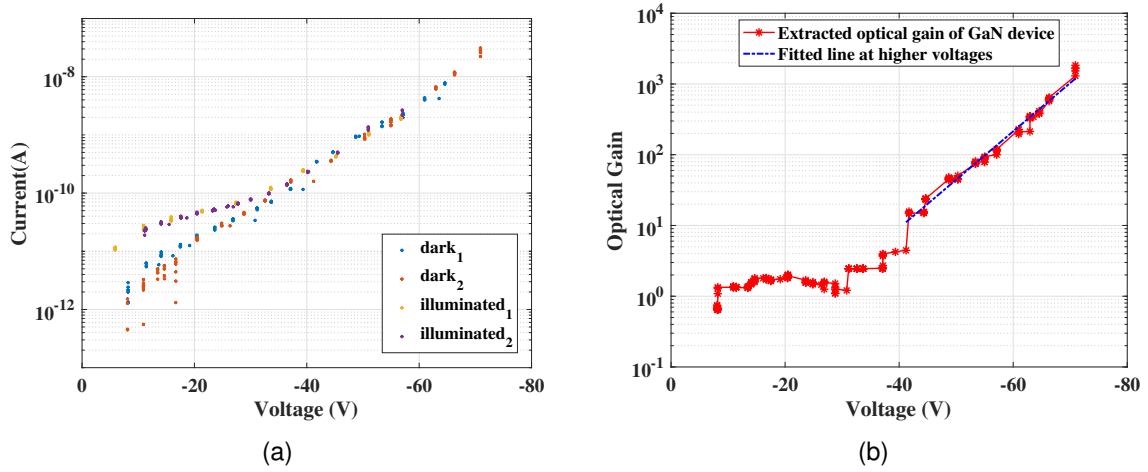
Figure 2.16 – GaN sensor characteristics under UV illumination. (a) It can be seen that, for lower voltages, the dark current is lower than the photocurrent by a factor of 10, while, for higher voltages (>40 V), the dark current also increases as the APD starts avalanching. This characteristic is similar to results shown in Figure 2.2b. (a) I–V curve under UV illumination; and (b) optical gain estimated from Figure 2.16a.

The expected exponential rise in the current can be seen as the APD avalanches beyond 80 V. A measurement was then performed under dark and illuminated conditions separately to demarcate dark versus photocurrent of the GaN APD. The raw oscilloscope waveforms from CTIA output node are shown in Figure 2.15b for three different APD bias voltages. Figure 2.16a shows the measured I–V curve of the GaN APD under UV illumination for increasing reverse bias voltages. Although we are able to distinguish the dark and the photocurrent in the UV region, the results also suggest that the tested GaN APDs have higher dark currents. Different colors in Figures 2.15a and 2.16a indicate results from different data sets with the same input condition. This was done in order to confirm reproducibility of the measurement.

Further, the avalanche gain is also estimated using a method described in [16], as follows:

$$Avalanche\ gain = \frac{I_{pd} - I_{dark}}{I_{pd\_nogain} - I_{dark\_nogain}}, \tag{2.10}$$

where I_{pd} is the photodiode current, I_{dark} is the dark current and I_{pd_nogain} and I_{dark_nogain} are their average values at the unity-gain point. The achieved avalanche gain was about 10 at lower bias voltages, while reaching up to 10^3 at 70 V, shown in Figure 2.16b. The results obtained so far successfully in turn also verify the voltage limiting functionality of the HV-NMOS, by allowing reverse bias voltages up to ≈80 V.

2.3.3 Noise measurement

The temporal noise sources identified in Section 2.2.3 are measured and the obtained results are discussed in this section.

Conversion gain and read noise– The APD, when biased at certain voltage, contributes to shot noise in the sensor and the readout chip contributes to the read noise of the sensor. Shot noise is the temporal variation in the electron-hole pairs generated inside the APD due to random arrival of the impinging photons which increases in proportion to the incident photon level [17]. The arrival of photons is governed by Poisson statistics such that the corresponding variance in number of photons, n, is given by $\sigma_n^2 = \overline{n}$. For the sensor with a gain, G (which includes the CTIA conversion gain and the APD avalanche gain), the output noise voltage V_{out} approaches

$$V_{out} = G.n. \tag{2.11}$$

The shot noise variance, σ_{sn}^2, at the output is then equal to the mean number of incident photons, $\overline{n}$, multiplied by the gain, G. Furthermore, the readout array contributes to the read noise (including the amplifier noise) such that the final noise variance, σ_n^2, is represented as,

$$\sigma_n^2 = \sigma_o^2 + G.\overline{V_{out}} = \sigma_o^2 + G.\overline{n}, \tag{2.12}$$

where the intercept of Equation (2.12), σ_o^2, gives the read noise power [18].

The plot of variance versus mean output voltage (commonly also referred to as the photon-transfer curve (PTC)) is very useful to extract conversion gain and noise contribution of the readout. The PTC plot is obtained by measuring the CTIA with an APD biased at 40 V under typical indoor illumination. Gain and noise parameters are extracted and the results obtained are compared with those obtained in Section 2.3.1.

The variance and mean were extracted for the difference obtained between the first sample ($S1$) and every subsequent sample ($S1 - S2$, $S1 - S3$ etc.) in the measured data. Interquartile estimate of variance was used to reduce the effects of outliers on the sampled data. The plot in Figure 2.17 shows the measured temporal variance for increasing mean output voltages against a fitted line.
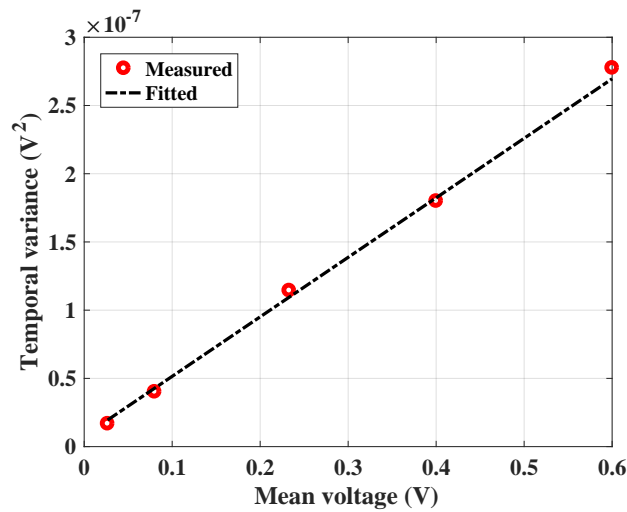
Figure 2.17 – Temporal variance versus mean output voltage extracted from measurement.

The slope of the plot in Figure 2.17 gives the value of conversion gain. For a feedback capacitance of 402 fF used in this measurement, the obtained conversion gain is 0.43 µV/e$^-$ from the slope. This value, as expected, is close to what is estimated in Section 2.3.1 from the measured feedback capacitance of 402 fF (0.39 µV/e$^-$), with <10% deviation, thus, confirming the accuracy of the measurement. The intercept obtained from the line plot measures a read noise voltage of 88 µV. It must be noted that while the read noise includes thermal noise and 1/f noise, the latter contributes little to the total noise at the output. This is because the measurements are made by subtracting the first sample ($S1$) from every subsequent sample as mentioned above; this method is a form of correlated double sampling (CDS) performed off-chip which is a standard approach to combating 1/f noise.

Please note that there is also noise due to the random fluctuations in the gain, G which produces a gain dependent multiplicative "Excess Noise Factor" introducing a scale factor to the PTC curve [19]; however, this is assumed to be small for low gains where we are operating.

CTIA- reset noise– The variation in the voltage level after the release of the reset switch determines the reset noise of the CTIA. The data set used to plot Figure 2.17 is also used to estimate this. The variance at the CTIA output is a combination of the correlated noise from the reset mechanism and the uncorrelated read noise. The read noise obtained from the mean-variance plot is subtracted from the variance of sample, $S1$, acquired right after the release of reset switch [12].

This read noise estimated from mean-variance plot is however, measured from a pair of samples. Thus, to estimate the reset noise, only half of that read noise is considered. This resulted in a reset noise voltage of 121.6µV. Comparing this with the theoretical value (also as calculated in Section 2.2.3— $\sqrt{kT/C_{\text{fb}}} \approx 101.2$ µV and C_{fb} = 402 fF, used in this measurement) shows that the noise floor is dominated by the ADC quantization noise ($\approx$112 µV) from the oscilloscope. Since there was no ADC in the readout chip, the quantization noise from the oscilloscope limited the measurable noise voltage; for the same reason, the thermal noise estimated for the CTIA in Section 2.2.3 could not be measured directly and only the read noise from the mean-variance method was estimated, seen in Figure 2.17.

2.4 Next-generation readout improvements

Table 2.1 summarizes the measurement results obtained so far. The CTIA implemented in this readout chip is a proof-of-concept with GaN APDs where the basic functionality of the readout was successfully verified. However, there are several improvements which can be made in the next version of the readout. The operational bandwidth of the CTIA needs to be improved for which the bias current can be increased. Alternatively, including a cascode transistor is also an option, which will lower down the capacitance at the drain of the PMOS input transistor. However, the trade-off between increased output impedance and lower cutoff frequency needs more analysis at design level. The current version of the chip cannot draw currents above 1.5 µA, as seen in Section 2.3.1.

To accommodate higher avalanche currents of GaN APDs at higher voltages, it is necessary to improve the current sinking capabilities of the source follower stage. Furthermore, a programmable reset needs to be implemented in the next chip, given a clearer understanding of the timing behavior from the transient measurement results. An integrated analog-to-digital converter (ADC) is another future addition to provide on-chip solution for sampling the CTIA output voltage. A target SNR of 50 dB for the next chip results in an achievable loop gain of about 54 dB, making it suitable for this implementation.

Table 2.1 – Performance summary

Parameter	Results
Photodetector technology	GaN avalanche photodiode
APD bias voltage	0–80 V, proportional-mode
Readout technology	0.35 µm HV CMOS, Supply voltage = 3.3 V
Readout topology	Capacitive transimpedance amplifier (CTIA)
CTIA array size	1×8
CTIA area	≈ 5 mm $\times$ 1 mm
Input current range	150 pA–1.5 µA
Slew rate	2.8 V/µs
Conversion gain	0.43 µV/e$^-$
CTIA read noise	88 µV
CTIA reset noise	121 µV
Power consumption	1.5 mW
Avalanche Gain	10^3

2.4.1 Hybrid integration of GaN APDs
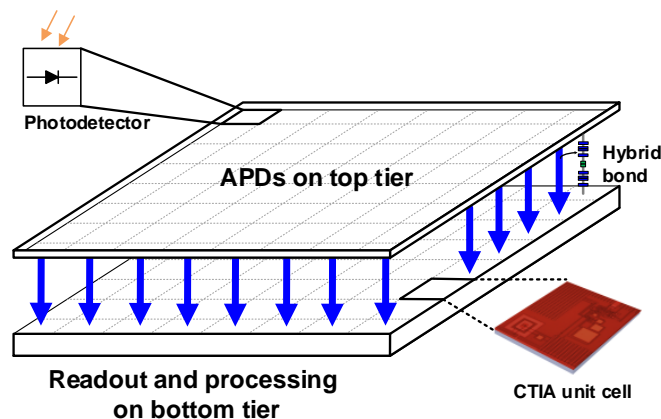
Figure 2.18 – Conceptual representation of 3D stacking.

There has been a growing interest in 3D integration technology in the last few years, thanks to the wafer-level stacking possible with this technology [8, 9, 10, 11]. Figure 2.18 visualizes the 3D stacking concept with photodetectors laid on the top tier and readout, processing and communication unit on the bottom tier. A direct vertical tier-tier connection not only reduces the interconnection parasitics compared to 2D wire-bonded connection but also permits massive parallelization between the two tiers. The work done so far is a proof-of-concept demonstrating the feasibility of our approach of combining GaN APDs with CMOS circuits. The results from this work can enable a 3D-integrated chip in future where the APDs will be stacked face-to-face with the custom CMOS circuit. Furthermore, based on the compatibility of the detector substrate with silicon, the hybrid concept can allow easy porting of the CMOS circuits to any kind of photodetector, including the GaN-based detectors discussed here.

The rest (and the majority) of this thesis focuses on Si-based SPAD image sensors implemented in 3D-stacked CIS/CMOS technology.

2.5 Geiger mode APD- a single photon detector

As described for GaN APD, a geiger-mode APD or SPAD is essentially a pn-junction diode, reverse biased at an excess bias voltage, V_{EB}, above its breakdown voltage, V_{BD}, operating in the Geiger mode. A common P+/N-well based SPAD device is shown in Figure 2.19.
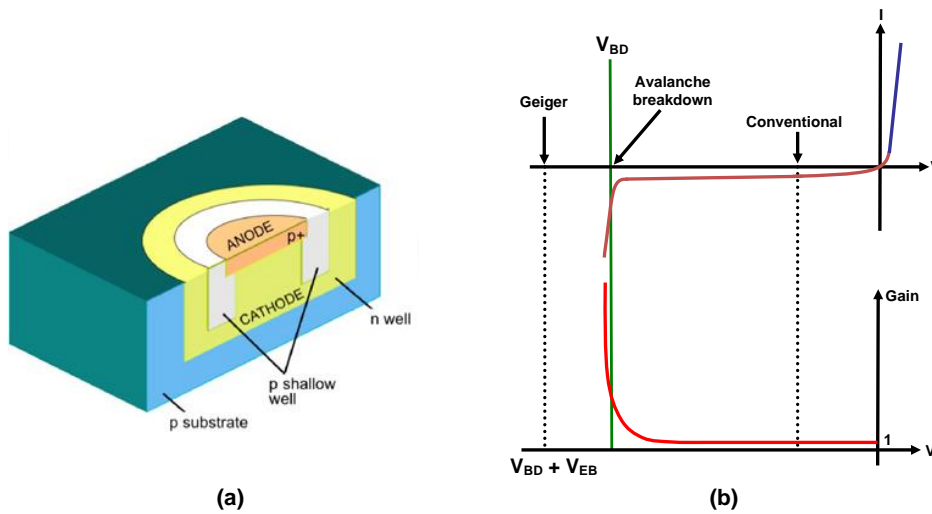
Figure 2.19 – (a) Cross-section view of a P+/N-well SPAD device [20] and (b) typical I-V characteristics of a reverse-biased photodiode.

In the Geiger mode of operation, SPADs provide near-infinite gain compared to linear-mode APDs, as also seen in GaN devices in the previous section. Under this condition, SPADs exhibit high electric field and any impinging photon in the depletion region of the diode may generate electron-hole pairs leading to avalanche self-sustained from impact-ionization between the carriers. The avalanche current rises quickly, within orders of a few nanoseconds and if left unchecked, these could reach

damaging levels (up to a few mAs). Therefore, typically a ballast resistor is connected in series with the diode, known as the passive quenching configuration, as seen Figure 2.20. This resistor quenches the avalanche, by lowering the bias voltage to V_{BD} and thus, lowering the avalanche current. Once the SPAD bias voltage drops to V_{BD}, the avalanche is no longer self-sustaining and the same resistor, R, is used to recharge the SPAD junction capacitance to bring the bias voltage to the initial $V_{BD} + V_{EB}$. The time during quenching and recharge cycle when the SPAD is not active or sensitive to impinging photons (to first order approximation) is referred to as the dead time, t_d. In the configuration shown in Figure 2.20, the dead time is dictated by the charging (and discharging) of the SPAD junction capacitance, C_J. Usually, this time is less precisely controlled in passive quenching when compared to its active counterparts [21]. In a CMOS implementation, the quenching resistor is usually implemented using a MOS transistor operating in the ohmic region, whose gate voltage is used to achieve a tunable recharge time (and thus, the dead time). This will be revisited in Section 2.6.2.
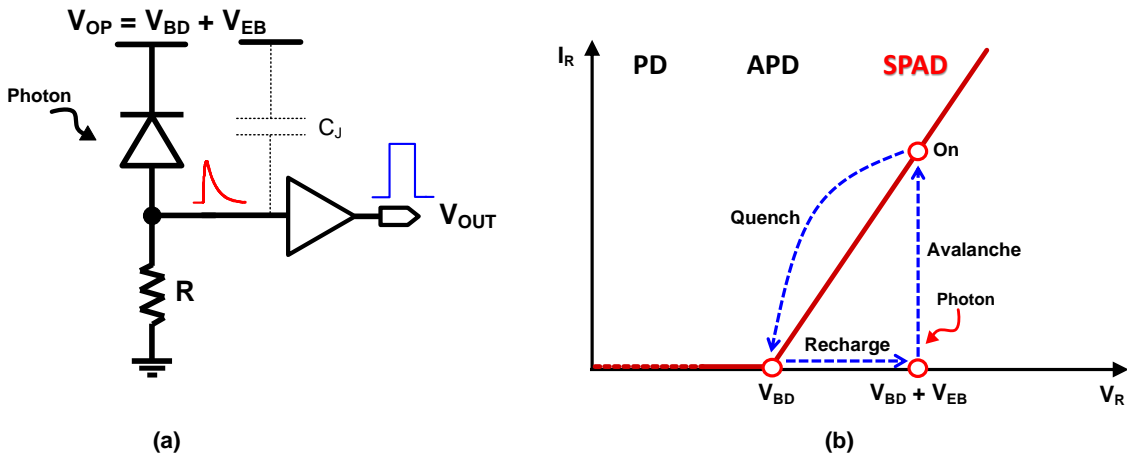
Figure 2.20 – (a) Passive quenching and recharge of a P+/N-well SPAD device using a ballast resistor, R and (b) I-V characteristics of the SPAD in the Geiger mode (redrawn from [22]).

The near-infinite gain of a SPAD thus, makes it sensitive to single photons. The anode of the SPAD in Figure 2.20, is used to sense and propagate any avalanche that may occur. The leading edge of this signal is then used to indicate the arrival of a photon. Typically, buffer/inverter element is utilized to propagate this signal further, without excessively loading the SPAD output capacitance.

SPADs can be implemented in a standard CMOS technology in a monolithic approach where the integrated electronics is housed on the same chip as the detectors, however at a reduced photon-sensitivity due to lower fill-factor arising from complex in-pixel electronic circuitry.

3D stacking technology on the other hand potentially promises a higher fill factor, due to the luxury of stacking all the electronic circuitry on a separate tier, different from the detector tier and thereby, eliminates any sharing of the photosensitive area. Furthermore, separation of the two tiers allows the use of an advanced CMOS technology node for the bottom tier capable of hosting more complex circuitry.

2.6 SPADs implemented in 3D stacked technology

SPADs in 3D stacked technology can be implemented in two forms namely, front-side illuminated (FSI) and/or back-side illuminated (BSI) technology. Conceptual representation of their cross-sections is shown in Figure 2.21.

<div align="center">(a)   (b)</div>
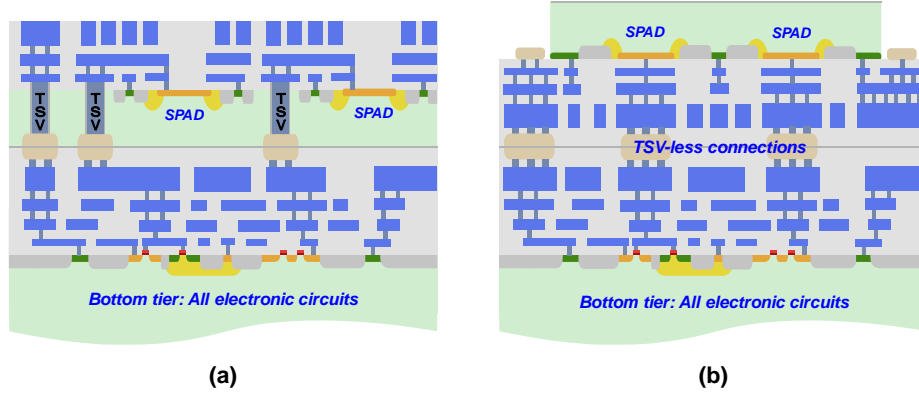
Figure 2.21 – Typical cross-sections of 3D-stacked technology– (a) Front-side illuminated (FSI) and (b) back-side illuminated (BSI) [23].

Both implementations include two tiers, where a dedicated SPAD chip is placed on top of a CMOS-based integrated chip. The primary difference between them is that in FSI technology, shown in Figure 2.21a, a through-silicon-via (TSV) is used to vertically connect the SPAD output to the pixel circuitry on the CMOS bottom tier and the amount of light reaching the photosensitive area is limited by the number of dielectric and metal layers in between, required to convert photons into electrons. In contrast to this, a BSI implementation, shown in Figure 2.21b, utilizes a TSV-free face-to-face connection between the SPAD tier and the circuit tier and the incoming light is therefore collected through the silicon substrate, which basically is the backside of the sensor in (a). The differences in the BSI and FSI structues also result in different PDP spectra. With shallower junction depths, SPADs in FSI technologies are more suitable for near-UV applications. SPADs in BSI technology achieve higher PDP in the red wavelengths and the near-infrared spectrum due face-to-face bonding and thereby, deepened junctions. Considering that the application of focus is LiDAR, FSI SPADs will not be explored any further and BSI SPAD-based sensors will be discussed in the subsequent sections.

2.6.1 3D stacked SPADs in 45 nm BSI CIS technology

The cross-section of the SPAD designed in 45 nm CIS technology is shown in Figure 2.22, which is face-to-face bonded with a 65 nm CMOS chip on the bottom tier [2]. The bottom tier consists of the pixel circuitry including passive quenching and recharge which is directly connected to the SPAD. The pixel output is then fed to time-resolving circuits which timestamp the photon-arrival. The DTOF sensor designed using these SPADs is described in Chapter 3 of this thesis.

The SPAD structure designed in this technology is circular in shape and is based on a P+/Deep N-well (DNW) junction and P-well (PW) guard ring (GR) to prevent premature edge breakdown, as seen in Figure 2.22. The DNW with a retrograde doping allows for a thicker multiplication region and provides high PDP and lower DCR. The active diameter of the implemented structure is 12.5 µm surrounded by a 2 µm GR. Dedicated technology development for every tier allowed thinning down of the substrate to the target thickness of about 3 µm as shown in Figure 2.22.
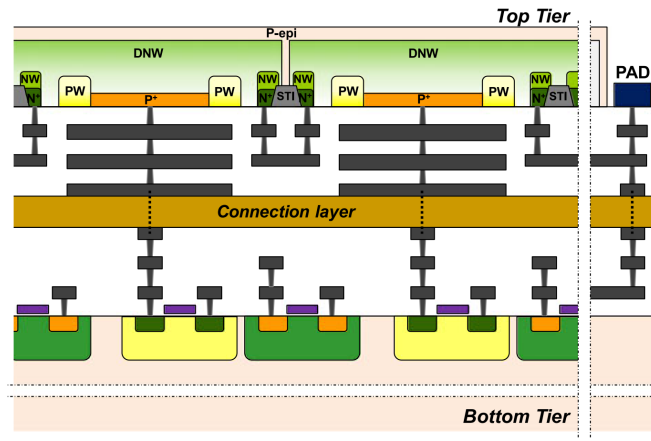
Figure 2.22 – Cross-section of BSI 3D-stacked SPAD in 45 nm CIS technology [2].

The SPAD active region is covered with metal layers in order to reflect the low energy photons back to the active region to enhance PDP at longer wavelengths. Further details on various optimization processes and the design can be found in [2].

SPAD characterization results– The measurement results of the relevant SPAD characteristics are presented in this section. The micrograph of the fabricated SPAD is shown in Figure 2.23 where the bottom tier itself is not visible due to BSI 3D bonding.
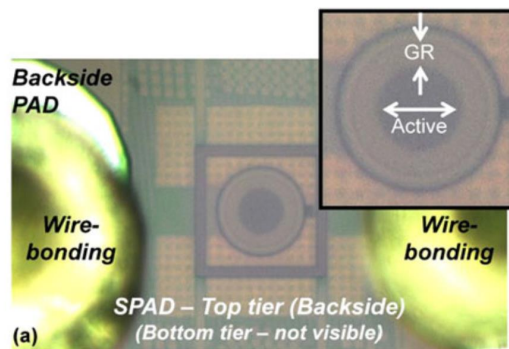
Figure 2.23 – Micrograph of the BSI 3D-integrated SPAD. The inset shows a magnification of active and guard-ring (GR) areas. [2].

Its current-voltage characteristics measure a low dark current, on the order of pA and avalanching at about 28.5 V. The DCR as a function of the excess bias voltage is shown in Figure 2.24a where a DCR of 55 cps/μm^2 at nominal temperature is obtained at excess bias, V_{EB} = 2.5 V. The cumulative distribution of DCR over 128 SPADs is shown in Figure 2.24b where it can be seen that the number of noisy SPADs is relatively small, about 4%.
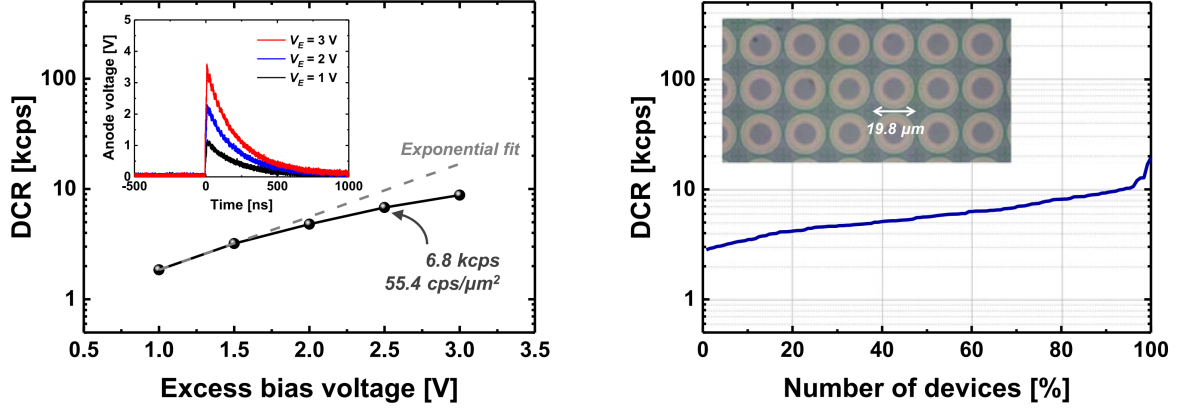
Figure 2.24 – (a) DCR as a function of the excess bias voltage, V_E, at room temperature where the inset shows the output pulses of the SPAD as a function of time and (b) cumulative DCR distribution of 128 SPADs. The inset shows a micrograph of the BSI 3D-stacked SPAD arrays used for this DCR distribution test [2].

The photon detection probability (PDP) of the SPAD is shown in Figure 2.25a. A PDP of 31.8 % at 600 nm is achieved at an excess bias voltage of 2.5 V. In general, as pointed previously, the BSI technology allows for a higher PDP at longer wavelengths compared to a FSI counterpart [24]. A larger depletion region and thinning of the substrate resulted in a uniform sensitivity over the entire visible range, between 400 and 600 nm and particularly, enhanced sensitivity at wavelengths over 700 nm, making them suitable for the targeted LiDAR applications.
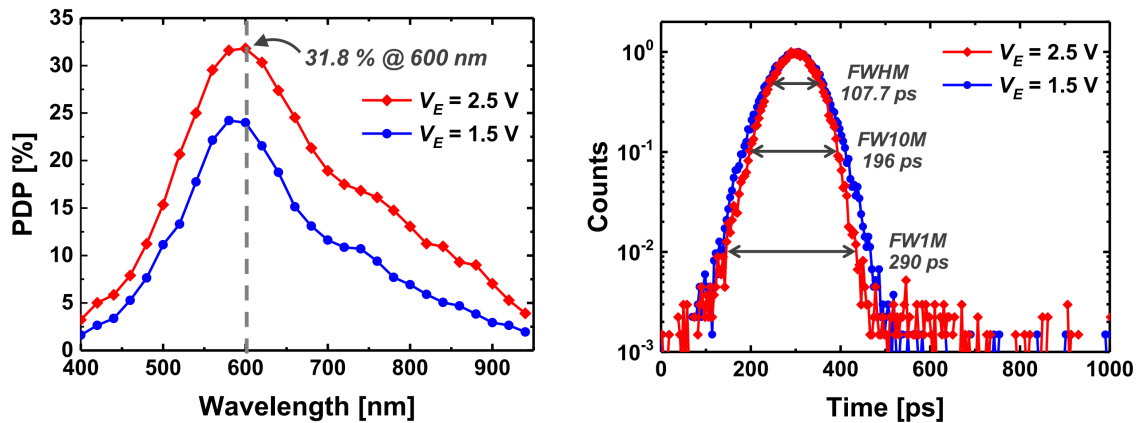
Figure 2.25 – (a) PDP at excess bias voltages of 1.5 V and 2.5 V and (b) timing jitter results using a 637 nm laser.

The timing jitter of the SPAD is characterized using a 637 nm laser source having a pulse width of about 35 ps, shone onto the SPAD at a 40 MHz repetition rate. The normalized histograms of the time interval between the laser output trigger and the SPAD output are shown in Figure 2.25b. A timing jitter of 107.7 ps FWHM is achieved at an excess bias voltage of 2.5 V. The SPADs presented in this work report improved PDP and DCR performance and negligible afterpulsing probability compared to state-of-the-art 3D stacked BSI SPADs. Furthermore, for LiDAR application, where other sources of noise (such as background light) may be dominant contributors, the reported results are adequately suitable. Table 2.2 presents the performance comparison of state-of-the-art 3D-stacked BSI SPADs.

Table 2.2 – State-of-the-art comparison of 3D-stacked BSI SPADs

	Unit	**45 nm CIS**	[25]	[11]	[26]
Top tier	nm	45	130	130	65
Bottom tier	nm	65	130	130	40
Active area	μm^2	122.7	28.3	28	27.6
Fill-factor	%	60.5	n.a.	23.3	45
V_{BD}	V	28.5	12.3	16.5	12
V_{EB}	V	2.5	4	1.5	3
DCR per active area	cps/μm^2	55.4	265.3	1250	391.4
PDP peak at	% at nm	31.8 at 600	11 at 725	13 at 700	27.5 at 640
Timing jitter (FWHM)	ps at nm	107.7 at 637	n.a.	505 at 750	205 at 773

2.6.2 From individual SPAD detectors to functional pixels

Pixel circuit– Expanding from an individual SPAD detector to an image sensor with an array of multiple SPAD pixels involves additional electronic circuitry. Apart from the basic quenching and recharging circuit, masking circuit is another functionality commonly required in imaging arrays which allows selective disabling of pixels. This is particularly useful in turning off noisy pixels (high DCR) which negatively impact the overall sensor performance. Pixel circuit along with passive quenching and recharge along with masking functionality is shown below in Figure 2.26. The shown pixel circuit is implemented in a low power 65 nm CMOS process in the bottom tier (Tier 2), 3D-stacked with the P+/N-well SPADs implemented on the top tier (described in Section 2.6.1). As seen in Figure 2.26, an NMOS transitor M_Q is used for passive quenching and recharge. The transistor M_Q is biased at voltage, V_Q to operate in the ohmic region, providing resistance on the order of 100–200 kΩ, while achieving SPAD dead times under 10 ns. Before any event, the SPAD sees the entire applied voltage, $V_{OP} = V_{BD} + V_{EB}$ as there is no voltage drop over M_Q (assuming no leakage currents). Upon a photon event, the incoming avalanche current of the SPAD causes a voltage drop which raises the SPAD anode voltage to V_{EB}, thus bringing the SPAD bias, $V_{OP} = V_{BD}$. Following this, a recharge process is initiated through M_Q which restores the SPAD bias to $|V_{BD} + V_{EB}|$.
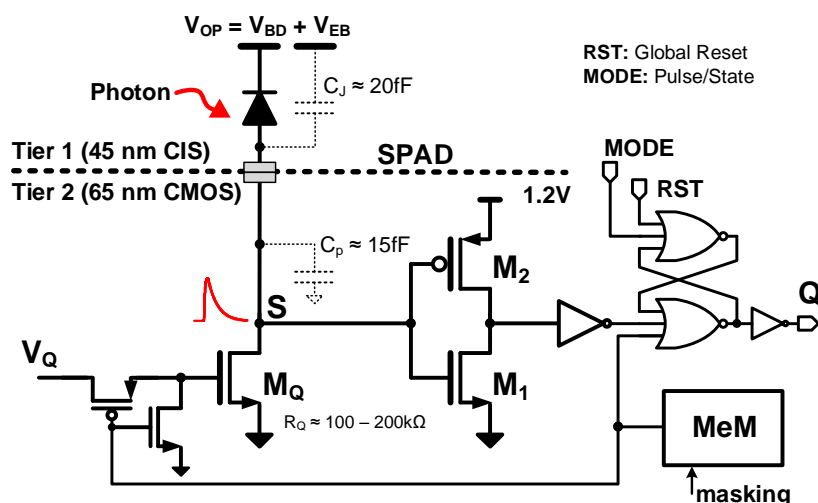
Figure 2.26 – Pixel circuit schematic for P+/N-well SPAD– Passive quenching and recharge along with masking block.

The SPAD anode acts as the sensing node which is connected at the input of a buffer element. The buffer is designed with two inverters which propagate the detected photon event. The quenching transistor, M_Q and the inverter (M1-M2) are all high voltage thick oxide transistors, which support voltages up to V_{EB} (limited by thick oxide breakdown voltage). The second inverter uses low-voltage transistors, supplied by a 1.2 V core voltage in the 65 nm technology design here. The masking function is obtained by programming a 1-bit internal SRAM memory at every pixel, which disables the pixel by setting the gate-source voltage of M_Q to ground when required. There is an SR latch which follows the buffer, designed with 3-input NOR gates as shown, to allow operation in two modes, namely, pulse and state. Depending on the mode chosen, a logic 1 at MODE implies a pulse state of operation, where the SR latch outputs a signal proportional to the dead time. When MODE is set to logic 0, the SR latch output goes to logic 1 and retains the state until an external reset signal (RST in Figure 2.26) is asserted. The implemented circuit occupies an area of 5.3 × 3.6 μm².

A passive quenching and recharge circuit was opted for its smaller area occupancy. Furthermore, a dead time on the order of a few nanoseconds could be achieved using passive quenching in this technology which was considered adequate for the targeted LiDAR application. For the aforementioned reasons, active quenching and recharge circuits are not explored for any sensor design in this thesis.

Bidirectional dual quenching for 3D stacked SPADs– 3D stacking technology enables optimization of the top tier and the bottom tier independently. Consequently, a configurable pixel circuit on the bottom tier, suitable for either SPAD orientation on the top tier is beneficial. In view of this, bidirectional dual quenching circuits are often useful. The higher area occupancy compared to a

basic pixel circuit (like in Figure 2.26), can be traded off for the advantages in terms of flexibility on the bottom tier to accommodate both the SPAD types. First of its kind bidirectional quenching circuit was reported in [27], implemented in a 65 nm /45 nm 3D stacked CMOS technology. A similar dual quenching circuit is implemented in an ultra-low power 22 nm CMOS technology in this thesis. The implemented dual-quenching topology is designed for a DTOF sensor where SPADs are designed in 45 nm CIS 3D-stacked BSI technology. The schematic of the pixel circuit showing passive quenching and recharge for bidirectional SPAD inputs along with masking functionality is shown in Figure 2.27a.
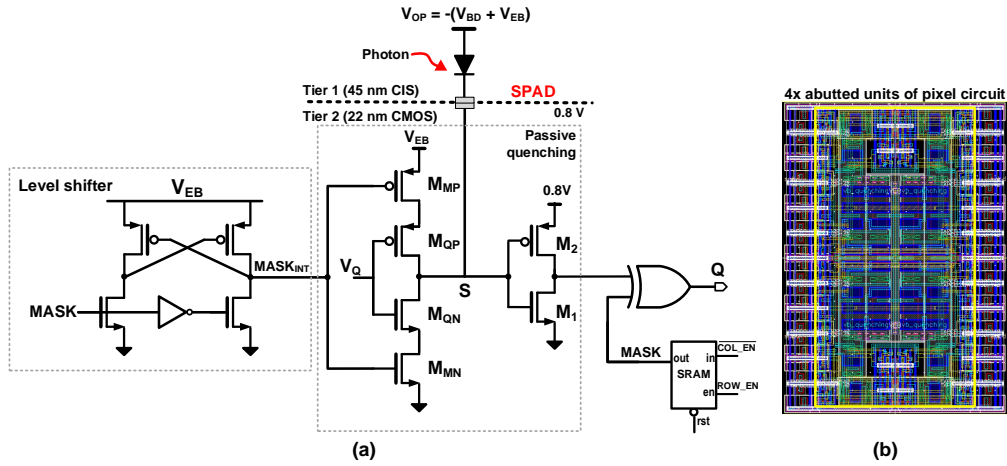
Figure 2.27 – (a) Bidirectional pixel circuit– dual passive quenching and recharge along with masking block and (b) Layout showing four abutted units of the pixel circuit.

If a N+/P-well (P+/N-well) SPAD type, biased with high negative (positive) potential at the anode (cathode), the branch with PMOS (NMOS) transistors, with M_{MP} and M_{QP} (M_{MN} and M_{QN}) perform the quenching and recharge action. The passive quenching and recharge operation itself is exactly as earlier described in Section 2.6.2. The dead time is adjusted by the common V_Q voltage, which controls the resistance of transistors M_{QP} (or M_{QN}) and thus, the decay constant. The transistors M_{MP}-M_{QP} and M_{QN}-M_{MN} along with the following inverter element (M1-M2) are all thick oxide transistors.

The masking signal is interpreted depending on the SPAD orientation. A level shifter is used to internally convert the masking bit to 2.5 V level to control the thick oxide transistors M_{MP} and M_{MN}. The SPADs implemented in 45 nm CIS/ 22 nm CMOS are of N+/P-well type. Therefore, PMOS transistors, M_{MP} and M_{QP}, are used to provide quenching and recharge function. As denoted in the Figure 2.27a, MASK bit of logic 0 is achieved by the logical AND operation, (ROW_EN & $\overline{COL\_EN}$), which turns the transistor M_{MP} on through $MASK_{INT}$ signal, activating passive quenching path required by N+/P-well type SPADs on top tier. Electrical masking of a certain pixel is achieved by turning the MASK bit to logic 1 which turns off M_{MP}. This action pulls the sense node S to ground, thus lowering the SPAD bias to breakdown voltage, V_{BD} and disabling the pixel. An XOR logic gate follows the passive quenching circuit, which combines the SPAD output with the MASK bit to always

propagate a leading edge pulse.

The layout of four abutted units of the pixel circuit is shown in Figure 2.27b, where the area occupancy of one quenching unit is 2.24 × 4.2 μm^2. The abutted layout was implemented in order to optimize area where thick oxide transistors require wider spacing compared to thin oxide transistors according design rules.

2.7 Conclusions

This chapter described two photodetector technologies where, a III-V based GaN detector technology developed at JPL, was first presented. A CMOS readout IC is proposed as a front-end circuit to read out the picoampere-range currents. A 1 × 8 linear array of capacitive transimpedance amplifiers implemented in a 0.35 μm CMOS technology is implemented and tested with GaN APDs. The results show that the readout chip provided a viable solution to operate the APDs at very high reverse bias voltages (≈80 V) without damaging the low-voltage front-end circuit. The current readout chip allows easy characterization of the APDs, while a next-generation of the readout is also being planned, which will enable higher avalanche gains. Future work will also include performing measurements in radiation-environment to ensure functionality in space applications.

Secondly, Si-based SPAD detectors implemented in 3D-stacked 45 nm CIS technology were presented along with their pixel circuits. Particular challenges with packing of dense electronics and consequently, its effect on fill-factor was discussed. The reported SPADs provide superior PDP and DCR performance compared to state-of-the-art BSI SPADs. An enhanced sensitivity (PDP) at wavelengths over 700 nm, < 10 ns dead times, and negligible afterpulsing probability make them promising candidates for LiDAR. Given that the dominant sources of noise is from ambient light, the obtained DCR of 55 cps/μm^2, makes it adequately suitable for the targeted LiDAR applications

REFERENCES

[1] P. Padmanabhan, B. Hancock, S. Nikzad, L. D. Bell, K. Kroep, and E. Charbon, "A hybrid readout solution for gan-based detectors using cmos technology," *Sensors*, vol. 18, no. 2, p. 449, 2018.

[2] M.-J. Lee, A. R. Ximenes, P. Padmanabhan, T.-J. Wang, K.-C. Huang, Y. Yamashita, D.-N. Yaung, and E. Charbon, "High-performance back-illuminated three-dimensional stacked single-photon avalanche diode implemented in 45-nm cmos technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, pp. 1–9, 2018.

[3] M. McGrath, P. Feldman, D. Strobel, K. Retherford, B. Wolven, and H. Moos, "Hst/stis ultraviolet imaging of europa," in *Bulletin of the American Astronomical Society*, vol. 32, p. 1056, 2000.

[4] P. D. Feldman, M. F. A'Hearn, J.-L. Bertaux, L. M. Feaga, J. W. Parker, E. Schindhelm, A. J. Steffl, S. A. Stern, H. A. Weaver, H. Sierks, *et al.*, "Measurements of the near-nucleus coma of comet 67p/churyumov-gerasimenko with the alice far-ultraviolet spectrograph on rosetta," *Astronomy & Astrophysics*, vol. 583, p. A8, 2015. [CrossRef].

[5] S. Nikzad, M. Hoenk, A. D. Jewell, J. J. Hennessy, A. G. Carver, T. J. Jones, T. M. Goodsall, E. T. Hamden, P. Suvarna, J. Bulmer, F. Shahedipour-Sandvik, E. Charbon, P. Padmanabhan, B. Hancock, and L. D. Bell, "Single photon counting uv solar-blind detectors using silicon and iii-nitride materials," *Sensors*, vol. 16, no. 6, 2016. [CrossRef].

[6] C.-C. Hsieh, C.-Y. Wu, F.-W. Jih, and T.-P. Sun, "Focal-plane-arrays and cmos readout techniques of infrared imaging systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 4, pp. 594–605, 1997. [CrossRef].

[7] Y. Bai, S. G. Bernd, J. R. Hosack, M. C. Farris, J. T. Montroy, and J. Bajaj, "Hybrid cmos focal plane array with extended uv and nir response for space applications," in *Proceedings of SPIE*, vol. 5167, pp. 83–93, 2003. [CrossRef].

[8] S. Kavusi, K. Ghosh, K. Fife, and A. El Gamal, "A 0.18 μm cmos 1000 frames/sec, 138db dynamic range readout circuit for 3d-ic ir focal plane arrays," in *Custom Integrated Circuits Conference, 2006. CICC'06. IEEE*, pp. 229–232, IEEE, 2006. [CrossRef].

[9] C. L. Keast, B. Aull, J. Burns, C. Chen, J. Knecht, B. Tyrrell, K. Warner, B. Wheeler, V. Suntharaligam, P. Wyatt, *et al.*, "Three-dimensional integration technology for advanced focal planes," *MRS Online Proceedings Library Archive*, vol. 1112, 2008. [CrossRef].

[10] D. Henry, J. Alozy, A. Berthelot, R. Cuchet, C. Chantre, and M. Campbell, "Tsv last for hybrid pixel detectors: Application to particle physics and imaging experiments," in *2013 IEEE 63rd Electronic Components and Technology Conference*, pp. 568–575, May 2013. [CrossRef].

[11] J. M. Pavia, M. Scandini, S. Lindner, M. Wolf, and E. Charbon, "A 1 × 400 backside-illuminated spad sensor with 49.7 ps resolution, 30 pj/sample tdcs fabricated in 3d cmos technology for

near-infrared optical tomography," *IEEE Journal of Solid-State Circuits*, vol. 50, pp. 2406–2418, Oct 2015. [CrossRef].

[12] H. Tian, B. Fowler, and A. E. Gamal, "Analysis of temporal noise in cmos photodiode active pixel sensor," *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 92–101, Jan 2001. [CrossRef].

[13] B. A. Fowler, J. Balicki, D. How, and M. Godfrey, "Low-fpn high-gain capacitive transimpedance amplifier for low-noise cmos image sensors," in *Sensors and Camera Systems for Scientific, Industrial, and Digital Photography Applications II*, vol. 4306, pp. 68–78, International Society for Optics and Photonics, 2001.

[14] B. Razavi, *Design of analog CMOS integrated circuits*. 2001.

[15] J. Hynecek, "Spectral analysis of reset noise observed in ccd charge-detection circuits," *IEEE Transactions on Electron Devices*, vol. 37, pp. 640–647, Mar 1990. [CrossRef].

[16] P. Suvarna, M. Tungare, J. M. Leathersich, P. Agnihotri, F. Shahedipour-Sandvik, L. Douglas Bell, and S. Nikzad, "Design and growth of visible-blind and solar-blind iii-n apds on sapphire substrates," *Journal of Electronic Materials*, vol. 42, pp. 854–858, May 2013. [CrossRef].

[17] A. J. Theuwissen, "Cmos image sensors: State-of-the-art," *Solid-State Electronics*, vol. 52, no. 9, pp. 1401–1406, 2008. [CrossRef].

[18] T. E. James Janesick, Kenneth Klaasen, "Ccd charge collection efficiency and the photon transfer technique," vol. 0570, pp. 0570 – 0570 – 13, 1985. [CrossRef].

[19] R. J. McIntyre, "Multiplication noise in uniform avalanche diodes," *IEEE Transactions on Electron Devices*, vol. ED-13, pp. 164–168, Jan 1966. [CrossRef].

[20] E. Charbon, "Single-photon imaging in complementary metal oxide semiconductor processes," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 372, no. 2012, p. 20130100, 2014.

[21] M. Ghioni, S. Cova, F. Zappa, and C. Samori, "Compact active quenching circuit for fast photon counting with avalanche photodiodes," *Review of scientific instruments*, vol. 67, no. 10, pp. 3440–3448, 1996.

[22] M.-J. Lee and E. Charbon, "Progress in single-photon avalanche diode image sensors in standard cmos: From two-dimensional monolithic to three-dimensional-stacked technology," *Japanese Journal of Applied Physics*, vol. 57, no. 10, p. 1002A3, 2018.

[23] M.-J. Lee, P. Sun, G. Pandraud, C. Bruschini, and E. Charbon, "First near-ultraviolet-and blue-enhanced backside-illuminated single-photon avalanche diode based on standard soi cmos technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 5, pp. 1–6, 2019.

REFERENCES

[24] M.-J. Lee, P. Sun, and E. Charbon, "A first single-photon avalanche diode fabricated in standard soi cmos technology with a full characterization of the device," *Optics express*, vol. 23, no. 10, pp. 13200–13209, 2015.

[25] E. Charbon, M. Scandini, J. M. Pavia, and M. Wolf, "A dual backside-illuminated 800-cell multi-channel digital sipm with 100 tdcs in 130nm 3d ic technology," in *2014 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–4, IEEE, 2014.

[26] S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf, and E. Charbon, "A high-pde, backside-illuminated spad in 65/40-nm 3d ic cmos pixel with cascoded passive quenching and active recharge," *IEEE Electron Device Letters*, vol. 38, no. 11, pp. 1547–1550, 2017.

[27] S. A. Lindner, *Time-resolved Single-photon Detector Arrays for High Resolution Near-infrared Optical Tomography*. PhD thesis, EPFL, 2018.

3 Resource sharing in DTOF sensors

Moving beyond a pixel circuit to designing a TOF sensor requires minimum of a timestamping circuitry such as a TDC, in addition to any other processing circuits depending on the complexity of the sensor. This necessitates an area-efficient sensor design under a given power budget without compromising on the timing performance of the sensor. While in-pixel TDC architectures have been implemented, often they are power hungry and their use is limited to photon-starved scenarios. Furthermore, it has a direct implication on the data throughput due to limited IO bandwidth. Consequently, resource sharing between pixels becomes inevitable. This chapter delves into this aspect and proposes a shared sensor architecture suitable for TOF applications. The DTOF sensor presented in this chapter is based on the work published in [1] and [2]. The content on analytical modeling and simulation is based on work published in [3].

3.1 Per-pixel and shared architectures

Direct time-of-flight (DTOF) sensors typically require timestamping circuits, such as, time-to-digital-converters (TDCs), to record the time of arrival of photons reflected from the target being mapped. As a result, most sensors involve much complex circuitry, going beyond the basic pixel circuit introduced in the previous chapter. While 3D-stacked architectures enable us to pack dense electronics by exploiting feature size in advanced technology nodes, fill-factor issues and high data volume often determine the upper limit on the integrable circuit density.

Typically, DTOF sensors are implemented either in a per-pixel-TDC architecture or within a shared-pixel-TDC design; in the former, every pixel has a dedicated TDC (shown in Figure 3.1a) and in the latter, multiple pixels share a common TDC. Per-pixel TDC designs mostly operate in an event-driven mode, where, the TDC starts upon a photon detection and stops at the end of reference time frame. Alternatively, the TDC can also operate in a reverse start-stop mode, where it is stopped by a photon event. Unlike monolithic implementations where, in-pixel TDC designs suffer from a very low fill-factor [4], 3D stacked in-pixel sensors have demonstrated up to 60.5 % [5]. However, their suitability is often limited to photon-starved applications and image sensors with smaller array. The increasing number of pixels also demands high bandwidth output channels to stream out the in-pixel

timestamps requiring GHz bandwidth in flash LiDAR applications which almost never operate in photon-starved regime. Additionally, the power consumption is also higher as a result of the high incoming photon activity in LiDAR applications. Consequently, on-chip data processing such as histogramming is necessary [6, 7].

While advanced technology nodes allow packing denser circuitry required in such on-chip processing, the trend towards decreasing pixel pitch still limits that density at pixel level. As a result, shared pixel architectures, promoting optimal resource sharing for timestamping and on-chip data processing are increasingly used.

Shared architectures can be operated in event-driven mode or sampled approaches where the TDC is operated continuously. The TDC itself is commonly based on ring-oscillator (RO) based designs for its lower area occupancy. A phase-locked loop (PLL) is then used to provides a reference and tracks for PVT variations.

3.1.1 Power consumption in shared architectures
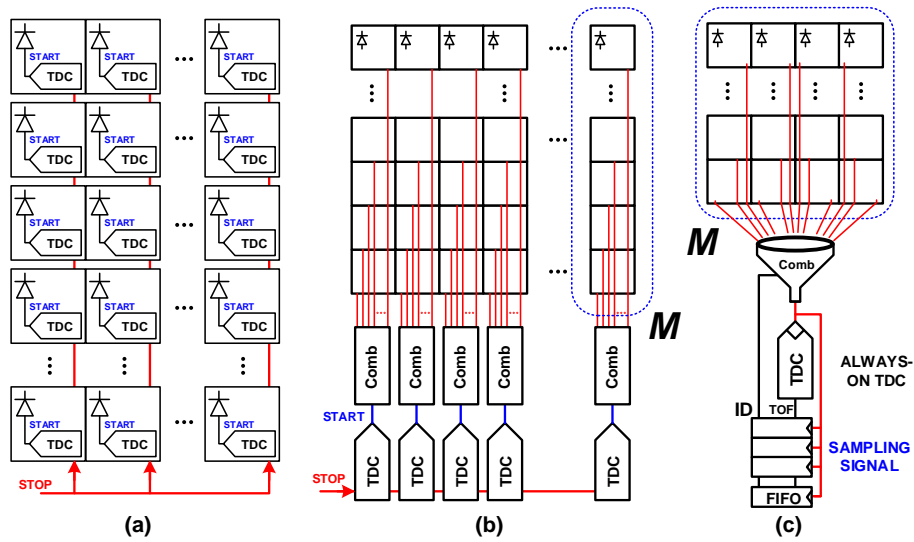
Figure 3.1 – TDC-pixel arrangement- (a) Per-pixel, event-driven, (b) column-wise, event-driven and, (c) always-on, shared TDC concept.

A shared architecture comprising M pixels, that share a single TDC is assumed and a total number of N TDCs in a sensor of $M \times N$ pixels. A combination circuit is required to combine multiple events among the M pixels. The generic concept of such a shared architecture is shown in Figure 3.1b,c. The average time a TDC remains activated is represented by a parameter, referred to as $\overline{\alpha}$, which in a noiseless system approaches the target location with respect to a reference time frame. In the presence of noise, $\overline{\alpha} \approx 0.5$, the middle point of the time frame, given that noise is uniformly

distributed. The presence of a target in the presence of noise will deviate $\overline{\alpha}$ from 0.5 depending on the returning signal intensity compared with noise levels. A parameter $\overline{\beta}$, represents the average activity rate of the pixel normalized to the laser repetition rate, F_{laser}. The product, $\overline{\alpha} \cdot \overline{\beta}$, alternatively also indicates the duty cycle of the TDC. If an upper limit of up to 1 event per pixel is assumed, this product is modified to $\overline{\alpha} \cdot min(\overline{\beta} \cdot M, 1)$, for a group of M pixels which share the TDC. The generic power consumption of this shared architecture is expressed as follows,

$$P_T = P_{PLL} + \#p \cdot C_{line} \cdot V^2 \cdot F + \overline{\alpha} \cdot P_{TDC} \cdot N \cdot min(\overline{\beta} \cdot M, 1)$$
$$+ E_{comb} \cdot N \cdot min(\overline{\beta} \cdot M \cdot F_{laser}, \tau^{-1}), \tag{3.1}$$

where, P_{PLL} is the power consumption of the PLL. The dynamic power of $|\#p|$ high-frequency (F) PLL phases, distributed over as many capacitive wires (C_{line}), with voltage swing V is expressed by the second term. The term, E_{comb} is the energy consumed per event by the circuit which combines multiple events, called the 'combination circuit'. τ is defined by the dead time of that combination circuit, limiting the activity among M pixels. P_{TDC} is the power of a single TDC. In Figure 3.1c, $\tau = \Delta t_{comb} \cdot \log_2 M$, where Δt_{comb} is the propagation delay of every binary combination stage.

In an in-pixel TDC architecture, the power consumed over M pixels containing a TDC each, is represented as,

$$P_{T,per-pixel} = \overline{\alpha} \cdot P_{TDC} \cdot M \cdot min(\overline{\beta}, 1). \tag{3.2}$$

The common term, P_{PLL} and the dynamic power of the distribution lines are ignored for direct comparisons between both architectures. The term E_{comb} is neglected since every pixel contains a TDC. The total power consumption in a shared architecture based on a sampled approach, is estimated over M pixels that share a single continuously running TDC in an array of $M \times N$ pixels. The combination circuit propagates a photon event to the TDC and the sampled timestamp is streamed out along with the corresponding pixel address through a FIFO-based readout circuit. The dead time for conversion and the system saturation is dictated predominantly by the combination circuit dead time (here, τ). The total power consumption over M pixels, is given by:

$$P_{T,shared\_sampled} = P_{TDC} + E_{comb} \cdot min(M \cdot F_{laser} \cdot \overline{\beta}, \tau^{-1}). \tag{3.3}$$

Although the TDC is continuously running, it contributes to a constant power independent of the incoming photon activity ($\overline{\beta}$). A dedicated power grid for the always-on TDC separated from the rest of the circuits will allow maintaining this constant power consumption and consequently, constant IR-drops, which are inevitable.

On comparing Equations (3.2) with (3.3), a relationship is obtained in equation 3.4 which establishes the condition to be satisfied for the shared always-on TDC architecture to offer better power efficiency

compared to in-pixel TDC architectures.

$$P_{T,per-pixel} \geqslant P_{T,shared\_sampled},$$

$$M \geqslant \frac{1}{\overline{\alpha} \cdot min(\overline{\beta}, 1) - \left(\frac{E_{comb} \cdot min(F_{laser} \cdot \overline{\beta}, (M \cdot \tau)^{-1})}{P_{TDC}}\right)}. \tag{3.4}$$

3.1.2 Sensitivity and saturation

Assuming an ideal condition of infinite IO bandwidth and data throughput, a shared architecture suffers from lower sensitivity compared to in-pixel architecture. The presence of the combination circuit dead time limits the maximum achievable sensitivity. Multiple events which can occur within the group of M pixels, may not be detected if they occur within this dead time. As a result of this, the combination circuit resets itself after τ, making it available for successive detections. In a shared architecture as this, the total dead time is a combination of τ, the SPAD dead time, τ_{SPAD} and the TDC dead time. However, the overall SPAD dead time in a shared case reduces to $\approx \tau_{SPAD}/M$ and the TDC dead time is neglible since it is always-on and the instantaneous states are sampled at the leading edge of an event. This makes τ, the dominant contributor. Thus, the maximum TDC conversion rate in a shared architecture as this is given by the inverse of the tree dead time, $\approx 1/\tau$. Assuming a non-paralyzable model [8, 9], the effective observable pixel activity rate is given by a reduction in the sensitivity,

$$\overline{\beta}_{shared} = \frac{\overline{\beta}}{1 + M \cdot 1/T_{win} \cdot \overline{\beta} \cdot \tau}, \tag{3.5}$$

where, T_{win} is the observation window and τ the combination circuit dead time.

While in-pixel TDC architectures do not suffer the aforementioned saturation, a non-ideal situation of limited output bandwidth and a typical LiDAR scenario of high incoming photon flux can blind the sensor by keeping TDCs occupied with noise events continuously. A comparison between in-pixel and shared architectures is made to evaluate their suitability with an example LiDAR scenario. A F_{laser} = 1 MHz allowing up to 150 m LiDAR measurement. An $\overline{\alpha} \approx$ 50 % (0.5) is assumed. A simple combination circuit based on an OR tree is assumed and its power consumption is estimated by the switching of $\log_2 M$ capacitors such that thus $E_{comb} \approx 2 \cdot (1/2 \cdot C \cdot V^2) \cdot \log_2 M$. A 1 fF capacitor per gate is assumed. The TDC power consumption considered is 500 µW, assuming a 65 nm CMOS process (also used in this design). The observation window, T_{win} introduced earlier is assumed to be 5 ns. The dead time of every binary stage, t_{comb} forming the combination tree, is assumed to be 80 ps.

Following the above mentioned parameters, the relationship between power, number of pixels, M, sharing a TDC and the pixel activity, $\overline{\beta}$, is plotted in Figure 3.2. The maximum observable activity, after compression (Equation 3.5) due to the combination circuit is plotted in Figure 3.2b. The data points in black indicate the observable activity ($\overline{\beta} \cdot M$) in an in-pixel TDC architecture operating in an

event-driven mode. It is capable of detecting only a single event in a given time frame. The data points on the grey curve indicate the observable activity in a shared-TDC architecture where an evident saturation is noticeable due to the inevitable combination circuit dead time, absent in in-pixel TDC design. Conditions above the blue line makes it more power-efficient to share a TDC instead of using a single TDC per pixel. An obvious reduction in power consumption with more pixels sharing a single TDC comes at the cost of fewer photons detected compared to in-pixel architecture. However, this is imminently valid for shorter observation cycles. For longer observation windows, the conversion rate in the proposed shared approach is inversely proportional to the combination circuit dead time which can reach up to Gtimestamps/s for a defined group of M pixels while the in-pixel approach is still limited to a maximum conversion rate of F_{laser} timestamps/s per pixel.

A typical flash LiDAR system detecting over long ranges and wide FOV operates under low detection probability due to limited SBR. Assuming the previously mentioned simulation parameters and $\bar{\beta}$ of $\approx 10\,\%$ (0.1), it is more power efficient to share a TDC than an in-pixel approach for a group of $M >= 5$ pixels. For $M = 64$ pixels, the power consumption is 3.2× lower than an in-pixel arrangement for the same number of pixels. However, as pointed out before, the effective detection drops to 62 % of incoming photons, assuming 5 ns T_{win} (see Figure 3.2). In this case, the combination circuit dead time amounts to $\tau = 80$ ps·$\log_2 64$), resulting in a 2 Gtimestamps/s. On the contrary, the in-pixel arrangement is limited to a single event conversion per time frame and therefore, only 1 Mtimestamps/s for the given $F_{laser} = 1MHz$.

(a)   (b)
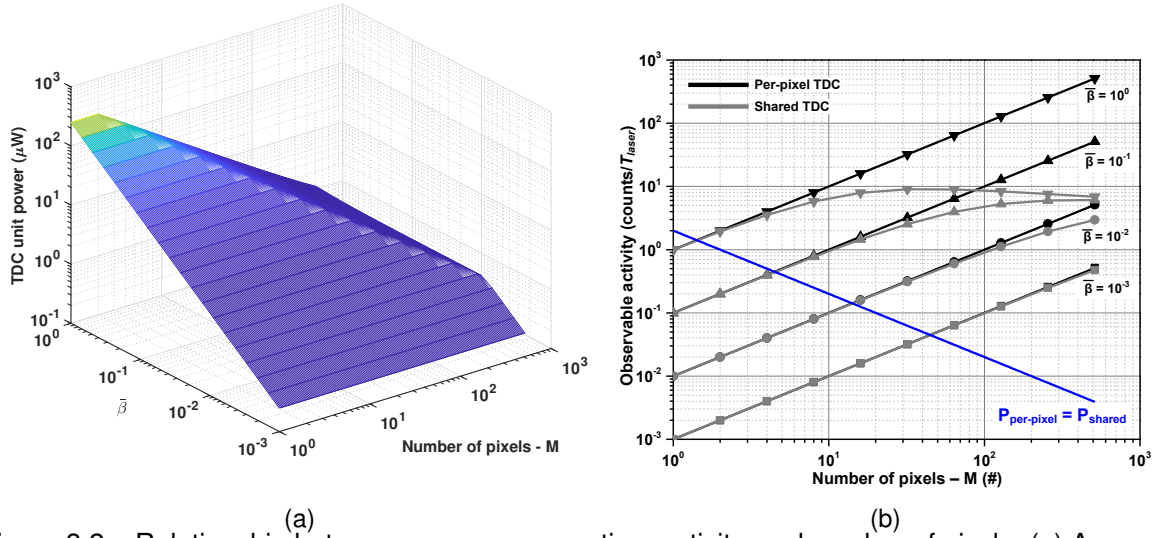
Figure 3.2 – Relationship between power consumption, activity, and number of pixels. (a) Average power per TDC unit; (b) $\bar{\beta}$ compression due to combination dead time, within a laser pulse (T_{laser}) of 5 ns.

In LiDAR applications where high background noise is an apparent challenge, it is necessary to increase timing throughput at lower power and consequently, a shared approach is more suitable.

3.2   A shared approach towards a DTOF sensor

A shared DTOF sensor is designed based on the sampled approach described in the previous section. The block diagram of the sensor is shown in Figure 3.3. The sensor is designed and fabricated in a 3D-stacked BSI 45 nm CIS/ 65 nm CMOS technology process. The SPADs designed in the 45 nm CIS process were described in the previous chapter. Given the first design attempt in this process, a conservative pitch of 19.8 µm was chosen. The architecture of the bottom tier, hosting the processing electronics is designed to suit a 3D-stacking technology. It consists of a module of 8×16 pixels formed from two subgroups, consisting 8×8 pixels each. The two subgroups share an always-on TDC where the subgroup size has been chosen based on the achievable activity rate for a given incoming photon flux and power efficiency, as analyzed in the previous section.

Every pixel has its dedicated passive and quenching recharge circuit, described in Section 2.6.2, directly beneath every SPAD. The in-pixel electronics are laid out respecting the 19.8 µm pitch. Every subgroup has a combination tree, which is referred to as the decision tree capable of managing multiple photon events within the subgroup. The signal propagated by the decision tree, (shown in Figure 3.3 as dTOF) samples the states of the continuously-running TDC and also generates the address/ID of the associated pixel. This ID acts as a pointer for the in-pixel memory which stores the sampled TDC timestamp.
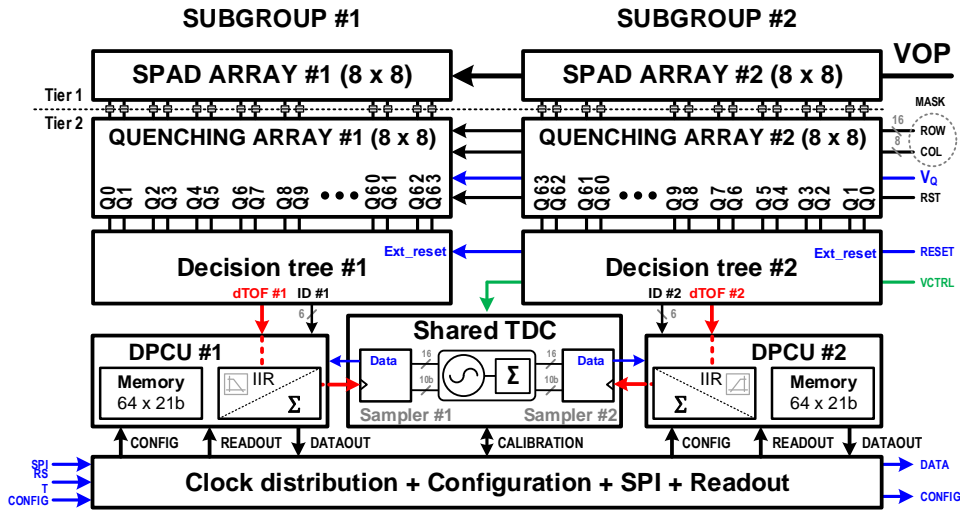
Figure 3.3 – Block diagram– A module comprised of two subgroups of 8 x 8 pixels (SPADs), shared TDC, in-locus digital processing and communication unit (DPCU), and memory.

The module is self-contained in nature, in that, there are in-locus data processing units for additional arithmetic and logical operations to be described later. The entire module is digitally synthesized using Cadence (R) and Synopsys tools. A selected number of blocks, such as the passive quenching and recharge circuitry, decision tree, 1-bit in-pixel SRAM memory and the TDC were all custom-designed and laid out manually. These cells were then placed as MACRO unit cells during

the top-level digital implementation of the module. The remaining circuits including the in-locus processing were all digitally described using RTL followed by automatic placement and routing. The modularity and digital nature of this architecture allows easy scaling to larger sensor sizes while accelerating the design process by avoiding excessive analog verification flows.

The subsequent section describes the building blocks of the DTOF sensor in detail.

3.2.1 Decision tree

Multiple events could occur in the subgroup within a short temporal window. The decision tree is a combination circuit composed of decision makers which act as binary arbiters. Decision makers manage multiple events by detecting the first photon event within a burst of events ("first-come-win-all" policy). The digital pixel outputs directly after quenching and recharge are connected to the decision makers at the first level. At each level, earlier of the two events in comparison is detected and the winning event gets propagated down to the next level and continues to the sixth ($\log_2 [64]$) level (64 pixels in a subgroup), where a single output is provided. This signal, referred to as the dTOF (see Figure 3.3), samples the always-on TDC. In addition to this, an address bit identifying the winning pixel gets generated at every level using a chain of multiplexers and at the end of the tree, corresponds to a 6-bit binary word, providing the address (ID) of the winning pixel. The dTOF signal generated is then used to resample the address word and an adequately delayed version of the dTOF signal is used to self-reset the decision tree to make it available for subsequent detections. Alternatively, an external reset, RESET in Figure 3.3, may also be asserted to limit the maximum activity to a defined number. The conceptual representation of the decision tree is shown in Figure 3.4 for 8 pixel inputs going over three-level ($\log_2 [8]$) propagation and 3-bit address generation.
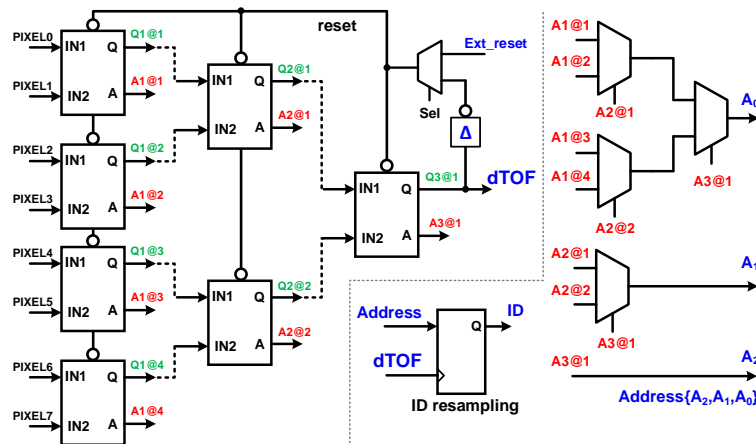
Figure 3.4 – Decision tree concept for 8 pixel inputs.

The decision tree dead time is basically the propagation time of the first event through the 6 binary levels, including the signal processing and reset time, amounting to less than 2.4 ns. This dead time provides a maximum conversion rate of up to 830 Mevents/s over the module of 128 pixel ($2 \times 8 \times 8$) or 6.5 Mevents/pixel/s. It is desirable to lower the tree dead time to increase the number

of conversions to cope up with high background noise in a typical LiDAR scenario.

The decision maker circuit is shown in Figure 3.5.
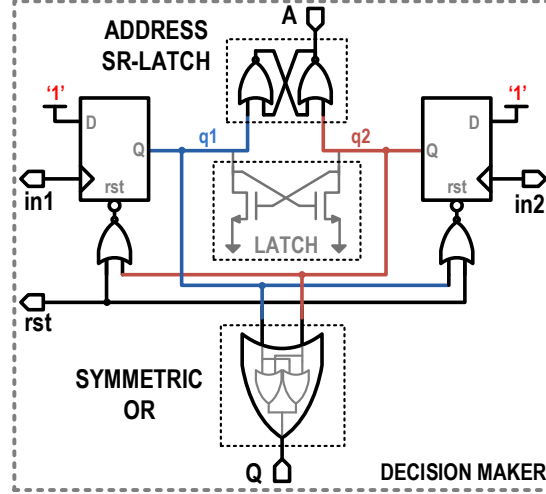
Figure 3.5 – Decision maker schematic.

Upon photon detection in any of the inputs, in1 or in2, logic 1 gets sampled and the earlier D-type flip flop (DFF) resets the later one to avoid detections after the earlier event. The DFF outputs are combined through a symmetric OR gate, equalizing their input loads, to provide a single output, Q through equal delays to either inputs to outputs. The internal nodes, q1 and q2 connect to an SR-latch which provides the address bit of the source event. The decision maker circuit can be entirely reset by a global reset (either self-reset or external).

While there is no metastability observed between the inputs, possible conflicts at the DFF outputs may result in different input-output delays, ($\tau_{in-to-Q}$) which directly impact the timing. This issue was resolved by adding the NMOS latch as shown in the circuit, which reduced the delay variation from 120 ps to 7.5 ps ($\pm$5%) within similar window ($\Delta_{in} = \pm 7$ ps). The metastablity window itself is very small, $\Delta_{in} = \pm 7$ ps. Its contribution compared to other sources of temporal noise from the TDC or jitter such as from the SPAD (reaching up to 100–150 ps) is very minimal.

Figure 3.6 shows the layout of the subgroup and final position of the building elements placed via script in the digital flow. The symmetric connections between the pixels enable a maximum of 1% uniformity variation among the pixels, obtained through Monte Carlo simulations. The same can be calibrated during post-processing.

3.2.2 Time-to-digital converter (TDC)

The always-on TDC is shared between two subgroups and hence carefully placed in between them, as shown in Figure 3.6. The location is placed so that the TDC is readily available to be sampled by dTOF signals from both subgroups. Due to area constraints, imagers often use TDC based on ring oscillators (ROs).
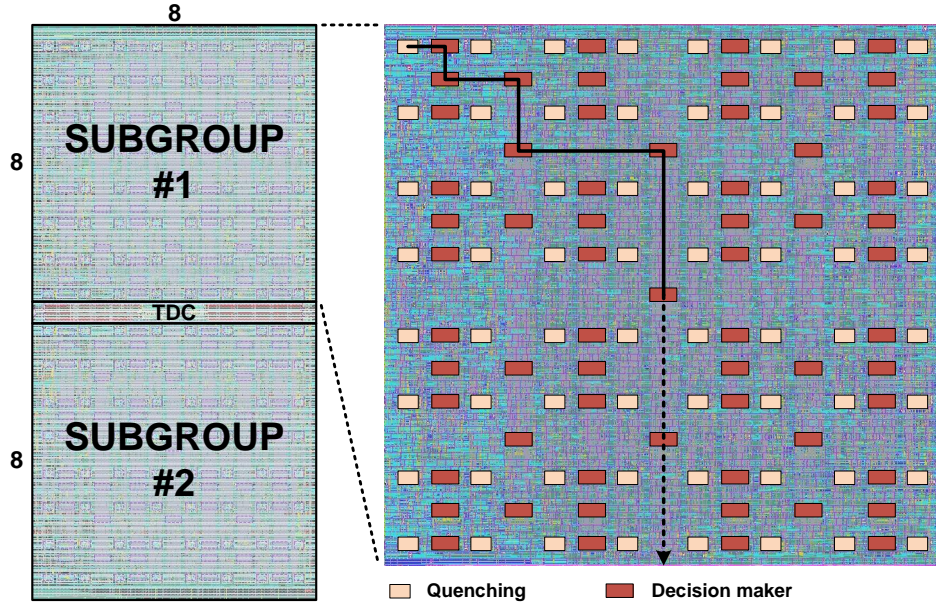
Figure 3.6 – Layout of a module consisting two subgroups obtained after place and route

For the intended LiDAR application, a maximum achievable range of 150 m and a resolution of 9 mm is defined as a target specification. This resolution in position translated into an equivalent timing resolution of about 60 ps. In any TOF measuring system, the achievable resolution is dictated by different noise sources in a system, as seen in Chapter 1, Section 1.3.7. These sources arise mainly from the SPAD jitter, TDC quantization noise and laser trigger delay and its jitter. Considering the noise sources part of the sensor design and assuming statistical independence between them, the total contribution from each of them can be written as a sum of their variances.

$$\sigma^2{}_{total} = \sigma^2{}_{SPAD} + \sigma^2{}_{TDC,jitter} + \sigma^2{}_{TDC,quantization} \tag{3.6}$$

Translating Equation 3.6 in terms of full width at half maximum (FWHM) expression gives,

$$FWHM_{total} \approx \sqrt{FWHM^2{}_{SPAD} + 2.335 \times \sigma^2{}_{TDC,jitter} + 2.335 \times \sigma^2{}_{TDC,quantization}} \tag{3.7}$$

where, the TDC quantization noise is given by the following equation [10].

$$\sigma^2{}_{TDC,quantization} = \frac{t^2{}_{res}}{12} \tag{3.8}$$

Here, t_{res} is the resolution of the TDC. The core of the TDC used in this design consists of a voltage controlled ring oscillator (VCO) which is used to estimate the fractional part of the TOF being measured and a counter which is used to measure the integral part of TOF. VCO, thus determines the resolution of the TDC. The target resolution of 60 ps translates to an equivalent standard deviation of about 17.32 ps from the TDC quantization noise according to Equation 3.8.

Further, in order to quantify the TDC jitter ($\sigma^2_{TDC,jitter}$), the figure of merit (FOM) is derived by considering the oscillation frequency of the VCO, its power dissipation and the offset frequency at which the phase noise is measured as shown below [11].

$$FOM_{VCO} = 10log\Big(\mathscr{L}_{VCO}(f_m).\frac{f_m}{f_{osc}}.\frac{P_{VCO}}{1mW}\Big) \tag{3.9}$$

where, $\mathscr{L}_{VCO}(f_m)$ is the phase noise of the VCO, f_m is the offset frequency at which the phase noise is measured, f_{osc} is the oscillation frequency of the VCO and P_{VCO} is the power dissipation. The unit of FOM_{VCO} is in dBc/Hz. A FOM of -160 dBc/Hz was used as a target specification to derive the phase noise requirements on the VCO at 1 MHz offset frequency and 1 GHz oscillation frequency based on Equation 3.9. With a power budget of about 200 µW for the VCO, the resulting phase noise requirement was -93 dBc/Hz at 1 MHz offset frequency. Finally, $FWHM_{total}$ is obtained as $\approx$ 80ps by evaluating Equation 3.7. With this estimation, the TDC resolution was still enough for the application requirement of a few millimeters accuracy.

The architecture of the TDC is based on a RO-based circuit, where a 8-stage pseudo-differential current-starved ring oscillator provides a 4-b fractional resolution, sampled by sense-amplifier flip-flops. Always-on TDC is independently accessible by both subgroups and therefore, there are two sets of SAFFs with independent sampling lines. The RO schematic along with SAFF arrangement is shown in Figure 3.7 where the oscillator frequency itself is controlled by a PMOS current source.

Nominally desirable to operate at 1 GHz frequency, the RO clocks a standard 10-b asynchronous counter thus, providing a total range of 14-b with 1 µs temporal range and 61 ps resolution. The counter schematic, also shown in Figure 3.7b, has every bit clocked by its previous stage and resulting delay accumulated through this chain may give rise to sampling errors. This is circumvented by resampling flip-flops which are clocked by the same input clock through a chain of buffers.

It is sufficient to guarantee that the buffer delay is shorter than the DFF propagation time and large enough to compensate the DFF delay errors since the input clock is nominally about 1 GHz. This is easily achieved since buffer delays are typically shorter than DFF delay when technology-library standard cells are used. The sampling lines from the decision tree, dTOF signals, also propagate through exact copies of the aforementioned structure of buffer+DFF to provide the final 14-b TDC code. The layout of the implemented TDC is shown in Figure 3.7c. The TDC occupies an area of 550 µm^2 shared between two subgroups. About 40% of the TDC area is dedicated to decoupling capacitors, while providing an equalized and calibration-free binary output. The TDC is periodically sampled using an external signal for calibration which is performed off-chip. Due to continuous operation, the TDC consumed a constant current and its power consumption does not depend on the activity. Thus, the main purpose of calibration is to track slow variations. Moreover, larger sensor arrays, designed using several such modules, can be synchronized by mutually coupling the TDCs which reduces the burden on calibration as will be analyzed in Chapter 5.
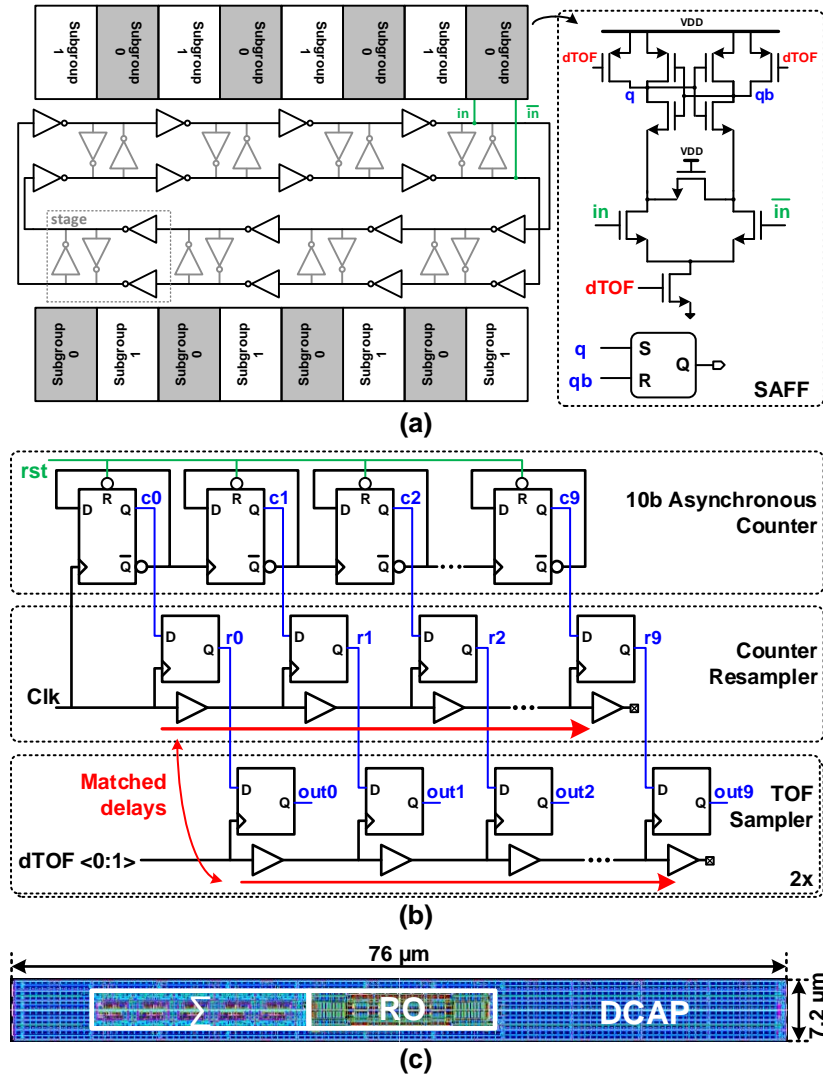
Figure 3.7 – TDC block diagram. (a) Pseudo-differential stages and SAFF arrangement for the two subgroups, (b) Counter schematic, (c) Layout.

3.2.3 Digital processing and communication unit (DPCU)

Figure 3.8 shows the block diagram of the subgroup including the digital processing blocks and the already described elements. The DTOF signal and the ID from the decision tree is fed to the DPCU where the DTOF signal itself acts as a clock and the ID is used to access the pixel memory where a prior stored information is read and the new timestamp data sampled by the DTOF is combined. The combined result of the current processing information is then stored at the next arriving uncorrelated event. A simplified timing diagram is shown in Figure 3.8. The arithmetic and logic unit (ALU), at the core of the DPCU provides dual functions by optionally switching between its low-pass-filtering and intensity counting features. The low-pass filter is a classic digital infinite impulse response (IIR) whose function is to reduce uncertainty over multiple events between readouts, by providing a result around their average value, whose frequency characteristics are described by the following

equation;

$$y[k] = (1 - \lambda) \cdot y[k-1] + \lambda \cdot x[k] \tag{3.10}$$

where, λ is the attenuation factor. However, such filtering is effective only in a low-noise background scenario. Otherwise, a noise-suppression technique is required to firstly eliminate background noise and then this filtering may be applied. Therefore, this technique in the current sensor is not suitable for LiDAR applications where high background noise is an inevitable challenge. This feature will hence not be explored any further and the primary purpose of the ALU will be limited to intensity counting.
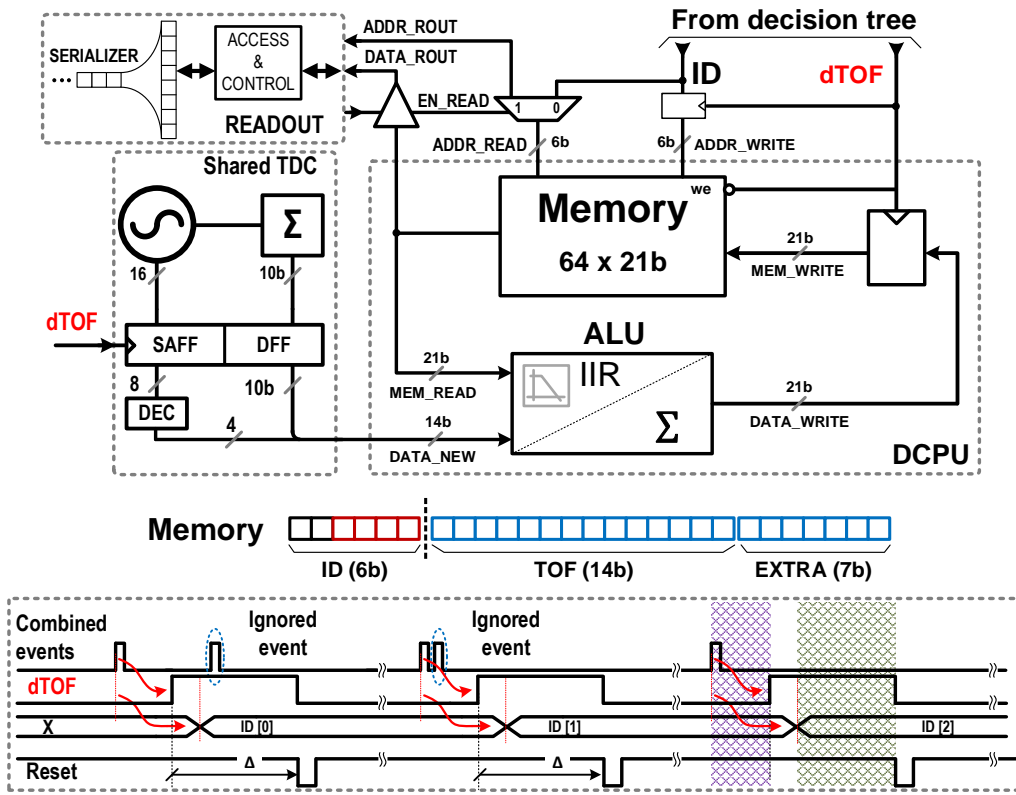
Figure 3.8 – DPCU block diagram for subgroup with the shared TDC and timing diagram.

The 21-bit in-pixel memory hosts the 14-b TOF information from the TDC and configurable 7-b intensity counter information. The 6-b ID information simply acts as a pointer to the memory, not requiring separate storage. The memory array is generated using custom-designed 1-bit SRAM circuit, schematic of which is shown in Figure 3.9. Read and write times achieved are 1.6 ns and 100 ps respectively, which are both minimized by utilizing tri-state buffers, capable of driving the whole bank without additional use of sense amplifiers or comparators. The overall organization and access to memory in shown alongside in Figure 3.9.
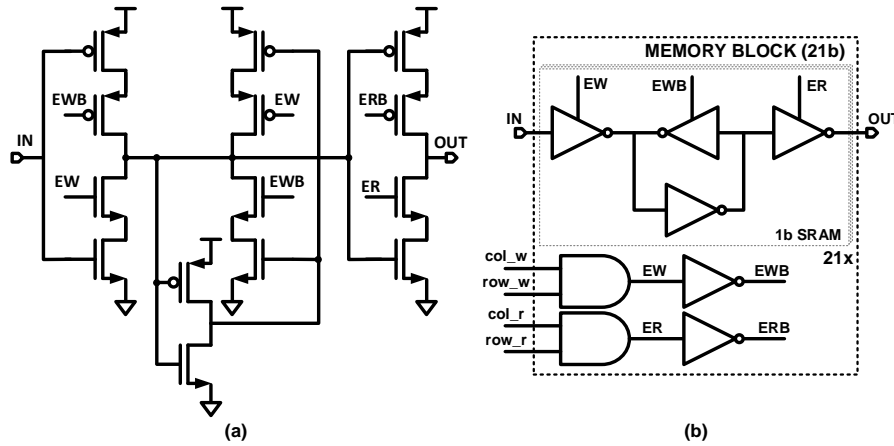
Figure 3.9 – Custom-designed pixel memory- (a) single-ended, tri-state SRAM and (b) 21-bit block memory per pixel.

3.2.4 Laser signature

Multiple LiDAR systems could co-exist in a real scenario where every system may present itself as interference to each other (Section 1.3.3, Chapter 1), although the reasonable occurence of such a situation is currently low. In order to circumvent such scenarios, many solutions have been proposed in the past based coded modulation techniques (CDMA) [12] and using pseudo-random sequences of the illumination to improve robustness in a multi-camera environment [13]. However, most of these techniques cost high computation and power, thus adding to system latency.
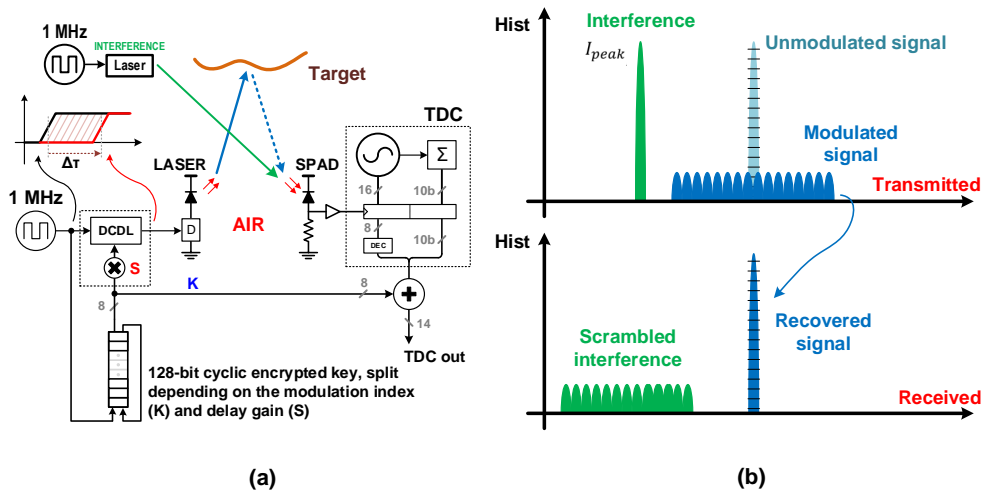
Figure 3.10 – (a) Laser signature concept- Implementation via encrypted key, divided according to modulation index and directly combined with digital TDC output and (b) Laser signature histogram.

Alternatively, a simpler laser signature is implemented directly on the laser trigger, by adding a digitally controlled delay line (DCDL) to it as well as to the TDC timestamp using arithmetic calculation. The principle is derived from a typical pulse position modulation (PPM) technique, shown in Figure 3.10. The discrete nature of the system allows controlling the position of the pulse with a known value which can then be used to recover the received signal distinctly without loss of information while scrambling down the interferences to lower levels. The conceptual histogram representation of such a technique is explained in Figure 3.10b. where the outgoing laser is spread into 16 equidistant chunks while the intereference signal is oblivious to this. On the receiver (DTOF sensor system), the modulation is applied to every detected signal and as a result of this, the interference signal is scrambled down to lower levels and the desired TOF information is recovered distinctly.

The number of discrete laser positions is defined by the modulation index K and the the amount of time shift itself, by the gain, G. The DCDL is implemented on the FPGA-based PLL where a desired time-shift can be provided. The same concept can however be extended to an integrated implementation on-chip. The block diagram of the implementation is shown in Figure 3.10. In order to provide maximum spectrum efficiency or synonymously, interference suppression over spread in histogram, the delay offset, produced by the modulation, should correspond to the system uncertainty (FWHM). Furthermore, the modulation is preferably chosen to be a multiple of the TDC LSB (Δ_{LSB}) for ease of correction. The delay gain S, see Figure 3.10, should therefore be chosen as the next integer of Δ_{LSB}, either in number of histogram bins or seconds, as:

$$S = \left\lfloor \frac{FWHM}{\Delta_{LSB}} \right\rceil \quad \text{and}$$
$$\Delta\tau = S \cdot K, \quad (3.11)$$

where $\Delta\tau$ is the time delay, in picoseconds, applied to the laser trigger. The index K, is selected based on the number of discrete positions to be utilized which reaches up to 8 bits in this implementation (256-PSK). In order to increase security, a unique 128-bit encrypted key can be added to the system, and subdivided in words of 8 or less bits, depending on K. If optimized, the system provides interference suppression of about $20 \cdot \log_{10}(0.89 \cdot K)$ [2].

3.3 Characterization results

The proposed sensor was implemented in a 3D-stacked BSI technology with 45 nm CIS used for the design of SPADs on the top tier and 65 nm low-power CMOS technology utilizing 5 metal layers for the readout circuit (ROIC) on the bottom tier. The fabricated die was packaged onto a ceramic QFP-120P package to be inserted onto a zero-insertion-force socket. The chip micrograph is shown in Figure 3.11. As can be seen, the ROIC on the bottom tier is not visible due to the use of 3D-stack technology and only the top tier with circular SPAD array is visible. The hybrid bonding connection between the two tiers, occupies 5 % of the pixel area, leaving behind rest of the area for laying out an equally-distributed power mesh created using top metals of both tiers, dedicated to reducing

excessive IR-drops. The results described through the following sections are characterized using a 532 nm PicoQuant VisUV for depth measurements, and a 637 nm ALDS PiL063X for SPAD characterization and laser signature.
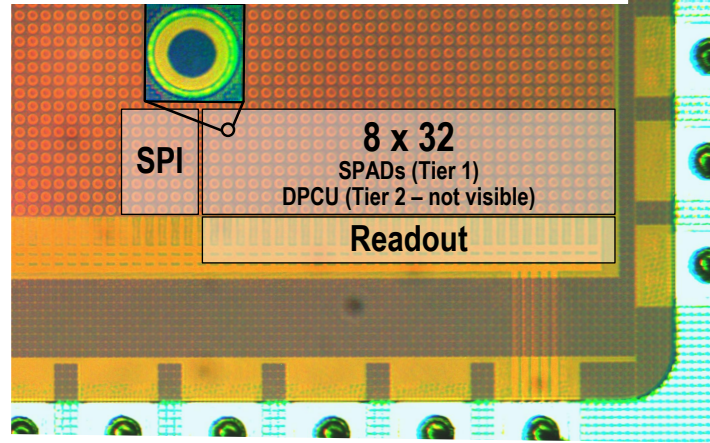
Figure 3.11 – Photomicrograph of the sensor

3.3.1 SPAD characterization

SPADs used in this sensor were desrcibed in Chapter 2 where the characterization results were also reported in Section 2.6.1. The SPADs exhibit a 108 ps FWHM timing jitter with a 31.8% peak photon detection probability (PDP) at 600 nm (see Figure 2.25). The dark-count rate (DCR) when operating under excess bias voltage (above breakdown) of 2.5 V is 55 cps/µm^2 (Figure 2.24), adequately suitable for a LiDAR application where background noise is a major source of noise.

Since space sector was one of the target LiDAR applications, the suitability of the implemented SPADs was analyzed by testing them under high dosage of radiation. A Co-60 Gamma source was used to irradiate the sensor and the DCR performance of the SPAD was monitored, as shown in Figure 3.12.

The DCR increases from 2.8 to 5.8 kcps at a dose rate of 73 krad/h over a 90-min exposure and returned to the original value after annealing. The applied dose is much higher than required, thus, allowing the possibility for further investigations on use of this sensor for space applications.

3.3.2 Depth measurements

Depth performance was evaluated by characterizing the sensor with single-point ranging measurements. These measurements were performed using flat targets with uniform reflectivity of 50 %, placed perpendicular to the sensor optical axes. In the first mode, the TDCs in the sensor operate in high resolution mode, providing $\Delta_{[LSB]} = 61\,ps$ over a 14-b TDC range of 1 µs and equivalently, up to 150 m in distance.

(a)   **(b)**
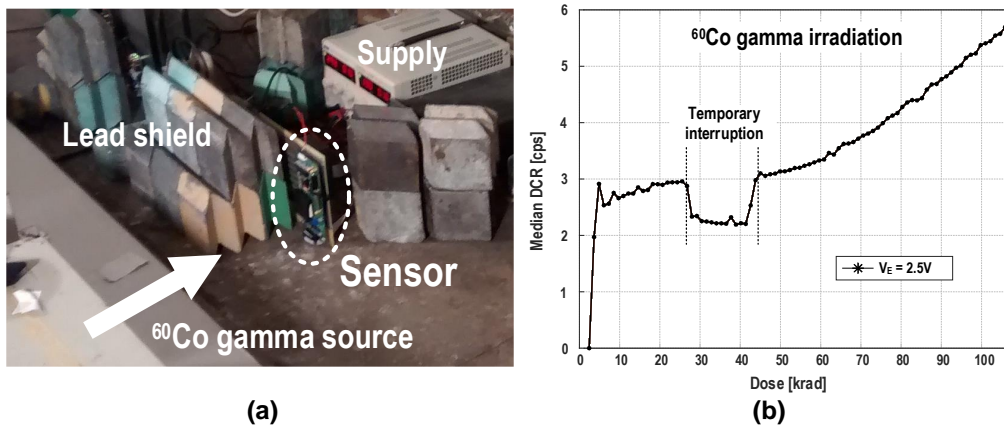
Figure 3.12 – Irradiation measurement, (a) setup and (b) DCR increase with accumulated dose.

(a)   **(b)**
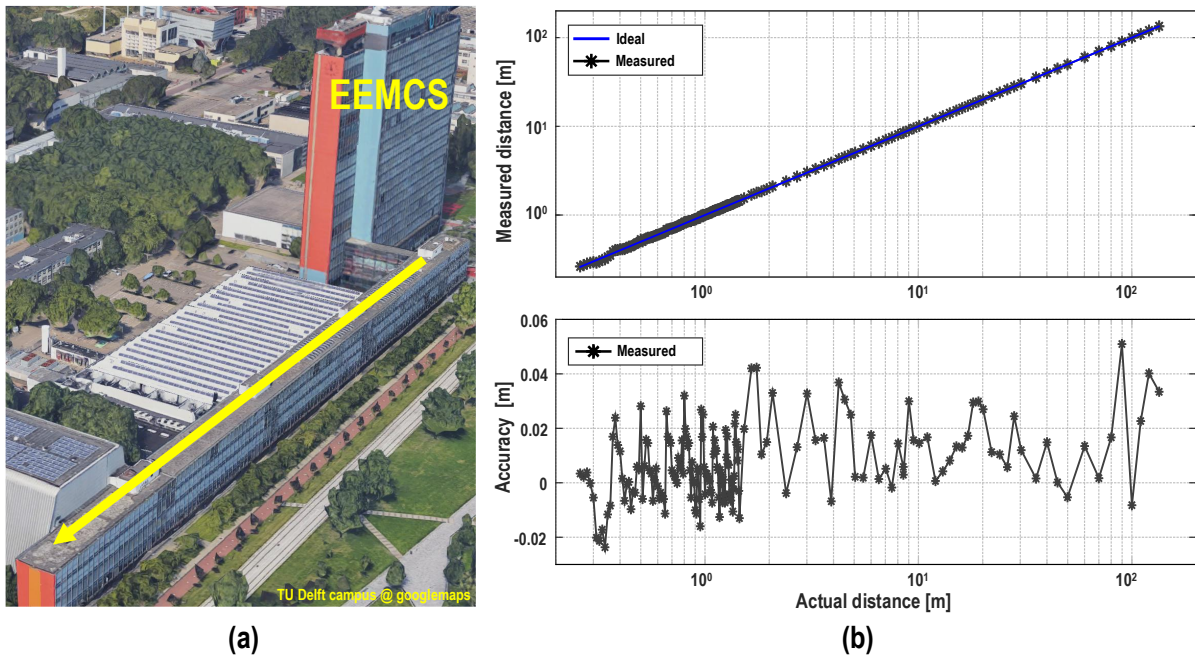
Figure 3.13 – High-resolution range measurement- (a) aerial view of measurement location and (b) Measured distance and accuracy.

Alternatively, a second low-resolution mode allowed extension of the range measurements up to 500 m, by tuning the ROs in the TDC to provide a resolution of 204 ps, covering a temporal range of 3.34 μs. The mean laser power used was 4 mW at 1 MHz repetition rate in the high resolution measurements and 1.4 mW at 300 kHz repetition rate in the low resolution mode. An approximately constant 4 nJ energy per pulse is maintained at a pulsewidth of 80 ps FWHM and 47 W peak power. Each range measurement point was obtained by accumulating 100 chip readouts and by combining information from all pixels, operating the 128 pixels in the module like a digital SiPM. The histogram was then calculated in MATLAB without any post-processing (such as filter). The maximum chip readout is limited to 2000 fps, resulting in 20-fps depth measurements. All measurements were physically performed, where high resolution mode characterization was done indoor along a corridor

of known length using a portable optical bench (see aerial view in Figure). The resulting accuracy and precision of these measurements are reported in Figure 3.13, where a maximum accuracy error (deviation from ground truth) of less than 7 cm (0.3 % nonlinearity) was measured and a precision (worst-case standard deviation) of 15 cm (0.1 % nonlinearity).

The low resolution mode was characterized outdoor, the aerial view of which is shown in Figure 3.14. The maximum accuracy error of 80 cm (0.3 % nonlinearity). A code-density test performed on the TDC showed a differential non-linearity (DNL) of less than 2 LSB and integral non-linearity (INL) of less than 3 LSB. This linearity arose from the mismatches between the sampling signals and phases of the RO+counter. Although calibration helps alleviating these issues to an extent improving the results, there was no calibration performed in this design due to tight area and power constraints within the module.

(a)   (b)
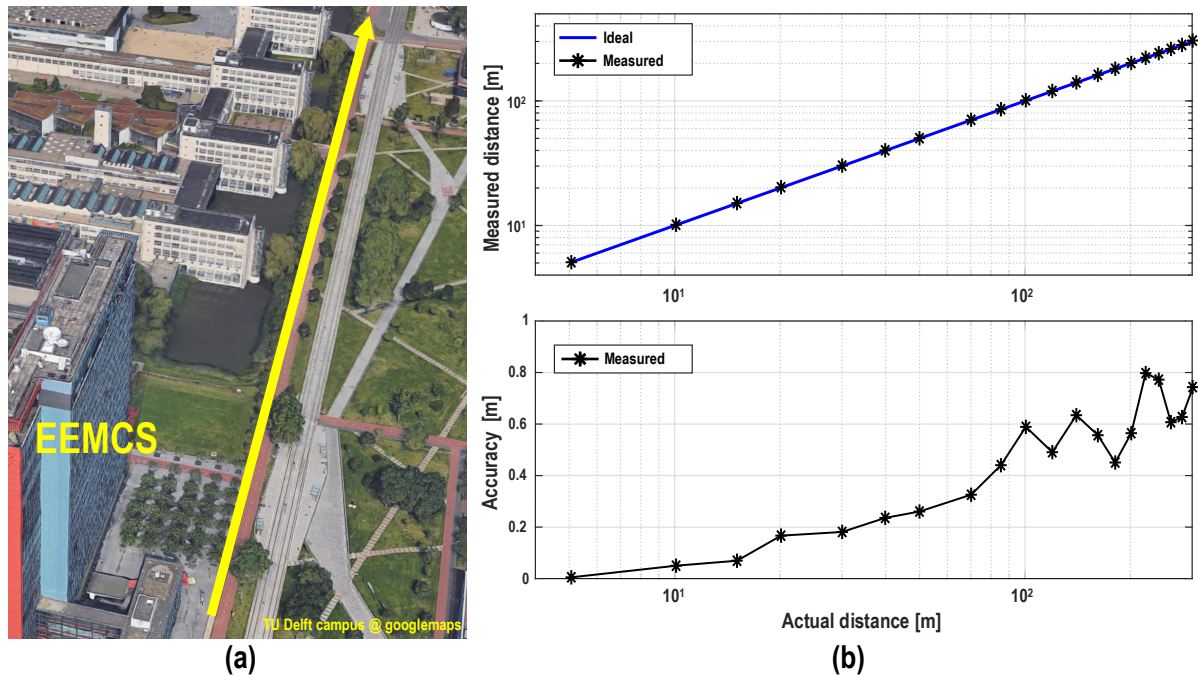
Figure 3.14 – High-resolution range measurement- (a) aerial view of measurement location and (b) Measured distance and accuracy.

3.3.3 Laser signature

Laser signature technique described in Section 3.2.4, was measured at three indices of K, i.e., K = 2^3, 2^4, and 2^5 and gain, S = 16.Δ_{LSB} using two lasers– a primary laser at 532 nm wavelength, focused directly onto the sensor and a 637 nm laser acting as an interference.

Absence of color filters in these measurements increased the noise floor due to background sunlight. Therefore, measurements in Figure 3.15a show the effects of PPM without any background illumination, where interference suppression of close to expected value ($20\log_{10}0.89K$) was measured.

Figure 3.15b shows the PPM applied with external background light of 3 klux where the interference suppression is less effective due to the increase in the noise floor which adds to the bias level of the acquired histogram and secondly, the nature of decision tree causes collisions between noise and signal events, where noise is propagated more often, thus, lowering the overall signal peaks (primary laser and the interference).
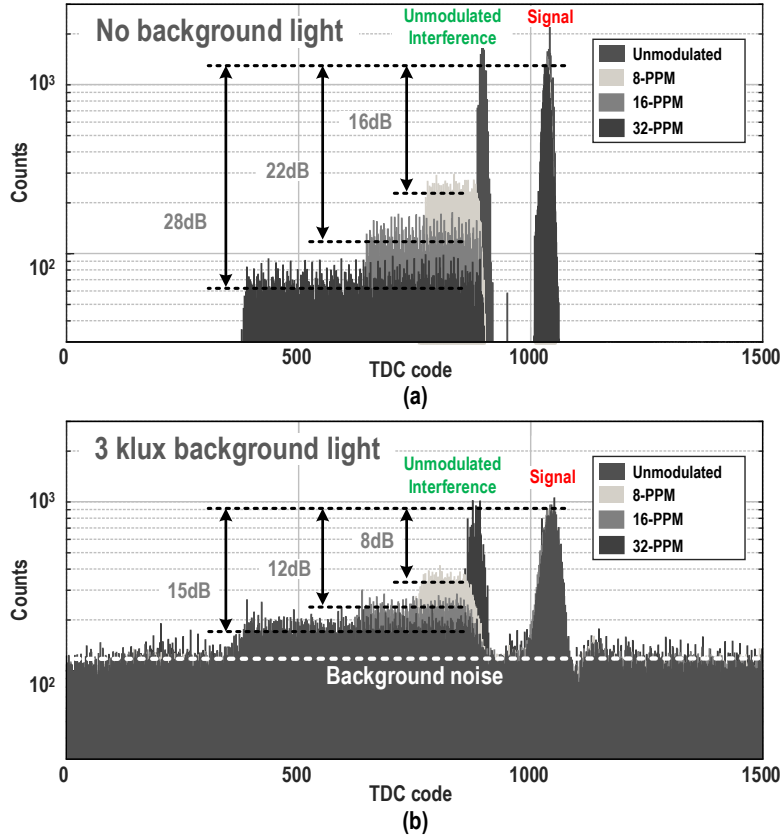
Figure 3.15 – Laser signature measurement- (a) no background illumination and (b) 3 klux background illumination.

3.3.4   3D image reconstructions

A dual-axis laser scanner was used to obtained flexible lateral resolutions while reconstructing 3D images. Particularly, due to the small size of the module, scanning was necessary for reasonable spatial resolutions. A Thorlabs Large Beam Diameter Galvo scanner GVS212, with broadband mirrors was controlled using a standalone waveform generator. The control was implemented via MATLAB. The scanner was synchronized with the laser and readout to provide spatially efficient illumination and reconstructions. Figure 3.16 shows a 32 × 32 image of wide dynamic range scene featuring targets with varying reflectivities from as low as 8 % on the black sections of the wall to as high as 60 % on a white pillar, ranging from 4 to 10 m and about 30° FOV.

The integration time per point of acquisition was 5 ms (or 10 chip readouts) resulting in 1280 TOF

measurements per point over the module of 128 pixels. As can be seen from the depth map as well as its cross-section along row 30, the acquired TOF measurements were successfully obtained irrespective of varying target reflectivities.

Different target reflectivities
[1] – White wall – 50%
[2] – Black wall – 8%
[3] – White pillar – 60%
[4] – Aluminum bin – 54%
[5] – Cardboard box – 21%
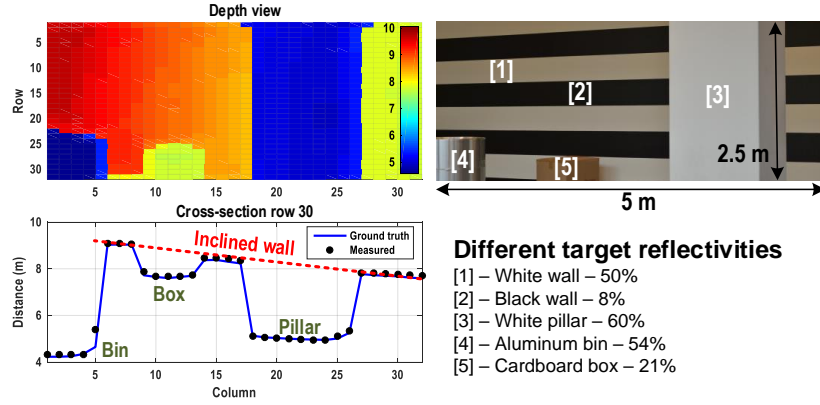
Figure 3.16 – A 32×32 image featuring multiple targets with different reflectivities.

A 256 × 256 fine resolution 3D image was also reconstructed with a 7° FOV. A higher target reflectivity and a smaller FOV enabled a smaller integration time, reaching only 0.5 ms per point in an SiPM-like module operation. Although the SPI-based data readout offered flexibility in sensor control, this also limited the maximum data throughput. Consequently, a maximum chip readout of 2000 fps was obtained, requiring 32 s to obtain the full 3D image shown in Figure 3.17. The TOF and intensity information were acquired simultaneously where the depth map shown in the figure is an effective superimposition of intensity and TOF. Table 3.1 shows the performance comparison of this sensor with the state-of-the-art LiDAR-targeted designs. Reference [14] operates using a 870 nm laser, thus considerably reducing the integrated solar noise and uses a narrow bandpass filter with a 40 mW laser which enhances the SBR. Reference [15] uses the same wavelength as in the above measurement and has advantages of filtering background noise by using its smart-triggering feature, however, still obtaining results with a high power laser guided through a fiber over emulated distances.
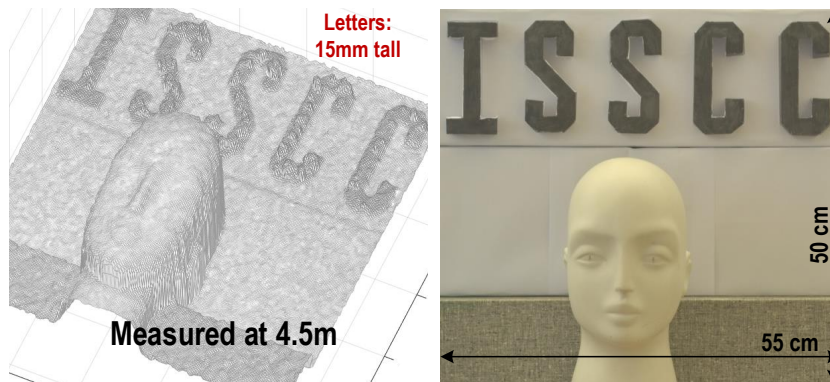
Figure 3.17 – A 256×256 depth data superimposed with intensity image.

Table 3.1 – Performance comparison of state-of-the-art DTOF sensors (2018)

Parameter	Unit	**This Work**	[15]	[14]	[16]	[17]
Technology	–	45/65 nm CMOS	150 nm CMOS	180 nm CMOS	130 nm CIS	0.35μm CMOS
Architecture	-	Always-on, shared TDC	Start/Stop, per-pixel TDC	Column-wise shared TDC	Histogramming shared TDC	Start/Stop, per-pixel TDC
Sensor characteristics						
Pixel count	-	8×16^a	64×64	340×96	32×32	32×32
Pixel pitch	μm	19.8	60	25	21	150
Pixel fill factor	%	31.3	26.5	70	43	3.14
SPAD DCR@V_E	cps/μm^2	55.4 @ 2.5 V	57 @ 3 V	6 @ 3.3 V	N/A	120 @ 6 V
TDC depth	bit	14	16/15	12	8	10
TDC resolution	ps	61 – 204	250 – 20000	208	71.4	312
TDC power	mW	0.5 – 0.2	N/A	N/A	14.1	0.35/pixel^f
TDC area	μm^2	550	N/A	31,000^d	30,000	5,600^d
TDC linearity — DNL [LSB]		+0.9 / -1	+1.2 / -1^b	+0 / -0.52	+0.75 / -0.61	+0.06 / -0.06
TDC linearity — INL [LSB]		+3 / 0	+4.8 / -3.2^b	+0.73 / -0.49	+0.65 / -0.2	+0.22 / -0.22
Measured distance performance						
Distance range	m	150 – 300	367 – 5862^c	128	2.82 – 3.375	48
Precision — m		0.15 – 0.47	0.2 – 0.5^c	0.1^e	N/A	0.04^g
Precision — %		0.1 – 0.11	0.13 – 0.14^c	0.1^e	N/A	0.8^g
Accuracy — m		0.07 – 0.8	1.5 – 35^c	0.37^e	N/A	N/A
Accuracy — %		0.3 – 0.4	0.37 – 1.9^c	0.37^e	N/A	N/A

^{a} Up to 256×256 resolution achieved by flexible scanning system. ^{b} Measured over 5% of the total range. ^{c} Emulated results with optical fiber.
^{d} Estimated by layout. ^{e} Measured at 100 m. ^{f} DLL and TDC power. ^{g} Measured at 5 m.

3.4 Challenges with decision-tree based DTOF sensor

The design based on decision tree (DT), while offering an effective way to combine events from multiple pixels, has its own limitations. The primary limitation is its nature of winner-take-all propagation. A typical LiDAR system is overwhelmed with high background noise and operates predominantly in a low SBR regime. The winner-take-all characteristic of the decision tree is naturally prone to propagating more noise events compared to signal events on an average, thereby keeping the tree busy with noise events most often and resulting in saturation of the sensor sometimes. A more detailed analysis on this decision-tree based architecture provides more insight on this inherent limitation. Consequently, a robust alternative DTOF sensor architecture is required to operate under extreme background noise condition, often reaching up to 50–100 klux from bright solar irradiation.

3.4.1 Analytical model of a DTOF Sensor in a flash LiDAR

With the established sensor architecture and performance through previous sections, an analytical model of the same design is firstly written on MATLAB to identify various challenges in the DT-based design. A flash LiDAR scenario is constructed using the DT-based sensor and the findings of this model are then used to migrate towards an alternative sensor architecture, to be discussed in the next chapter.

For the DT-based design, the photon-detection process is first modeled analytically and the probability of detection is calculated from noise and signal events. The spatial arrangement of the pixels is exactly the same as before, with a subgroup clustered into an array of $M = 8 \times 8$ SPADs. For analysis purposes, the subgroup is arrayed to scale up to a spatial resolution of 32×32 SPAD pixels while there is no particular limitation of scaling to larger formats. The analysis is performed only on a single subgroup due to the modularity of the sensor architecture which consists of multiple identical subgroups. A high-level block diagram of the subgroup with the shared TDC is shown in Figure 3.18.
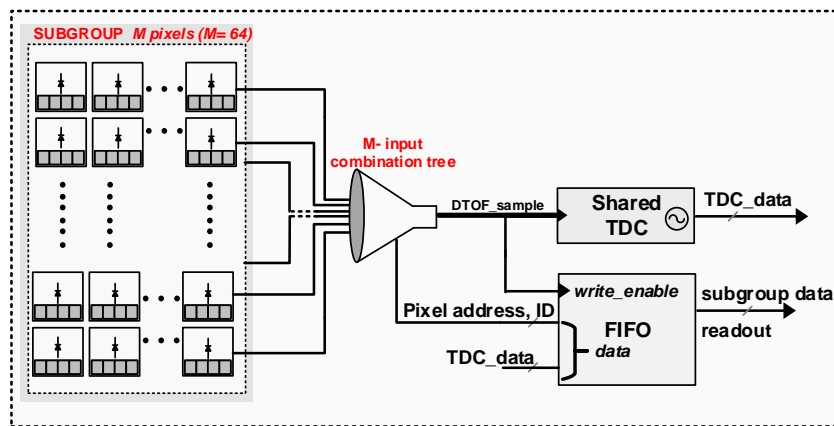
Figure 3.18 – A block diagram of a DTOF sensor in a shared architecture.

The combination tree is exactly modeled like a decision tree circuit implemented in the sensor (seen in Section 3.2), with its dead-time, $t_{d,comb}$, after which it resets itself making it available for successive detections. As already discussed, in a shared architecture as this, the total dead time, t_d for detection is a contribution of $t_{d,comb}$ from the DT circuit and the SPAD dead time, $t_{d,spad}$. The overall SPAD dead time in a shared case reduces to $\approx t_{d,spad}/M$, thus making $t_{d,comb}$ the dominant contributor. Consequently, the maximum TDC conversion rate in a shared architecture as this is given by the inverse of the tree dead time, $\approx 1/t_{d,comb}$, which can reach up to Gtimestamps/s. The ID information of the first event along with the associated timestamp (TDC code) is modeled to be read out using a digital readout logic, such as a first-in-first-out (FIFO) bus, as shown in Figure 3.18.

The DT-based architecture is modeled within a flash scenario as shown in Figure 3.19.
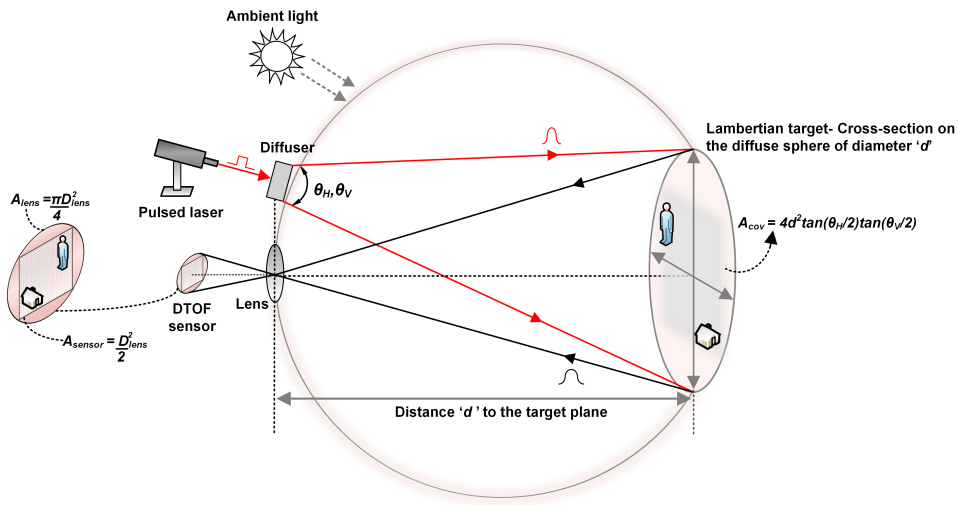
Figure 3.19 – Flash LiDAR operation.

As seen in Figure 3.19, one can estimate the effective SBR from the number of noise events versus the number of signal events on a per-pixel basis. Background noise is modeled with Planck's law of blackbody radiation and Poisson statistics. Assuming a solar irradiance, P_{solar} W/m^2, the returning power per pixel (in units of Watts) back-reflected from a flat target with uniform reflectivity, r, received through the lens with an efficiency, T_l and filtered using an optical bandpass filter with a passband wavelength, Δ_{bw} and efficiency, T_f is written as follows

$$P_{noise,pixel} = P_{solar} \cdot A_{cov} \cdot r \cdot \left(\frac{D_{lens}}{2d}\right)^2 \cdot T_l \cdot T_f \cdot \Delta_{bw} \cdot \left(\frac{2}{\pi}\right) \cdot \left(\frac{1}{N}\right), \tag{3.12}$$

where, N is the number of pixels in the sensor. It has to be noted that due to the rectangular geometry of the sensor array, the entire area projected by lens which is circular, is not entirely useful. The effective area is calculated by multiplying the above equation by the fraction, $A_{sensor}/A_{lens} = 2/\pi$, as shown in Figure 3.19. On expanding the term, A_{cov} as seen in Figure 3.19, it is observed that $P_{noise,pixel}$ is independent of the distance to the target, d, as expected. The reflected power per

pixel from the laser pulse, with an average power, P_{avg} is similarly estimated as follows,

$$P_{signal,pixel} = P_{avg} \cdot r \cdot \left(\frac{D_{lens}}{2d}\right)^2 \cdot T_l \cdot T_f \cdot \Delta_{bw} \cdot \left(\frac{2}{\pi}\right) \cdot \left(\frac{1}{N}\right). \tag{3.13}$$

Unlike noise, the returning power of the signal is dependent on the distance, d and decreases with d^2, following the inverse square law. Given a certain photon detection probability (PDP) for the SPAD and fill-factor, FF, Equations (3.12) and (3.13) reduce to

$$P\_eff_{noise,pixel} = P_{noise,pixel} \cdot PDP \cdot FF, \tag{3.14}$$

$$P\_eff_{signal,pixel} = P_{signal,pixel} \cdot PDP \cdot FF. \tag{3.15}$$

From the wavelength, λ_{laser}, of the laser, the number of noise and signal events per second is estimated by dividing Equations (3.14) and (3.15) by the energy of the photon at λ_{laser},

$$N_{pixel} = \frac{P\_eff_{noise,pixel}}{hc/\lambda_{laser}}, \tag{3.16}$$

$$S_{pixel} = \frac{P\_eff_{signal,pixel}}{hc/\lambda_{laser}}, \tag{3.17}$$

where c is the speed of light (3×10^8 ms^{-1}) and h, Plancks's constant (6.626×10^{-34} Js).

SPADs have a certain dead time between successive detections which limits the theoretically estimated photon-count statistics. This dead time itself can be paralyzable or non-paralyzable in nature [9]. A non-paralyzable dead time, t_d, is assumed for all analyses in this thesis. Consequently, the event rates in Equations (3.16) and (3.17) are modified to provide the effective rates as follows,

$$N\_eff_{pixel} = \frac{N_{pixel}}{1 + N_{pixel} \cdot t_d}, \tag{3.18}$$

$$S\_eff_{pixel} = \frac{S_{pixel}}{1 + S_{pixel} \cdot t_d}. \tag{3.19}$$

For simplicity, we will continue using the terms, N_{pixel} and S_{pixel}, for the noise and signal event rates respectively. Based on the established equations, the effective noise and signal events per pixel per second are simulated for a flat target of $r = 10\%$ reflectivity over varying distances. An average laser power, P_{avg}, of 20 mW, wavelength, $\lambda_{laser} = 780$ nm and a repetition rate of 1 MHz and a pulsewidth of ≈ 500 ps were assumed to be uniformly illuminating the target with FOVs, $\theta_H = 20°$ and $\theta_V = 20°$. A 780 nm laser wavelength is used for analysis purposes simply because of the practical availability of such a laser in view of future measurements using that laser. However, a common choice for LiDAR is working with longer near-IR wavelengths, which would naturally perform better under high solar exposure compared to a 780 nm laser. A lens with with diameter, $D_{lens} = 11$ mm and a f-number of 1.4 (focal length ≈ 15 mm) is assumed to collect light onto the

assumed 32×32 SPAD sensor ($N = 1024$). Additionally, background light of different levels ranging from 5 klux ($\approx 50\ W/m^2$) to 100 klux ($\approx 1000\ W/m^2$) is imposed and simulations are performed. A summary of common simulation parameters used through the analysis is mentioned in Table 3.2.

Table 3.2 – Simulation parameters

Parameter	Value
Average laser power, P_{avg}	20 mW
Laser wavelength, λ_{laser}	780 nm
Repetition rate , f_{laser}	1 MHz
Total system FWHM	530 ps
Target reflectivity, r	variable, 8–60 %
Field-of-view, FOV	15°–40°
Background light	variable, 5–100 klux
Sensor resolution	32 × 32
SPAD detector PDP	10 %
Pixel fill-factor, FF	50 %
Diameter of collecting lens, D_{lens}	11 mm
f-number, f#	1.4
focal length, f	15 mm
Lens efficiency, T_l	0.8
Optical filter passband, Δ_{bw}	20 nm
Filter efficiency , T_f	0.7

Figure 3.20 shows the resulting noise and signal rates indicated per laser pulse per pixel. No particular noise filtering mechanism has been modeled for this simulation. As can be seen in Figure 3.20b, beyond 1 m at 100 klux background light and beyond 2 m at 50 klux background light ($\approx$ 503 W/m²), the system starts approaching a negative SBR regime within the assumed flash LiDAR conditions.
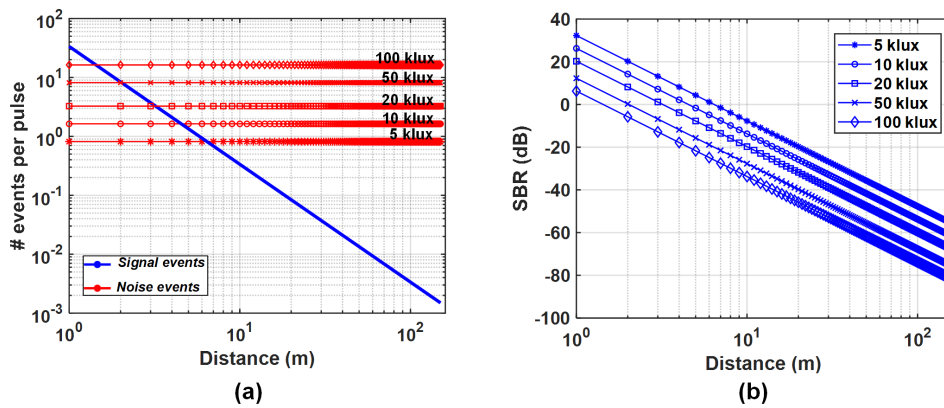
Figure 3.20 – Simulation results of (**a**) the number of events per pixel per laser pulse at different background noise levels and (**b**) the SBR for 1–150 m target distances, d.

Following this, the DT-based sensor is evaluated in the given flash scenario and its performance is simulated, where, the probability of signal detection is estimated analytically for the flash scenario and the count statistics per pixel in Equations (3.16) and (3.17) are used to estimate the number of signal events with one laser pulse for every pixel.

$$N_{pulse,pixel}(i) = N_{pixel}(i) \cdot t_{meas}, \tag{3.20}$$

$$S_{pulse,pixel}(i) = S_{pixel}(i) \cdot (1/f_{laser}). \tag{3.21}$$

The number of signal photons are directly related to the repetition rate of the laser; i.e., the lower the repetition rate, the higher is the energy per laser pulse, in that, all the laser photons are concentrated within the pulsewidth of the laser (FWHM) over the laser period following a Gaussian distribution (duty cycle ratio = $FWHM \cdot f_{laser}$). Whereas, the number of integrated noise photons are uniformly distributed and need to be calculated based on the actual measurable window, t_{meas}, as shown in Equation (3.20).

Due to the presence of SPAD dead time, the incoming photon events, otherwise modeled as Poisson arrival processes need to be modified to account for this [9]. The probability of detecting k number of events over a measurable window, t, taking the SPAD dead time into account, is,

$$p(k) = \frac{(\lambda(t - kt_{d,spad}))^k \exp(-\lambda(t - kt_{d,spad}))}{k!}; t_k < t - t_{d,spad}, \tag{3.22}$$

$$p(k) = \frac{(\lambda(t_k - (k-1)t_{d,spad}))^k \exp(-\lambda(t_k - (k-1)t_{d,spad}))}{k!}; t_k > t - t_{d,spad}, \tag{3.23}$$

where λ is the average photon arrival rate (not to be confused with λ_{laser}, the wavelength of the laser) and t_k is the arrival time of these photons. Equation (3.23) is for the case when $kt_{d,spad}$ may fall outside the measurable window, t, arriving at a time t_k. For the given DT-based sensor architecture, the subgroup detects and propagates every first event ($k = 1$) in an array of M pixels. A constant dead time of $t_{d,comb}$ is assumed during which the subgroup cannot detect any event and at the end of this duration, it is assumed that it is ready to detect the next incoming event. Thus, any detection within the subgroup following this time instant will now include the absolute dead time of the SPAD, $t_{d,spad}$. Under these conditions, Equations (3.20) and (3.21) are modified to include the dead time of the SPAD as follows,

$$N_{pulse,pixel}(i) = N_{pixel}(i) \cdot (t_{meas} - t_{d,spad}), \tag{3.24}$$

$$S_{pulse,pixel}(i) = S_{pixel}(i) \cdot (1/f_{laser} - t_{d,spad}). \tag{3.25}$$

The probability of detecting a noise event per pixel within a subgroup will be determined by the

probability that the combination tree propagates this event through the subgroup. For every pixel, i in the subgroup, this is given by the conditional probability of detecting a noise photon in pixel, i, given that no other noise event has been propagated through rest of the subgroup. In order to calculate this, the total number of photons per pulse within the subgroup, sg, is first estimated. The number of photons in a subgroup, sg, with M number of pixels, is calculated as a summation of the pixel-wise photon number for noise and signal per laser pulse.

$$N_{pulse,sg} = \sum_{i=1}^{i=M} N_{pulse,pixel}(i), \tag{3.26}$$

$$S_{pulse,sg} = \sum_{i=1}^{i=M} S_{pulse,pixel}(i). \tag{3.27}$$

The probability of detecting a noise event in pixel, i is given by

$$p_{n,pixel}(i) = N_{pulse,pixel}(i) \exp(-N_{pulse,pixel}(i)). \tag{3.28}$$

The probability of detecting a noise event in subgroup sg in rest of the $M-1$ pixels excluding the ith pixel is given by

$$p_{n,sg}(i) = (N_{pulse,sg} - N_{pulse,pixel}(i)) \exp(N_{pulse,sg} - N_{pulse,pixel}(i)). \tag{3.29}$$

The conditional probability is then calculated which provides the final effective probability of detecting and propagating a noise event through the subgroup.

$$p_{n,eff}(i) = p_{n,pixel}(i) \cdot (1 - p_{n,sg}(i)). \tag{3.30}$$

For the signal events, the final conditional probability can be calculated as follows

$$p_{s,eff}(i) = (1 - p_{n,pixel}(i)) \cdot (1 - p_{n,sg}(i)) \cdot p_{s,pixel}(i), \tag{3.31}$$

where $p_{s,pixel}(i)$ is,

$$p_{s,pixel}(i) = S_{pulse,pixel}(i) \exp(-S_{pulse,pixel}(i)). \tag{3.32}$$

For the flash LiDAR scenario described in Section 3.4.1, the probabilities of noise ($p_{n,eff}$) and signal ($p_{s,eff}$) detections are calculated for $k = 1$, where the DT in the subgroup propagates every first incoming event. Figure 3.21 shows the MATLAB simulation results. All simulation parameters remain the same as mentioned in Section 3.4.1 except FOVs which are now increased to $\theta_H = 40°$ and $\theta_V = 40°$. The integration window for noise events is $t_{meas} = 20$ ns. A 50 klux background light condition is assumed for most analysis unless specified otherwise.
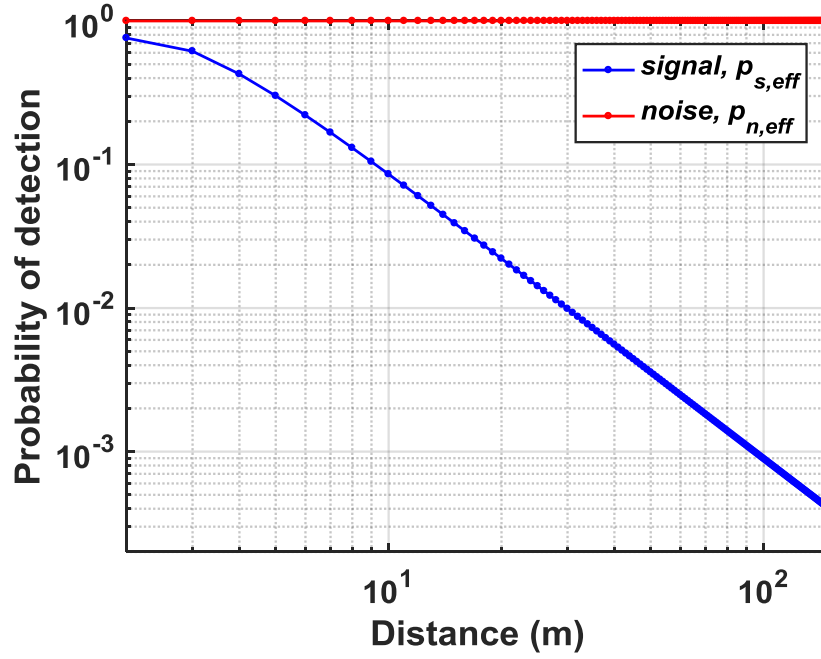
Figure 3.21 – Probability of detecting signal and noise events in a flash scenario using DT-based DTOF scheme.

As can be seen, due to the winner-take-all nature of the DT-based subgroup and no particular noise-filtering mechanism, all the noise events are integrated over all pixels being illuminated during the entire measurement window, t_{meas}, therefore, making it practically impossible to detect signal events. Also, in a flash LiDAR which integrates background noise over a wide FOV (in this example, $40°$), an optical bandpass filter becomes ineffective at the assumed 50 klux ambient light condition.

Even in a case where the DT dead time, t_{comb}, is lower than the inverse of the background noise events, a very high bandwidth readout channel would be required to propagate both signal and noise events impinging on the sensor surface. This in fact limited the operating condition of this DT-based DTOF sensor, as seen through the results presented in this chapter through Section 3.3 which were mostly measured under controlled (and low) background light condition as already mentioned. Thus, it is paramount that the sensor architecture has embedded noise-filtering which adds to the optical filtering provided by bandpass filters, especially when high background noise as this example is inevitable in a LiDAR system.

3.5 Conclusions

This chapter discussed the concept of resource sharing necessary in DTOF sensor design. A comparison between event-driven and always-on shared TDC architecture was made depending on the power consumption and area. The analysis and simulations show that shared (and sample) approach offers better power efficiency for moderate to high photon activity at a slightly lower saturation. Furthermore, the always-on TDC concept results in a uniform and (almost) constant power consumption throughout the sensor, independent of the activity, removing the IR-drop

uncertainty typical of event-driven systems.

Secondly, a modular DTOF sensor, based on the proposed TDC sharing was described. A decision tree is used to manage multiple events, while propagating the first incoming event over every detection cycle. The module provides in-locus data processing and storage on a per-pixel basis. Each module is digitally synthesized and completely autonomous, which enables scaling to a desirable sensor size, without affecting its operation. The design was implemented in a TSMC 3D-stacking technology, featuring a BSI SPAD array on the top tier, connected to a readout and processing circuit on the bottom tier. The sensor measured ranges up to 300 m with accuracy error lower than 0.4 %. 3D images were obtained by a two-axis galvo scanning system, for up to 10-m range and 30° AFOV. All long range measurements in this chapter have been performed using a laser in the visible spectrum, at 532-nm wavelength, due to lab availability. Conversely, commercial LiDARs usually use non-visible lasers in the near-infrared spectrum, above 700 nm where, typically, CMOS detectors have lower sensitivity. However, the ideal operation wavelength depends on the system architecture and application. In order to limit the amount of integrated solar noise in the sensor as well as interaction with atmosphere, it is desirable to choose NIR-IR illumination wavelengths [18].

Finally, an analytical model is also presented identifying various challenges in the decision-tree based approach due to the absence of noise filtering. This is used as a premise to develop an alternative noise-resilient design, aimed at rejecting high background light. The next chapter proposes a new design for which an analytical model is described. The improvised architecture is based on coincidence detection and gating. Various simulations are presented where the benefits of migrating towards the newer architecture will be apparent.

REFERENCES

[1] A. Ronchini Ximenes, P. Padmanabhan, and E. Charbon, "Mutually coupled time-to-digital converters (tdcs) for direct time-of-flight (dtof) image sensors," *Sensors*, vol. 18, no. 10, p. 3413, 2018.

[2] A. R. Ximenes, P. Padmanabhan, M.-J. Lee, Y. Yamashita, D.-N. Yaung, and E. Charbon, "A modular, direct time-of-flight depth sensor in 45/65-nm 3-d-stacked cmos technology," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 3203–3214, 2019.

[3] P. Padmanabhan, C. Zhang, and E. Charbon, "Modeling and analysis of a direct time-of-flight sensor architecture for lidar applications," *Sensors*, vol. 19, no. 24, p. 5464, 2019.

[4] C. Bruschini, H. Homulle, I. M. Antolovic, S. Burri, and E. Charbon, "Single-photon avalanche diode imagers in biophotonics: review and outlook," *Light: Science & Applications*, vol. 8, no. 1, pp. 1–28, 2019.

[5] M.-J. Lee, A. R. Ximenes, P. Padmanabhan, T.-J. Wang, K.-C. Huang, Y. Yamashita, D.-N. Yaung, and E. Charbon, "High-performance back-illuminated three-dimensional stacked single-photon avalanche diode implemented in 45-nm cmos technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, pp. 1–9, 2018.

[6] C. Zhang, S. Lindner, I. M. Antolović, J. M. Pavia, M. Wolf, and E. Charbon, "A 30-frames/s, 252×144 spad flash lidar with 1728 dual-clock 48.8-ps tdcs, and pixel-wise integrated histogramming," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 1137–1151, 2018.

[7] R. K. Henderson, N. Johnston, S. W. Hutchings, I. Gyongy, T. Al Abbas, N. Dutton, M. Tyler, S. Chan, and J. Leach, "5.7 a 256 × 256 40nm/90nm cmos 3d-stacked 120db dynamic-range reconfigurable time-resolved spad imager," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, pp. 106–108, IEEE, 2019.

[8] G. F. Knoll, *Radiation detection and measurement*. John Wiley & Sons, 2010.

[9] S. H. Lee and R. P. Gardner, "A new g–m counter dead time model," *Applied Radiation and Isotopes*, vol. 53, no. 4-5, pp. 731–737, 2000.

[10] S. Henzler, *Time-to-digital converters*, vol. 29. Springer Science & Business Media, 2010.

[11] X. Gao, E. A. Klumperink, P. F. Geraedts, and B. Nauta, "Jitter analysis and a benchmarking figure-of-merit for phase-locked loops," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 56, no. 2, pp. 117–121, 2009.

[12] T. Fersch, R. Weigel, and A. Koelpin, "A cdma modulation technique for automotive time-of-flight lidar systems," *IEEE Sensors Journal*, vol. 17, no. 11, pp. 3507–3516, 2017.

[13] B. Buttgen and P. Seitz, "Robust optical time-of-flight range imaging based on smart pixel structures," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 6, pp. 1512–1525, 2008.

REFERENCES

[14] C. Niclass, M. Soga, H. Matsubara, S. Kato, and M. Kagami, "A 100-m Range 10-Frame/s 340×96-Pixel Time-of-Flight Depth Sensor in 0.18-μm CMOS," vol. 48, no. 2, pp. 559–572, 2013.

[15] M. Perenzoni, D. Perenzoni, and D. Stoppa, "A 64 × 64-pixels digital silicon photomultiplier direct tof sensor with 100-mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6 km for spacecraft navigation and landing," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 151–160, 2016.

[16] T. Al Abbas, N. A. Dutton, O. Almer, N. Finlayson, F. M. Della Rocca, and R. Henderson, "A cmos spad sensor with a multi-event folded flash time-to-digital converter for ultra-fast optical transient capture," *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3163–3173, 2018.

[17] F. Villa, R. Lussana, D. Bronzi, S. Tisa, A. Tosi, F. Zappa, A. Dalla Mora, D. Contini, D. Durini, S. Weyers, *et al.*, "Cmos imager with 1024 spads and tdcs for single-photon timing and 3-d time-of-flight," *IEEE journal of selected topics in quantum electronics*, vol. 20, no. 6, pp. 364–373, 2014.

[18] ISO-9845-1, "Solar energy—reflectance solar spectral irradiance at the ground at different receiving conditions—part 1: direct normal and hemispherical solar irradiance for air mass 1.5," 1992.

4 Coincidence-based noise-resilient DTOF sensor

High background noise from ambient light is a primary challenge of depth sensing within a LiDAR system. This challenge is exacerbated specifically in a flash LiDAR where the entire FOV is illuminated at once, resulting in relatively lower achievable signal-to-background-noise ratio (SBR) compared to a scanning system which operates over a much smaller FOV. However, irrespective of whether it is a scanning or a flash implementation, integrated noise suppression on chip becomes necessary to improve signal acquisition under high ambient light. This chapter proposes a shared DTOF sensor architecture to cope up with high-background noise. An analytical model is developed on MATLAB to define a suitable architecture to cope up with the challenge. Simulation results from this model set the target specifications for the IC design of the actual sensor. Modeling work and analysis presented in this chapter is based on the work published in [1].

4.1 Overview– coincidence detection

High background noise has been mitigated by implementing a well-known concept referred to as "coincidence detection" or alike [2, 3, 4, 5]. Coincidence detection is a technique which utilizes spatio-temporal correlation (and closeness) of photons within a laser pulse to filter out background noise photons which are uniformly distributed in time. Figure 4.1 conceptually explains this technique with an example scene and its corresponding 3D image, earlier presented in Section 3.3.4, Figure 3.16. The main idea is to exploit the fact that the signal photons reflected from the target are temporally correlated and thus, most likely to be concentrated within a time-window coarsely equal to the total system full width at half maximum, $FWHM \approx 2.355\sigma_{total}$. Here, σ_{total} is the contribution of individual sources of timing uncertainty ($\sigma_{total} = \sqrt{\sigma_{laser}^2 + \sigma_{SPAD}^2 + \sigma_{TDC}^2 + \sigma_{other}^2}$).

Instead of letting the sensor integrate events over the entire measurement window, t_{meas}, imposing a time constraint, referred to as the "coincidence window", reduces the likelihood of acquiring noise events whose probability of occurence within that narrow window is very low. Consequently, this results in electrically enhancing the SBR. Coincidence may be implemented at the sensor level over clusters/groups of closely-spaced pixels, exploiting a "more-likely" fact that neighboring pixels may belong to similar target depths (and thus, TOFs). This is also observed in the 3D reconstruction in

Figure 4.1 for all objects in the scene (see inset of the object labelled [4]).
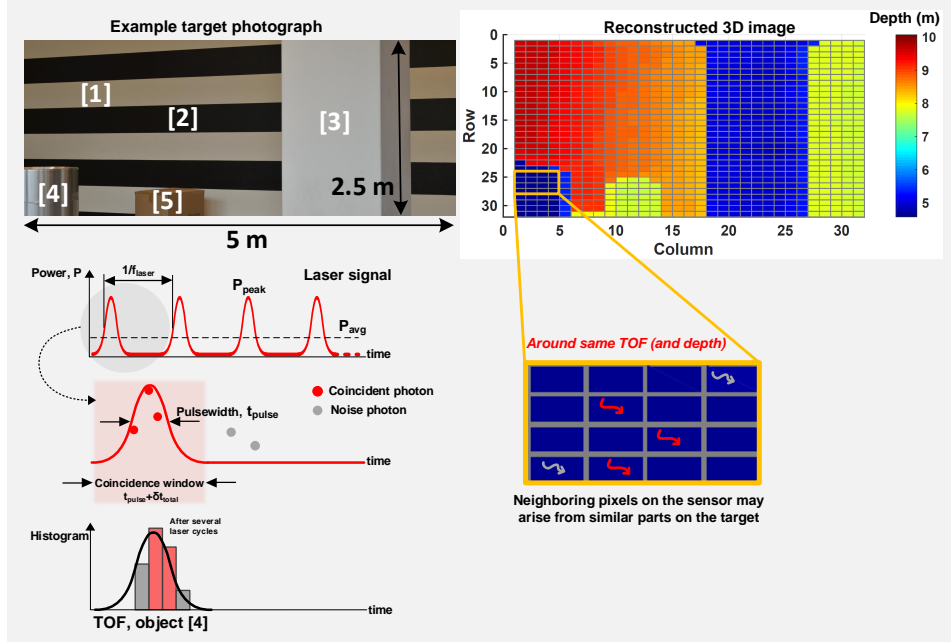
Figure 4.1 – Conceptual representation of coincidence detection.

4.2 Proposed DTOF sensor based on coincidence

The DTOF sensor implemented in the previous chapter operated predominantly in the negative SBR regime when the background light reaches high levels, evident both from the measurements and the analytical results. To circumvent these issues, a new architecture is proposed in this chapter, adapted to detect coincidence. The spatial arrangement of pixels and the subgroup are similar to the decision-tree-based (DT-based) architecture (Chapter 3), retaining the benefits of sharing, modularity as well as scaling, while the combination tree itself is significantly modified to enhance SBR. Apart from coping with high background noise, the sensor is designed to improve robustness in wide dynamic range scene where incoming photon flux can be vastly varying. The proposed sensor architecture is firstly analyzed using a MATLAB model, where performance is evaluated under wide range of simulation conditions, as seen in a typical LiDAR scenario. The analysis and simulation results from this chapter form the basis for the actual sensor implementation, to be discussed in the next chapter.

Figure 4.2a shows a high-level block diagram of the proposed DTOF sensor model. Figure 4.3 shows a timing diagram with the flow of operations by taking an example of three incoming events which are marked in red as 1, 2, 3, as shown in Figure 4.2a.

In the proposed model, the subgroup, sg (M pixels), is further clustered into N minigroups, mg, comprising of (M/N) number of pixels each. Unlike the approach in DT-based architecture (Figure 3.18), where the first incoming event is propagated while blocking/ignoring successive events,

the tree in the proposed architecture has a non-blocking nature. Arrival of the first event starts a tunable coincidence window, t_{window} (see event 1 in Figure 4.3). While a M-input combination tree propagates the first event (DTOF_sample in Figures 4.2a and 4.3) to the TDC, the successive events within the window are preserved and locally processed in the minigroups by the M/N-input combination trees without being ignored or completely lost. This results in up to N detectable events to be detected compared to the previous 1-event detection in the decision tree. The combination tree within the minigroups is referred to as the "coincID" tree. Every minigroup, mg, is modeled to provide coarse timestamps with a resolution, T_{coarse}, along with generating the binary ID ($log_2(N)$ bit ID) of the pixels contributing events. For N number of minigroups, the subgroup is able to provide data for up to N number of detections within the coincidence window.
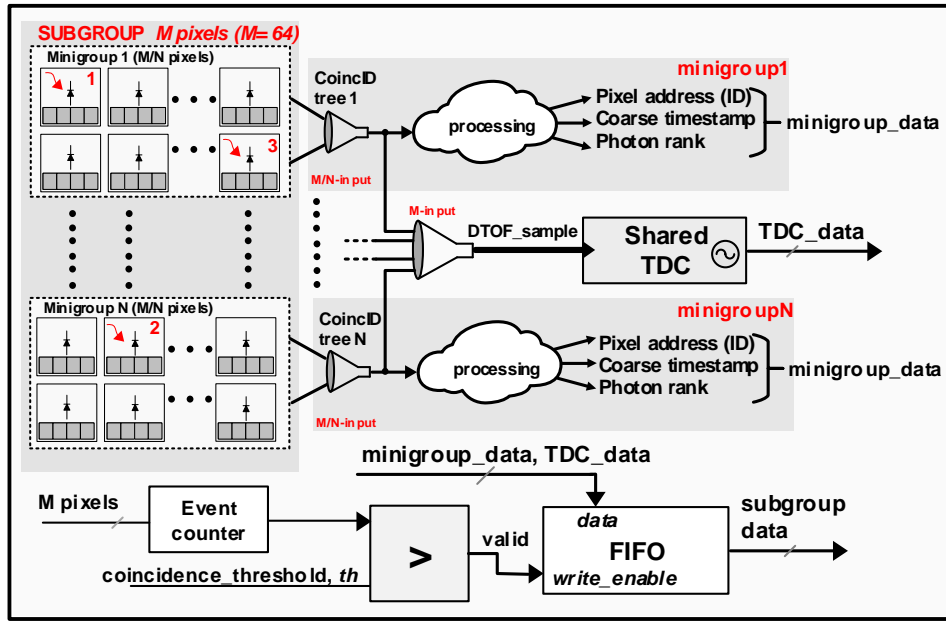
(a) Block diagram- Coincidence-based DTOF sensor
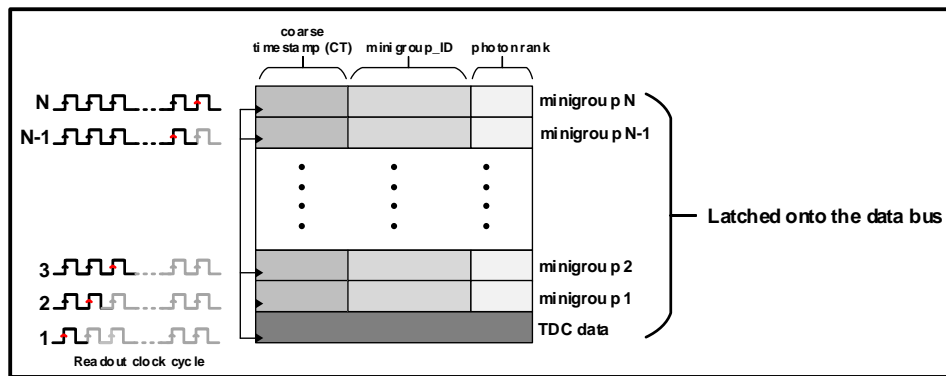
(b) Data description

Figure 4.2 – (**a**) A block diagram of a DTOF sensor adapted to detect coincidence and (**b**) description of data from subgroup, that is, minigroups and TDC data.

The data description from the minigroup is shown in Figure 4.2b. As shown in Figure 4.3, minigroup

1 generates data corresponding to event 1. This data consists of a coarse timestamp, CT-1 along with its address referred to as ID1 and the photon rank, which indicates the order of the photon in a coincidence window. Similarly, minigroup N provides data corresponding to event 2.

There is an event counter in every subgroup which tracks the number of photons within a coincidence window. A comparator logic is used to compare the output of the event counter with a predefined (and variable) coincidence threshold, th. As soon as the event counter output exceeds th, a signal is considered valid. In the example timing diagram in Figure 4.3, $th = 2$ and as the event counter output reaches 2, a valid signal is generated. The valid signal enables the data-writing process in the FIFO block following which the data from various minigroups and the TDC data is latched onto a data bus (see data description in Figure 4.2b).

Simplified timing diagram
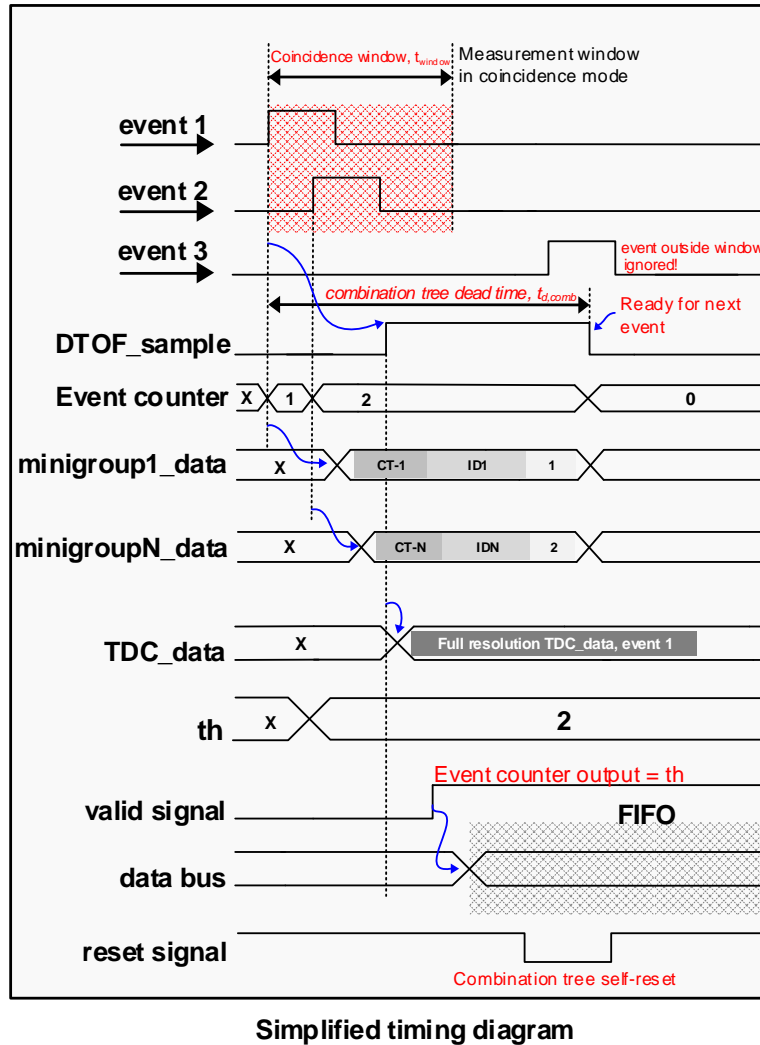
Figure 4.3 – Simplified timing diagram showing the operation of the proposed architecture.

Subjecting the validity of an incoming event to a comparator logic with a certain threshold thus allows to reject most of the background noise, whose likelihood of occurrence during the coincidence

window is low. The main purpose of implementing variable thresholds is to address the challenge in wide dynamic range targets where photons from lower reflective parts of the scene need to be captured in the presence of brighter targets. It will be shown in Section 4.3.2 that there is an evident relationship between the coincidence threshold and the photon activity rate where the benefit of configurable threshold is clearer.

Additionally, the filtering action through coincidence detection, along with multiple minigroups which enable timestamping of more than one event in a coincidence window also help mitigate pile-up distortion [6] by avoiding unnecessary sampling of the TDC when the threshold condition is not satisfied while also processing more than 1 photon through the minigroups in the presence of multiple events.

Usually, in TCSPC with coincidence detection, the combination logic combines multiple pixel events from multiple pixels on satisfying coincidence, however, often sacrificing granularity and achievable spatial resolution [2]. When a flash LiDAR is operating in a wide FOV, the sensor may see multiple different targets within a scene, unlike low-FOV systems. Under such scenarios, it becomes important to capture information from as many targets as possible within the FOV while also ensuring only minimal depth errors. The feature of multiple coarse timestamping in the minigroups facilitates this by allowing up to N detections within the subgroup by providing up to N *timestamps* and corresponding *ID* data of those N detections. This, consequently also enhances the timing throughput per subgroup by N times.

The proposed architecture described above is analytically modeled on MATLAB and the probability of noise and signal detections are calculated in the coincidence mode. Under high background noise conditions, noise events can potentially result in false coincidences as well, where a valid event may be a contribution of either all noise events or a combination of signal and noise events. This is in addition to the true coincidences where the contributing events are only due to signal photons. Thus, the probability of the true signal detection will depend on the probability of propagating the false coincidences as well.

For a pixel, i, the number of noise and signal events per pixel during a coincidence window, t_{window} with a coincidence threshold, th, is calculated as

$$N_{th,pixel}(i) = N_{pixel}(i) \cdot (t_{window} - th \cdot t_{d,spad}),$$
(4.1)

$$S_{th,pixel}(i) = S_{pixel}(i) \cdot ((1/f_{laser}) - th \cdot t_{d,spad}),$$
(4.2)

where, $N_{pixel}(i)$ and $S_{pixel}(i)$ are previously expressed in Section 3.4.1 in the previous chapter, through Equations 3.16, 3.17, 3.18 and 3.19.

Figure 4.4 shows a subgroup example with $M = 32$ pixels with $N = 4$ minigroups containing 8 pixels each, where, regions corresponding to specific probabilities are highlighted for the ease of understanding. The equations on the figure correspond to signal probabilities (Equations (4.6)

and (4.8)) while the pictorial representation itself is the same for analyzing both, noise and signal photons.

The probability of detecting a valid noise event (false coincidence) at pixel, i (marked as $i^{th} pixel$ in red in Figure 4.4a), with th number of coincident events in the subgroup, $sg(i)$, is calculated as the conditional probability of detecting a noise event at pixel, i, given that $(th-1)$ noise events are detected in rest of the subgroup (Figure 4.4a).

$$p\_n_{th}(i) = p\_n_{pixel}(i) \cdot p\_n_{th-1,sg}(i), \tag{4.3}$$

where $p\_n_{pixel}(i)$, the probability of detecting 1 event, is calculated as

$$p\_n_{pixel}(i) = N_{th,pixel}(i) \cdot \exp(-N_{th,pixel}(i)). \tag{4.4}$$

$$p\_s_{th}(i) = (1 - p\_n_{pixel}(i)) \times p\_s_{pixel}(i) \times p\_s_{th-1,sg}(i)$$

$$p\_s_{th-1,sg}(i) = p\_s_{th-1,mg} \cup p\_s_{th-1,sg-mg}$$
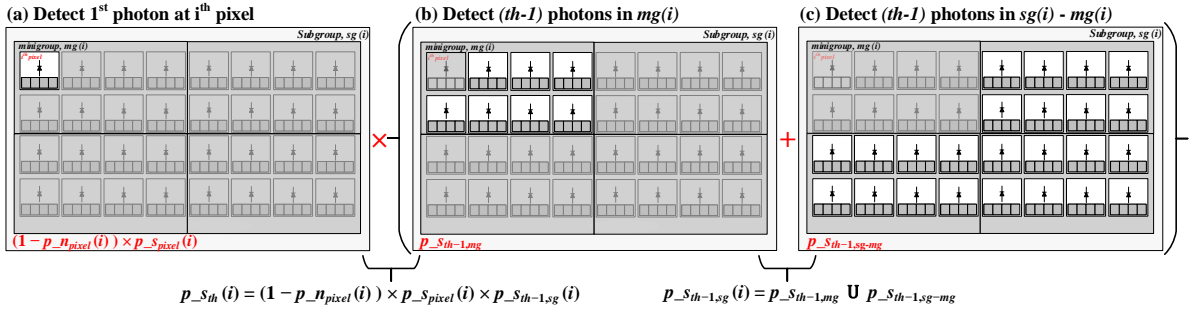
Figure 4.4 – Subgroup, $sg(i)$, demarcated to show various probabilities under coincidence mode to detect (th) number of signal photons.

The probability of detecting $(th-1)$ noise photons in the remainder of the subgroup is calculated from the union operation of individual probabilities of detecting $(th-1)$ noise photons in the minigroup, $mg(i)$ (see Figure 4.4b) and the remainder of the subgroup, $sg(i) - mg(i)$, (see Figure 4.4c).

$$p\_n_{th-1,sg}(i) = p\_n_{th-1,mg} \cup p\_n_{th-1,sg-mg}. \tag{4.5}$$

The effective number of photons in the minigroup, $mg(i)$ and in rest of the subgroup, $sg(i) - mg(i)$, excluding the minigroup to which pixel i belongs is calculated from the pixel-wise photon number calculated in Equation (4.1) and used to calculate $p\_n_{th-1,mg}$ and $p\_n_{th-1,sg-mg}$.

The probability of detecting valid signal events within t_{window} can be calculated as a conditional probability of detecting a signal event in a pixel, i, given that no noise photon is detected at i and $(th-1)$ signal events are detected in the rest of the subgroup (see Figure 4.4 showing combined probability).

$$p\_s_{th}(i) = (1 - p\_n_{pixel}(i)) \cdot p\_s_{pixel}(i) \cdot p\_s_{th-1,sg}(i), \tag{4.6}$$

where $p\_s_{pixel}(i)$, the probability of detecting 1 signal event, is calculated as

$$p\_s_{pixel}(i) = S_{th,pixel}(i) \cdot \exp(-S_{th,pixel}(i)). \tag{4.7}$$

Likewise, the probability of detecting $(th-1)$ signal photons in rest of the subgroup is calculated from the union operation of individual probabilities of detecting $(th-1)$ signal photons in the minigroup, $mg(i)$ and the rest of the subgroup, $sg(i) - mg(i)$ (see grouping in Figure 4.4),

$$p\_s_{th-1,sg}(i) = p\_s_{th-1,mg} \cup p\_s_{th-1,sg-mg}. \tag{4.8}$$

Equation (4.2) is likewise used to compute effective number of photons in the minigroup, mg_i and in rest of the subgroup, $sg(i) - mg(i)$ as mentioned for noise photons.

The final conditional probabilities in Equations (4.3) and (4.6) for noise (p_n_{th}) and signal (p_s_{th}) events respectively, are used to obtain simulation results described in the next section.

4.3 Simulation results

The proposed model was simulated within a flash LiDAR scenario already introduced in the previous chapter, redrawn here again for ease of comprehension. Simulation parameters remain the same as before; described through Section 3.4.1 and summarized in Table 3.2 of Chapter 3.
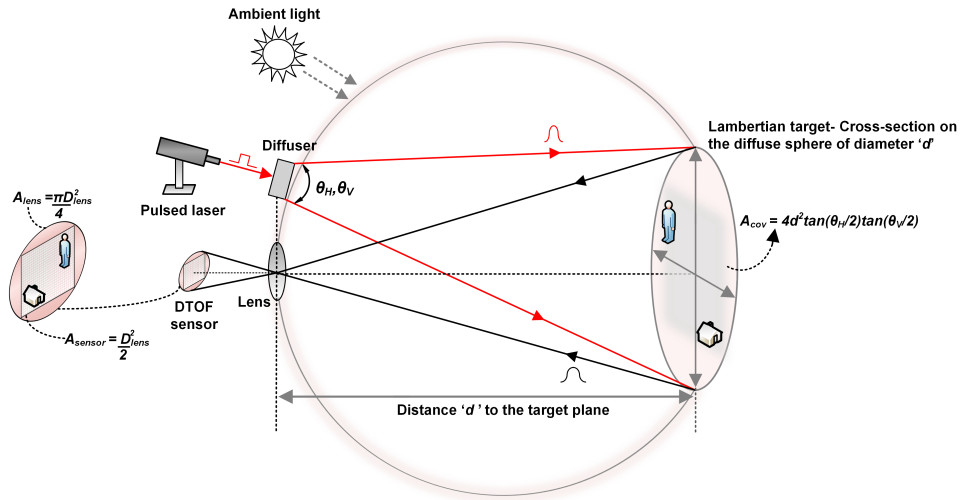
Figure 4.5 – Flash LiDAR operation.

4.3.1 Single-point ranging

Single-point ranging was simulated for a uniform flat target over increasing distances, d and the relative probabilities of noise-based coincidence and signal-based coincidence were calculated from the established equations. In these simulations, the subgroup, sg, consists of 64 pixels (8×8) and

the minigroup, mg consists of 4 pixels (2 × 2).

Figures 4.6a,b, show the simulated results on a log and linear scale respectively under a 50 klux background light condition.
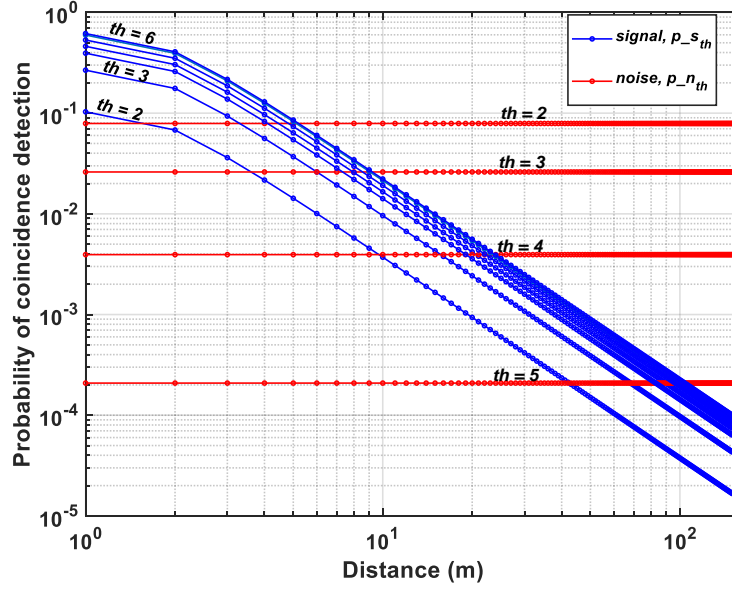
(a) Vertical axis log scale.
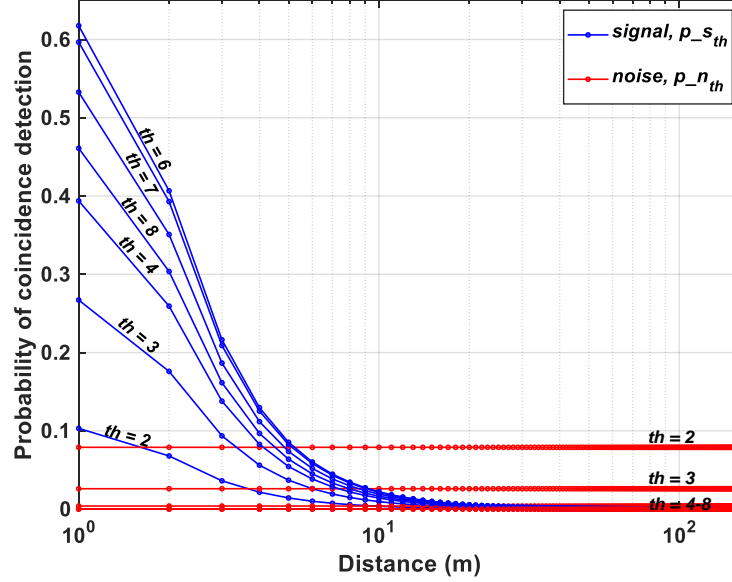
(b) Vertical axis linear scale.

Figure 4.6 – Probability of signal and noise detection at different coincidence thresholds- (**a**) log scale and (**b**) linear scale.

The probabilities are plotted for coincidence thresholds, $th = 2$ to $th = 8$ with a $t_{window} \approx 1$ ns.

A direct comparison with Figure 3.21 shows that coincidence evidently increases the probability of signal detection, particularly seen for $d = (1 - 10)$ m (clearer visibility on linear scale in Figure 4.6b). While this is true, it can also be seen that for $d > 10$ m, there is significant reduction in signal probability as well. The achieved SBR after detection, SBR_{det} (dB) is plotted for the proposed architecture without and with coincidence, for $th = 4$ in Figure 4.7. As seen earlier, the sensor operates in a negative SBR regime throughout the unambiguous range under no-coincidence. In coincidence mode, the results improve for shorter distances ($d < 11m$) with positive SBR, however, for longer ranges, d, the SBR enters the negative regime where the returning signal photons are also fewer (inverse square law) (see marking in Figure 4.7). Consequently, imposing a timing constraint through coincidence, t_{window}, does not provide any additional improvement in the SBR.
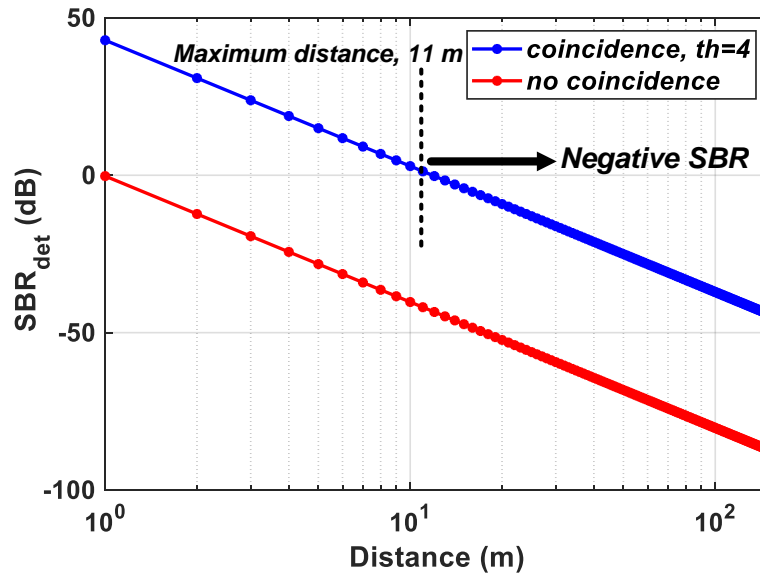
Figure 4.7 – SBR achieved after detection without and with coincidence with $th = 4$.

This points to two main corollaries— (1) as concluded in Section 3.4.1, noise-filtering and thus, SBR improvement provided by optical means such as, bandpass filters, is limited when background illumination is as high as 50-100 klux and (2) noise-filtering implemented at the sensor-level also has its limitations. Therefore, the components within a LiDAR system should not be understood as mutually-exclusive in operation and one component alone may not yield desirable results. In other words, for a robust operation of a LiDAR, the whole system should be considered as a closed-loop system where there is continuous feedback between the sensor and the rest of the components—illuminator and the optics. Intensity of incoming photon activity can be monitored and this data can be used to control the illumination power of the outgoing laser pulse and/or control the FOV through optics, depending on the operating distance to yield an optimal SBR. This closed-loop relationship is indicated in red dotted lines in Figure 4.8. In general, for longer distances (see Figure 4.6), the fewer number of returning photons can be overcome by a operating over a number of parameters including but not limited to,

- increasing the outgoing laser power without violating eye-safety regulations,
- decreasing the FOV, given that at longer distances, a full resolution sensing may not be

required,

- at the sensor level, data from multiple pixels may be combined at the expense of lower spatial resolution,
- time-gating may be another alternative to achieve target-dependent ranging at the sensor level (will be discussed in Section 4.3.4).
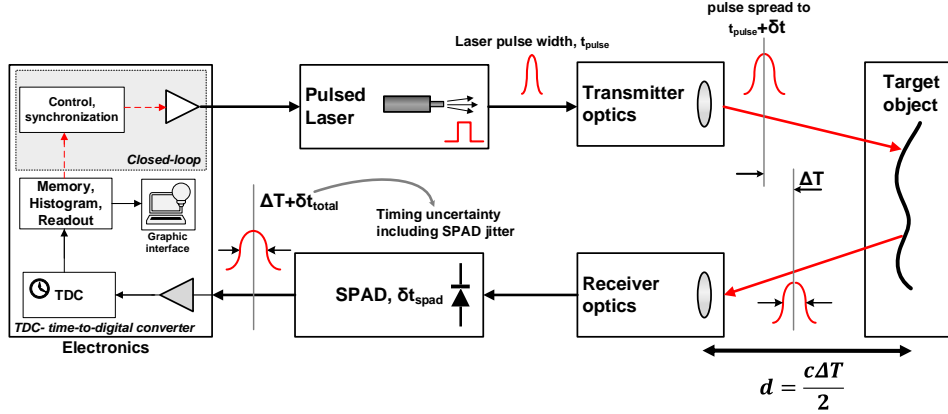
Figure 4.8 – DTOF system block diagram- red arrows indicate the feedback between the sensor and the illumination control.

Figure 4.9 shows the simulation of alternative (2) mentioned above to detect signal from a flat target at 150 m, which is the maximum unambiguous range at f_{laser} = 1 MHz (unless multiple frequencies are used).
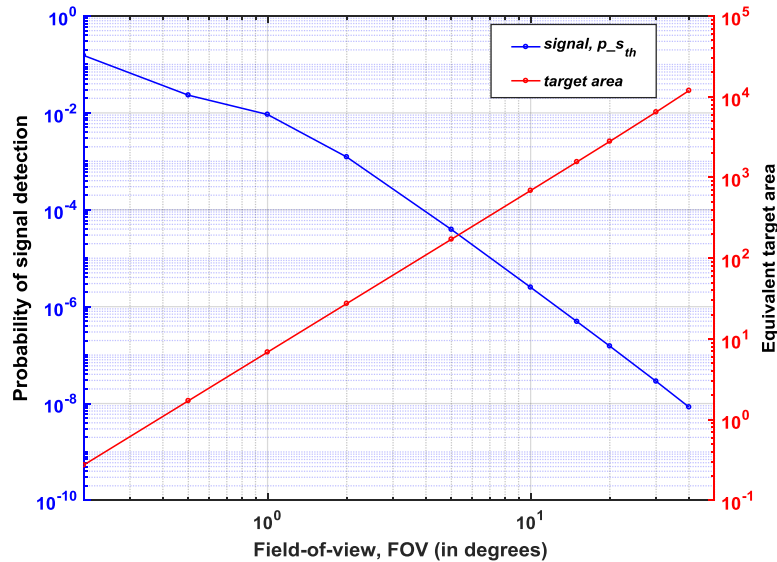
Figure 4.9 – Probability of detection at $d = 150$ m (left vertical axis) and equivalent target area (right vertical axis) at varying FOV (horizontal axis).

The coincidence threshold is set at $th = 2$. The probability of signal detection is plotted with varying FOVs. It can be seen clearly that the success of detection improves with decreasing FOVs as

expected, due to fewer number noise photons being integrated and laser energy concentrated over a small FOV together yields a higher SBR. In this example, peak detection is achieved at 0.2° FOV.

In a practical scenario, at distances of about 50–150 m, most often it may not be required to get high resolution 3D images at further distance and it may suffice to get the range estimation alone and this can be achieved with a very narrow FOV. Figure 4.9 also shows the evident improvement in SBR at a lower FOV and thus, lower integrated background noise. This naturally improves the signal detection probability. Certainly, as mentioned before, at 150 m, operating with a FOV of 40°, covers an area of approximately 10,000 m^2, which is totally an impractical scenario. Therefore, for longer distances, the illumination unit may be used as a point source, using all the energy of the laser pulse to provide range estimation alone while for shorter distances, the laser energy may be distributed into a wider FOV with an array of points to provide 3D depth map.

4.3.2 3D imaging with wide dynamic range targets

This section analyzes a typical flash LiDAR scenario with multiple objects within a scene covering a wide FOV of 40° at 0–10 m. Figure 4.10a shows a scene where targets range between 8–60% reflectivites, an apt example to examine a wide dynamic range scene. Alongside the example scene is a 32 × 32 3D image measured using the DT-based chip in a scanning LiDAR setup under a low-noise environment. This measured image (Figure 4.10b) is used as an input target scene and fed to the analytical models of the DT-based DTOF sensor scheme described in Section 3.4.1 of the previous chapter and the proposed coincidence-based DTOF scheme presented in 4.1.

Different target reflectivities
[1] – White wall – 50%
[2] – Black wall – 8%
[3] – White pillar – 60%
[4] – Aluminum bin – 54%
[5] – Cardboard box – 21%

(a)

(b)
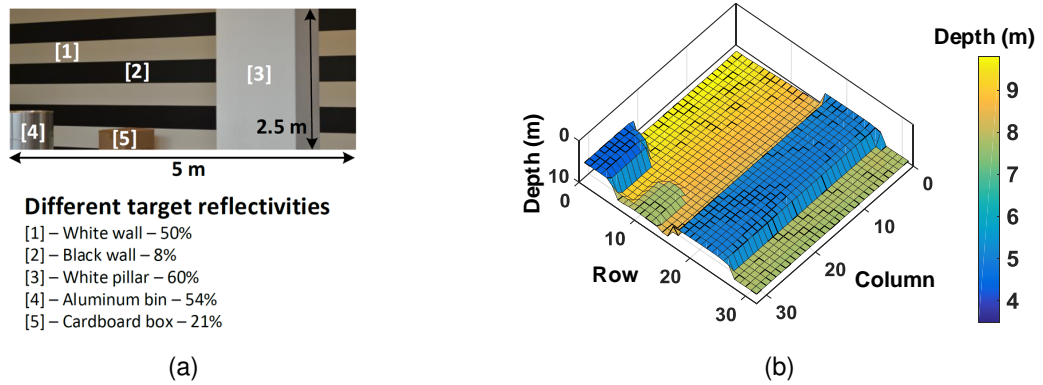
Figure 4.10 – (**a**) Photograph of the example target scene and (**b**) 32×32 image reconstructed in a scanning LiDAR setup [7].

A DTOF sensor with 1024 pixels (32 × 32) is assumed to map the given target scene. The subgroup, sg, consists of 64 pixels (8 × 8) as mentioned previously. The target reflectivities are as indicated in Figure 4.10a. The probabilities of detection are calculated based on the established equations seen through previous sections and using these, the histogram is computed for every pixel. The TOF for every pixel is calculated as the time bin with maximum number of counts in the computed histogram which is then used to reconstruct the combined 32 × 32 depth image. It is assumed that every module (2 × subgroups), in the 32 × 32 sensor array is read out through 8 independent channels

(as number of modules is $N_{sg}/2 = 8$) whenever there is a valid event. The readout clock, clk_{read} is assumed to be around typical digital I/O frequency of 100 MHz for all the simulations herein. The timing throughput per subgroup per second, t_counts_{sg}, is then calculated as follows,

$$t\_counts_{sg} = \frac{N_{sg}}{2} \cdot clk_{read} \cdot N_{mg},\tag{4.9}$$

where N_{mg} is the number of minigroups.

The histogram statistics is acquired for t_counts_{sg} number of events for signal and noise. The total number of valid events for noise and signal is referred to as N_{total} and S_{total} respectively. The noise events, N_{total}, are modeled as a uniform distribution over the unambiguous range, $1/f_{laser}$ with a bin resolution, t_{res} defined by the total system jitter, σ_{total}, ($\sigma_{total} = \sqrt{\sigma_{laser}^2 + \sigma_{SPAD}^2 + \sigma_{TDC}^2 + \sigma_{other}^2}$) and number of bins, $b = 1/f_{laser}/t_{res}$. The signal events, S_{total} are modeled considering the Gaussian nature of the laser pulse and are calculated from the error function of a Gaussian distribution.

The winner-take-all, DT-based approach, described in the previous chapter is first simulated. The measured image (Figure 4.10b) is fed to the model under a 5 klux background noise condition and in a flash LiDAR setup with θ_H and θ_V are 30° and 15° respectively. Please note that a 50 klux background noise condition on this simulation showed only noise in the reconstructed 3D image and therefore, a 5 klux noise condition was used to simulate this part alone, to distinguish between noise and signal events and understand the limitations of this scheme. The system timing uncertainty assumed is $FWHM_{total} \approx 530$ ps. The laser parameters remain the same as in previous simulations. Figure 4.11a shows the obtained 3D image.
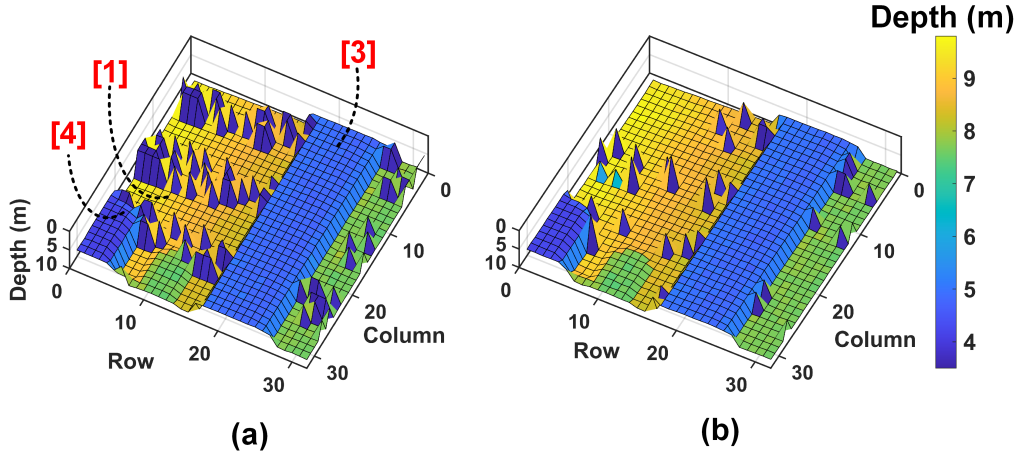
Figure 4.11 – (a) Simulated result of the 3D image reconstructed through DT-based scheme and (b) coincidence-based proposed architecture.

Since the DT-based approach described in Section 3.4.1 does not include any noise filtering mechanism, the expected degradation in the reconstructed (simulated) image in Figure 4.11a is evident compared to the measured image in Figure 4.10b, obtained under minimal noise in a scanning setup.

There is another important observation from this simulation result. It can be seen that the white wall (object (1)), the aluminium bin (object (4)) and the white pillar (object (3)) are readily reconstructed in comparison to the rest of the scene where there is up to 11.9 % incorrect sampling. This is also expected of the DT-based architecture which has the tendency to propagate events from higher reflective parts (and thus, more number of average events per second) compared to lower reflective parts of the scene.

The coincidence-based DTOF model in Section 4.1 is proposed to mitigate the limitations in a first-in-win-all scheme. The measured image (Figure 4.10b) is then fed as a target to the proposed model under 5 klux background noise first. The subgroup, sg, consists of 64 pixels (8 × 8) as mentioned previously and the minigroup, mg (absent in the first-in-win-all scheme) consists of 4 pixels (2 × 2). The reconstructed 3D image from the proposed scheme is shown alongside in Figure 4.11b for direct comparison with the first-in-win-all (no-coincidence) architecture. As can be seen, coincidence significantly improves the depth measurement with much fewer incorrect samplings.

Following this, the measured image (Figure 4.10b) is then fed as a target to the proposed model, now under the 50 klux background noise condition and with coincidence. The 3D image results from the simulations are shown in Figure 4.12. Interestingly, an apparent threshold-selective pattern in the reconstructed 3D image can be observed. In Figure 4.12a, with $th = 5$, it can be seen that objects (1), (3) and (4) are more accurately reconstructed compared to rest of the scene owing to their higher reflectivities, whereas, in Figure 4.12b, with $th = 2$, objects (2) and (5) take preference over the rest of the scene. This points to an important relationship between coincidence threshold, th and the target reflectivities.
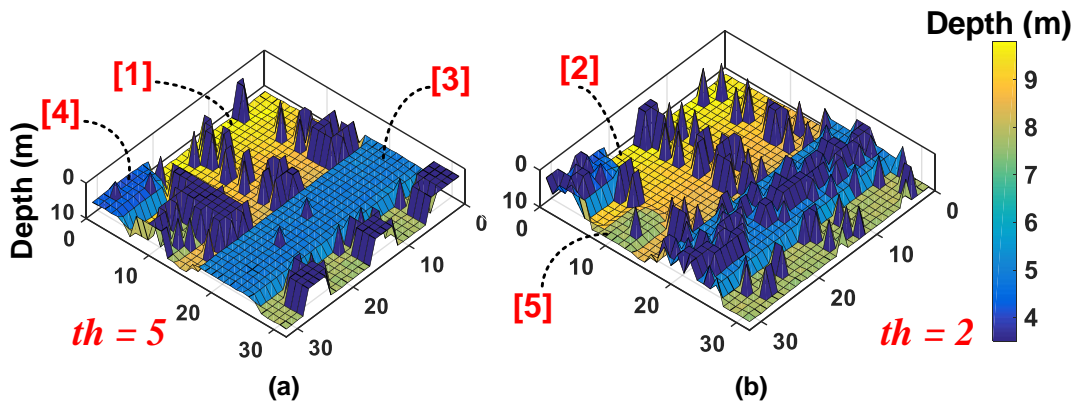
Figure 4.12 – Simulated result of the 3D image reconstructed through proposed DTOF model (a) coincidence threshold, $th = 5$ and (b) coincidence threshold, $th = 2$.

Figure 4.13 actually shows how coincidence threshold, th, increases for increasing photon activity, R, to provide a successful detection (temporal error, $\sigma_{total} < 230$ ps). This implies that a single threshold cannot yield an accurate reconstruction of a scene and in fact, a higher (lower) reflective target will imply higher (lower) photon activity rate requiring higher (lower) coincidence thresholds. On revisiting Figure 4.6b (redrawn for clarity in Figure 4.14), this relationship is already evident there. For targets at distances, particularly at $d = 1, 2, 3$ m, where the number of returning photons

is higher, the probability of signal detection increases for increasing thresholds, $p(d = 1, th = 6) > p(d = 1, th = 2)$ (see Figure 4.14). However, the probability ceases to rise for thresholds where the number of photons within the coincidence window is lesser than the threshold itself (see $p(d = 1, th = 7), p(d = 1, th = 8)$), given the fact that coincidence detection is performed utilizing photons within the laser pulse. The threshold, th, marked in Figure 4.2 can be chosen based on the activity rate of the subgroup to yield maximum success in signal detection. The modularity of the subgroups is used to configure them with unique thresholds depending on scene being imaged.
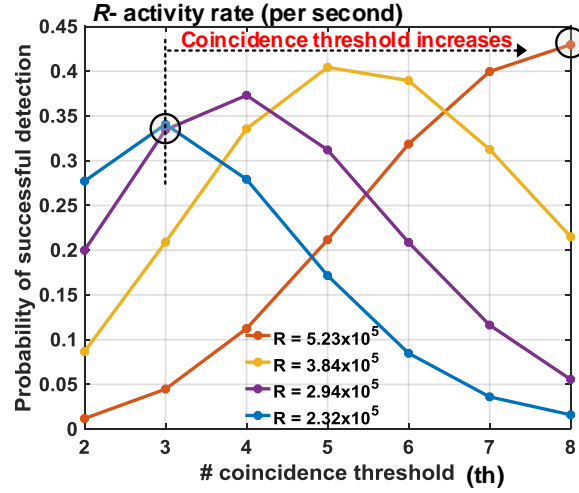
Figure 4.13 – The relationship between coincidence threshold and the incoming activity rate, R, received per second.

In fact, the 3D image earlier compared in Figure 4.11b was simulated with target-specific coincidence threshold in every subgroup over the sensor array.
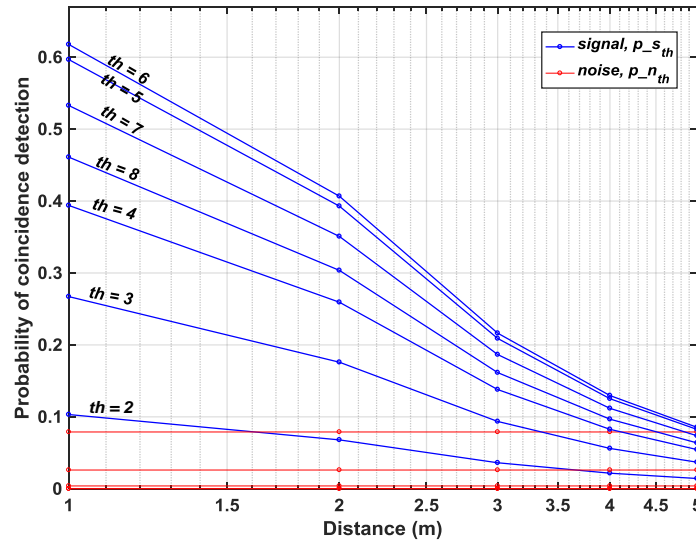
Figure 4.14 – Probability of signal detection in Figure 4.6b selected to show threshold dependence.

4.3.3   3D imaging and multiple timestamping

This section presents the multiple timestamping feature enabled through coarse time-taggers in the minigroups. Seen in Figure 4.15 are the 3D image reconstructions of the flash scenario earlier described. The grouping scheme is simulated for various subgroup and minigroup sizes where every subgroup has a unique coincidence threshold, th, chosen based on the incoming photon activity rate. For instance, the subgroup mapping the white pillar (object [3] in Figure 4.10a) is set to to $th = 5$ and cardboard box (object [5]), $th = 2$).

Subgroup- 16x16, Minigroup- 4x4   **Subgroup- 16x8, Minigroup- 4x4**   **Subgroup- 8x8, Minigroup- 2x2**

27.92% incorrect sampling   **14.64% incorrect sampling**   **7.03% incorrect sampling**
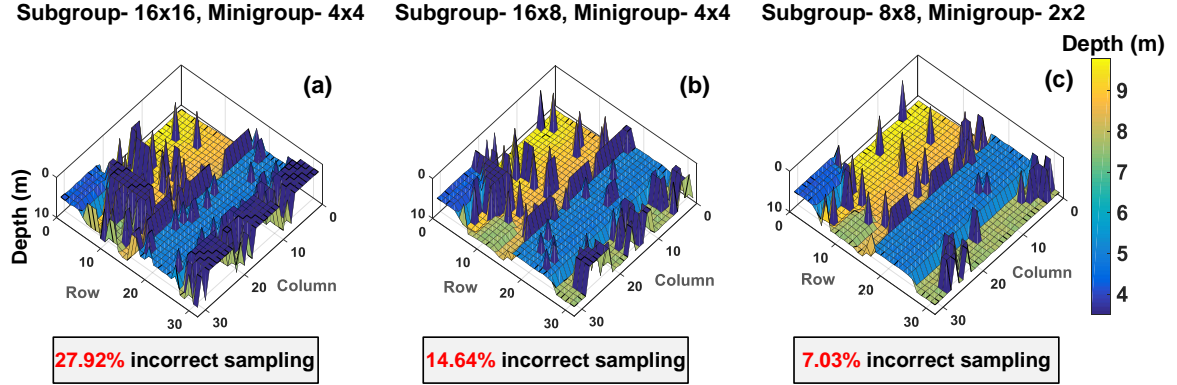
Figure 4.15 – The proposed grouping scheme illustrated for different subgroup and minigroup sizes.

It can be seen through Figure 4.15a–c that a coincidence within a subgroup of 8 × 8 pixels and minigroup of 2 × 2 pixels in Figure 4.15c allows improved 3D reconstruction with only about 7% incorrect sampling. The 16 minigroups containing 4 pixels (2 × 2) each enable up to 16 simultaneous TOF measurement within the coincidence window, t_{window}. Another objective of enabling multiple timestamping through the minigroups is to reduce the timing uncertainty, which could otherwise be limited to t_{window}. This is understood better by looking at a particular region of interest (ROI) in the example target introduced; Figure 4.16 highlights the ROI.
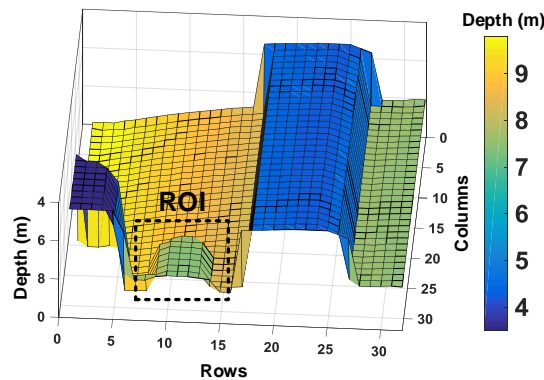
Figure 4.16 – 3D image input highlighted with the region of interest (ROI).

Absence of timing information on the contributing events within t_{window} results in an averaging effect in the reconstructed image, with a timing uncertainty $\approx t_{window}$. Although, this may not be of concern for very narrow pulses (with small FWHM), however when pulses become wide (on the

order of 2–5 ns, typical of commercial LiDARs), this uncertainty can result in depth error estimates. Figure 4.17a shows this effect where there is no multiple event timestamping. As can be seen, the reconstructed 3D image has averaged out with respect to the actual target input (see ROI in Figure 4.16) with an uncertainty, $\sigma = 1.93$ ns.
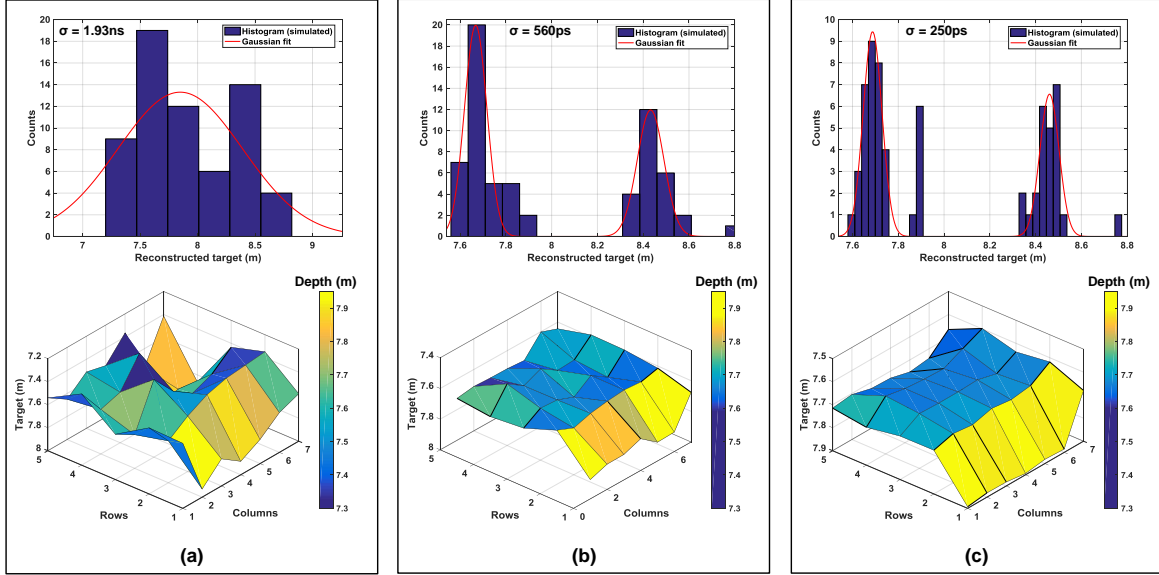
Figure 4.17 – Minigroup timestamping feature (**a**) no timestamping, uncertainty $\approx t_{window}$, (**b**) minigroup timestamping with a resolution, $T_{coarse} \approx 500$ ps and (**c**) minigroup timestamping with a resolution, $T_{coarse} \approx 200$ ps.

The corresponding histogram depicts the obtained Gaussian fit peaking around the mean value of 7.84 m, deviating from the actual input. Figures 4.17b,c show simulations with the proposed minigroup timestamping with a resolution of $T_{coarse} = 500$ ps and $T_{coarse} = 200$ ps respectively. The histograms in these cases show two distinct peaks with a bin resolution given by T_{coarse}. These peaks are then spatially correlated to the pixels in the minigroup from the IDs generated in each minigroup (as described in Section 4.1). The improvement in timing uncertainty is also seen in the corresponding 3D images below the histograms. In principle, the number of multiple peaks being detected is equal to the number of minigroups. In this simulation study, it is therefore, possible to detect up to 16 distinct peaks (number of minigroups = 16). Figure 4.18 shows another histogram simulation result on a different ROI, as marked in Figure 4.18a where up to 4 distinguishable peaks are detected concerning 4 distinct parts in the ROI.

4.3.4 3D imaging and time-gating

Time-gating is a useful technique to achieve range-selective depth sensing as well as 3D imaging under high background noise. A practical example may include a scene with a retro-reflective object in front of the actual desired target and it may become difficult for the sensor to estimate and reconstruct the actual target situated behind this retro-reflective object due to significant differences in their returning photon rates. Another scenario where time-gating may be effective is in an adverse

(a) ROI highlighted (different 3D perspective).

(b) Obtained histogram, $T_{coarse} = 500$ ps.
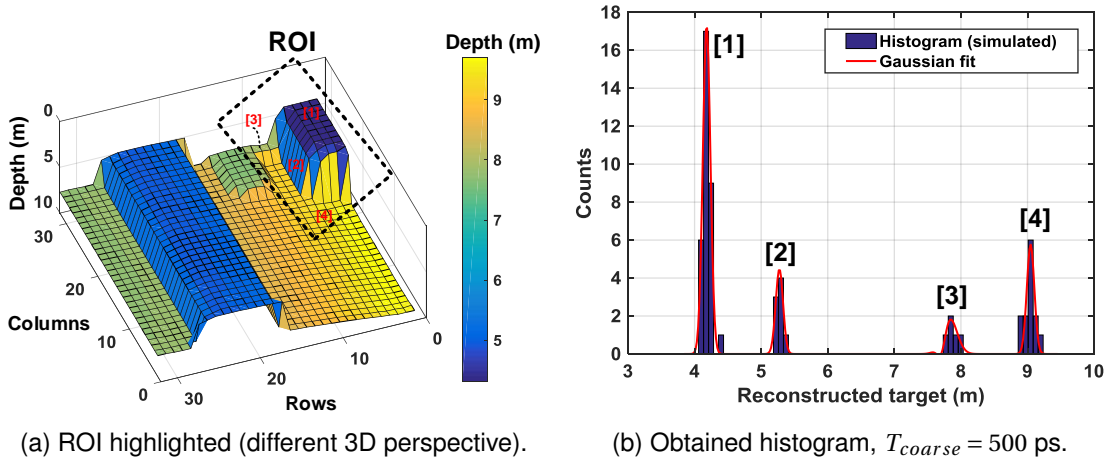
Figure 4.18 – Minigroup timestamping feature for up to four peaks.

weather condition due to the presence of scattering medium (such as fog, cloud). Particles in fog/cloud cannot be treated like background noise due to their non-uniform temporal distribution [8, 9] which may reflect as distinct peaks in the acquired histogram. Gated imaging can be useful to selectively eliminate unwanted peaks in the light propagation path. The proposed DTOF sensor is modeled to optionally operate in a time-gating mode to address aforementioned scenarios. Every subgroup, sg, is modeled to have unique gating windows. A different example target scene, shown in Figure 4.19a is used as an input to analyze this condition and fed to the proposed DTOF model, now operating under time-gating mode. The number of pixels in the subgroup is assumed to be 8×8 with 16 minigroups, each consisting 4 pixels (2 × 2).

(a)

(b)
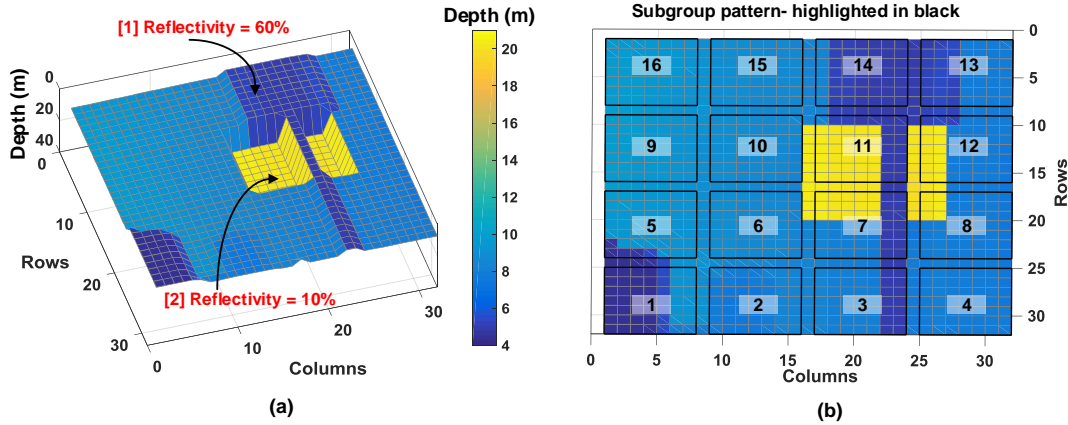
Figure 4.19 – (**a**) Input 3D image used to evaluate time-gating and (**b**) subgroups highlighted with the target scene.

Labelled [1] is an object with 60% reflectivity (retro-reflective equivalent) at about 5 m and the desired target, labelled [2], situated at 20 m with 10% reflectivity. This target is reconstructed by the proposed DTOF model with 1024 pixels (32 × 32); the parts of this target as seen by different subgroups is highlighted in Figure 4.19b for ease of visualization.

As numbered above, subgroup, sg_{11} is focused for this analysis. It covers a vast portion of the actual desired target (42 pixels) and a small fraction of the retro-reflective equivalent (17 pixels). The proposed DTOF model is simulated without and with gating. The histogram is then calculated for pixels in and around sg_{11} (rows-9:16, columns- 17:25) and the simulation results are shown in Figure 4.20.

The average photon rate from object (1) is much higher than object (2) and therefore, the subgroup combination tree will propagate more events from object (1) compared to object (2) for a given measurement time. Consequently, it can potentially mask the reflected photons from the desired target which reflects fewer photons (as object (2) in this example with lower reflectivity).
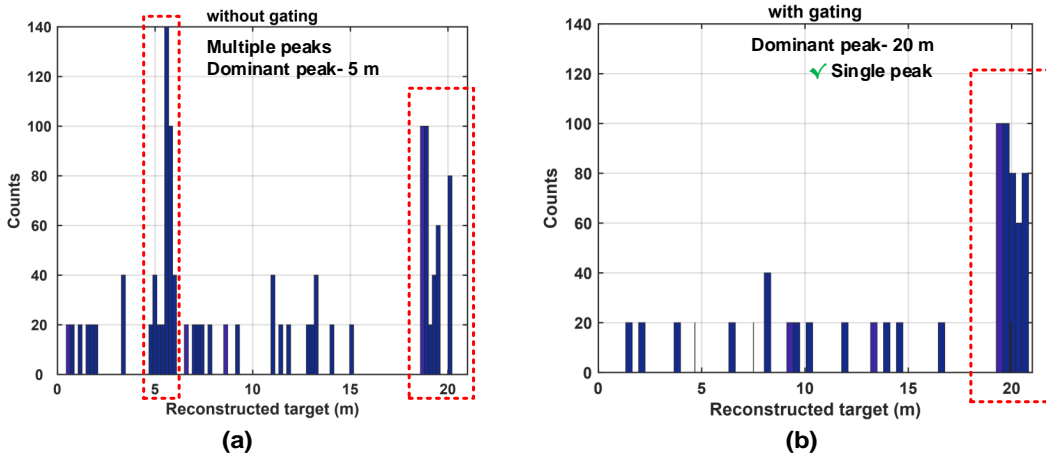
Figure 4.20 – Histogram of sg_{11} (**a**) without gating and (**b**) with gating.

Absence of gating results in a histogram with two peaks, with a dominant one at 5 m as seen in Figure 4.20a and another peak at 20 m. In fact, if a desired target was situated at a much longer distance, it may not even be possible to recover photons back from this target. Consequently, the presence of a dominant peak from the retro-reflective equivalent in the acquired histogram may mask the presence of the farther target. While the dominant peak from the retro-reflective target is not wrong, the low probability of recovering the farther target peak is of important concern. Gating will thus, allow us to select and propagate only "desired timestamps" within a gating window, in that way we can mitigate situations as in this example case. In Figure 4.20b, the subgroup operates in the gating mode with a gating window of 10 ns around the desired target situated at $TOF \approx 133$ ns (20 m) and the resulting histogram generates only one peak, around the desired target. In general, the sensor can have a moving gate with tunable window lengths to scan through the entire maximum ambiguous range ($\approx c/(2f_{laser})$) to gather an estimate on the desired target and choose the gating window accordingly.

4.4 Conclusions

Given that depth sensing involves a number of interdependent challenges, it is imperative to visualize the entire LiDAR system as a closed-loop system wherein, there is a continuous feedback from the

sensor to the illumination control logic and/or the optics. Therefore, an alternative (to DT-based approach) DTOF sensor model was proposed in this chapter, with an architecture suitable for such an implementation. An activity-based coincidence detection is proposed where the sensor is modeled to estimate the returning photon activity from different parts of the scene. Information as this can be used to select an appropriate coincidence threshold as seen in Section 4.3.2, to improve 3D imaging in wide dynamic range scenarios. Passive imaging in the sensor can be used to accumulate incoming photon activity on the sensor for a desirable observation time-window. The output after accumulation can be fed back to a control unit to configure subgroups with appropriate coincidence thresholds. While the feedback and configuration process itself can be very fast in an integrated implementation of the sensor, the major latency will be dictated by the observation window used to accumulate the incoming photon activity.

The feature of subgroup clustering into multiple minigroups enabled with independent timestamping increases the overall timing throughput (up to number of minigroups). This has a direct implication on the histogram processing as well, especially in long distance ranging, by providing multiple depth estimation simultaneously. Additionally, this feature also decreases the timing uncertainty within a detection window (coincidence or gating), the absence of which can otherwise lead to large depth errors, particularly when windows become longer than $\approx$ 2–10 ns ($\approx$ 0.3–1.35 m position error). Multi-pixel timing, counting and ID information as proposed preserves multiple pixel-data even in a shared architecture, which otherwise is typically lost due to the way they are combined [2, 7]. Furthermore, the information from multiple pixel events can be used in efficient photon-by-photon processing directly at the hardware-level instead of streaming out the entire raw data for processing off the chip [10].

The analytical model and simulations in this chapter provide an early validation and aid in the conceptual understanding of the sensor architecture. The next step involves transforming this analysis into a SPICE-compatible model to support actual design of the sensor in an integrated circuit. The next chapter presents the IC design of a (second) coincidence-based DTOF sensor along with its characterization.

REFERENCES

[1] P. Padmanabhan, C. Zhang, and E. Charbon, "Modeling and analysis of a direct time-of-flight sensor architecture for lidar applications," *Sensors*, vol. 19, no. 24, p. 5464, 2019.

[2] C. Niclass, M. Soga, H. Matsubara, M. Ogawa, and M. Kagami, "A 0.18-μ m cmos soc for a 100-m-range 10-frame/s 200×96-pixel time-of-flight depth sensor," *IEEE Journal of solid-state circuits*, vol. 49, no. 1, pp. 315–330, 2013.

[3] D. Portaluppi, E. Conca, and F. Villa, "32× 32 cmos spad imager for gated imaging, photon timing, and photon coincidence," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 2, pp. 1–6, 2017.

[4] M. Beer, J. Haase, J. Ruskowski, and R. Kokozinski, "Background light rejection in spad-based lidar sensors by adaptive photon coincidence detection," *Sensors*, vol. 18, no. 12, p. 4338, 2018.

[5] M. Perenzoni, D. Perenzoni, and D. Stoppa, "A 64 × 64-pixels digital silicon photomultiplier direct tof sensor with 100-mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6 km for spacecraft navigation and landing," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 151–160, 2016.

[6] W. Becker, *Advanced time-correlated single photon counting techniques*, vol. 81. Springer Science & Business Media, 2005.

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

[8] G. Satat, M. Tancik, and R. Raskar, "Towards photography through realistic fog," in *2018 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–10, IEEE, 2018.

[9] T. G. Phillips, N. Guenther, and P. R. McAree, "When the dust settles: the four behaviors of lidar in the presence of fine airborne particulates," *Journal of Field Robotics*, vol. 34, no. 5, pp. 985–1009, 2017.

[10] J. Rapp and V. K. Goyal, "A few photons among many: Unmixing signal and noise for photon-efficient active imaging," *IEEE Transactions on Computational Imaging*, vol. 3, no. 3, pp. 445–459, 2017.

5 A 256×128 DTOF sensor with coincidence detection and progressive gating

Modeling and simulations are left unreliable unless the supporting concepts are tested in a real integrated circuit implementation. This chapter presents a 256 × 128 coincidence-based DTOF sensor which is designed based on the concepts discussed in the previous chapter. Availability of multiple pixels is particularly exploited to perform photon-by-photon processing including coincidence detection and progressive time-gating, to mitigate high background noise up to 10 klux. The content about this sensor is based on the work (accepted and) to be published in [1]. Precise timing, a fundamental requirement for depth estimation is achieved via injection-locking and mutually-coupled TDCs. The proposed timing solution results in an improved jitter performance by providing a superior overall phase noise. Various building blocks of the sensor are described followed by their characterization results within a flash LiDAR setup. The chapter begins with the discussion on the proposed timing solution, the content of which is based on the work published in [2].

5.1 Mutually-coupled TDC array

There are multiple ways to achieve a precise integrated timing reference on chip. The typical way is to implement a feedback system such as phase-locked loop (PLL) or delay-locked loop (DLL) [3], capable of frequency or delay scaling, synchronized with an off-chip crystal oscillator. Although several oscillator topologies exist including LC-tank oscillators with high quality factor, their use in image sensors is limited due to area constraints and consequently, ring-oscillator (RO)-based PLLs/DLLs are preferred. As the sensor array grows in size, it becomes challenging to maintain a precise and stable time reference over a large silicon area while not burdening the power budget.

In Chapter 3, Section 3.1.1, a power-efficient TDC sharing concept was presented. In addition to this aspect, it is also important to provide frequency/phase synchronization among the large number of ROs to obtain a reliable timing measurement. Inspired from the clock distribution solution in microprocessors [4, 5], concept of phase-coupling and injection locking is proposed as a timing solution to obtain a well-known and a stable reference independent of mismatches and PVT variations. The principle is extended to RO-based TDCs where the TDC itself operates continuously, based on the sampled approach presented in Chapter 3, Section 3.1.1.

Activity-dependent systems, where power consumption varies with incoming light (e.g., in event-driven approaches), are typically hard to predict and constant foreground calibration is required. In our proposed architecture, where the TDC power consumption is constant due to continuous operation, this is less of an issue. However, such designs are still subject to mismatch and PVT variations.

Thus, our proposed approach exploits the availability of continuously running oscillators (ROs) by operating them mutually coupled, through a single phase, in a process of injection-locking at the fundamental frequency. When combined, the oscillators provide a much lower phase noise, while operating synchronously (phase/frequency locking), even under potential oscillator mismatches, without any external circuit or additional power consumption. A single PLL is then implemented (using any node of the array as reference for the feedback path) to track PVT variations.

The concept of mutual coupling is shown in Figure 5.1, where the unit cell is highlighted. The coupling impedances are represented by $Z_{h,L}$, $Z_{h,R}$, $Z_{v,T}$, and $Z_{v,B}$. The oscillators are based on ROs, where capacitive and resistive coupling are studied, as depicted in Figure 5.2. Inductive coupling was not considered due to practical layout implementations, and the parasitic inductance of the wire was neglected due to relatively low operation frequency and short length.
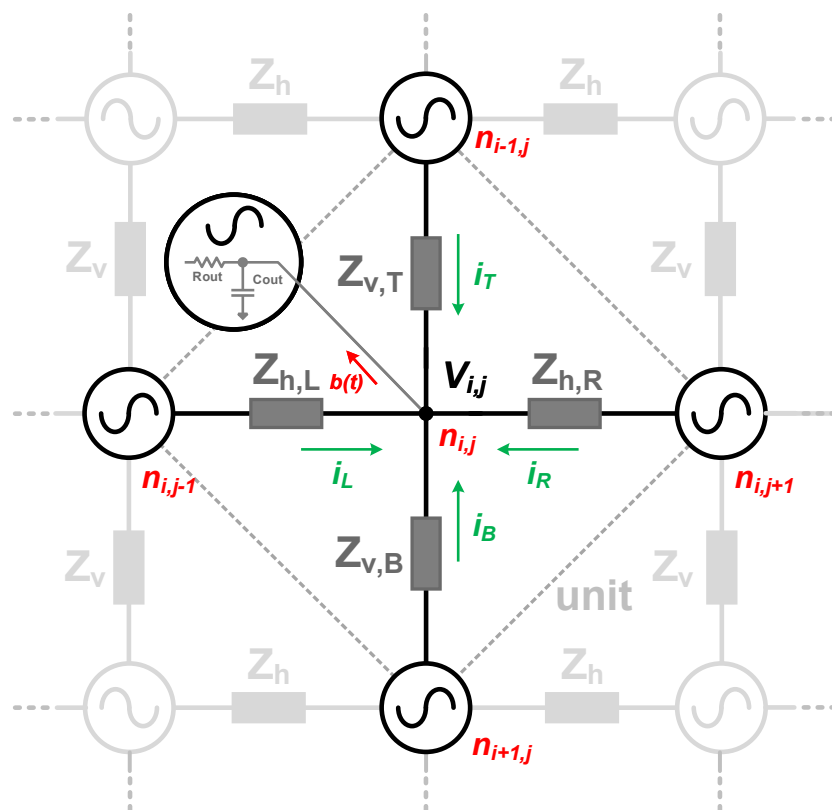
Figure 5.1 – Generic mutually coupling oscillators concept.

5.1.1 Non-linear modeling

Injection locking has been successfully used in many applications, such as high-frequency clock division [6], quadrature generation [7], clock distribution [4], etc. The effect has been extensively studied by several authors, based mostly on the generalized Adler's equation [8, 9], and it is not in the scope of this thesis to revisit the physics of the process further. Instead, we intend to provide a useful tool by applying this concept to the design of DTOF image sensors.

The dynamics of the system can be analyzed by performing a nodal analysis on the model shown in Figure 5.2. The process of synchronization occurs by injection-locking through the fundamental frequency, at a single node of each oscillator. The strength of the coupling element and the quality factor (Q) of the oscillator will define the maximum injection bandwidth, settling time, and sensitivity to neighboring disturbances, which will be discussed further.
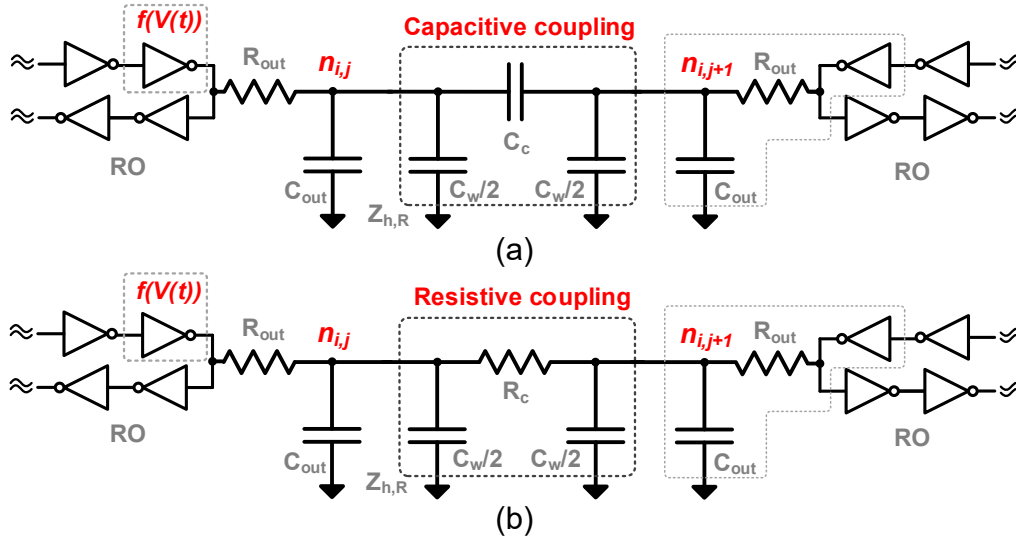
Figure 5.2 – (a) Capacitive and (b) Resistive coupling elements between two generic ring oscillators (ROs) (only $Z_{h,R}$ shown).

A non-linear phase macromodel is used to investigate the injection phenomenon [10]. The ROs dynamics are solved through ordinary differential equations at node $n_{i,j}$, shown in Figure 5.1, under the influence of its neighboring oscillators, at nodes $n_{i-1,j}$, $n_{i+1,j}$, $n_{i,j-1}$, $n_{i,j+1}$, and extrapolating it to the entire system. The numerical analysis of the perturbations is based on the Floquet theory of periodically time-varying systems [11] of ordinary differential equations.

The steady state voltage response of an oscillator, in the absence of any perturbation, can be represented by the time-dependent function $V_s(t)$. Under an external perturbation, $b(t)$, the RO response becomes:

$$V_{(i,j)} = V_s(t + \alpha(t)) + y(t), \tag{5.1}$$

where the term $\alpha(t)$ is the phase deviation caused by the disturbance $b(t)$. The perturbation $b(t)$ in this model is represented by currents from the neighboring oscillators i_L, i_R, i_T, i_B, as shown in Figure 5.1. The term $y(t)$ is the orbital deviation reflecting any gain error, in the presence of this

external perturbation. However, this term will not be considered for further analysis, as amplitude variations are negligible and the effect of the injection mechanism on the phase of the oscillator is dominant [10]. Thus, the perturbed steady state solution can be approximated by $V_s(t + \alpha(t))$.

A current analysis of the capacitive coupling, shown in Figure 5.2a, at node $n_{i,j}$, can be obtained by:

$$\frac{dV_{(i,j)}}{dt} = \frac{f(V(t))}{R_{out}(C_{out} + 2C_w + 4C_c)} - \frac{V_{(i,j)}}{R_{out}(C_{out} + 2C_w + 4C_c)} + \frac{C_c}{(C_{out} + 2C_w + 4C_c)} \cdot \frac{d}{dt}(V_{(i+1,j)} + V_{(i-1,j)} + V_{(i,j+1)} + V_{(i,j-1)}),$$

(5.2)

where $V_{(i,j)}$ is the nodal voltage, and R_{out} and C_{out} are defined by the RO output impedance. C_w is the shunt parasitic capacitance from the coupling line, and C_c is the effective coupling capacitance. The term $f(V(t))$ models the RO stage non-linearity for the delay stage preceding the coupled node by a hyperbolic tangent function, $tanh(G_m V(t))$, where G_m is the large-signal stage transconductance.

Similarly, in the case of a resistive coupling element (Figure 5.2b), the voltage at node $n_{i,j}$ is given by:

$$\frac{dV_{(i,j)}}{dt} = \frac{f(V(t))}{R_{out}(C_{out} + 2C_w)} - \frac{V_{(i,j)}}{R_{out}(C_{out} + 2C_w)} + \frac{V_{(i+1,j)} + V_{(i-1,j)} - 2V_{(i,j)} + V_{(i,j+1)} + V_{(i,j-1)} - 2V_{(i,j)}}{R_c(C_{out} + 2C_w)}.$$

(5.3)

Equations (5.2) and (5.3) were numerically solved in MATLAB for TDC networks of 4 × 4, 8 × 8, and 16 × 16 elements, using seven-stage ROs, although the modeling holds true for any number of RO stages, just with an impact on its dynamics. The networks are terminated (at their boundaries) by the same coupling element, but open at one of its ends.

For the following simulation, the parameters R_{out}, C_{out} and G_m (refer Figure 5.2) were chosen (based on typical values) to obtain an average oscillation period of 2 ns (500 MHz). Random mismatches were also included, impacting on about ±15% period variation among the oscillators, in order to verify the robustness of the method.

The steady state voltage for a 16 × 16 RO array, using coupling resistance R_c = 250 Ω, is shown in Figure 5.3a. The ROs started with a random period of 2 ± 0.3 ns (500 ± 77 MHz) and completely arbitrary phases. After 18 cycles (36 ns), the ROs reached locking with a steady-state phase skew of 114 ps. Any disturbance on chip, such as supply spikes and charge injection on the ROs phases, directly affects the attained steady state. Although open-loop TDCs cannot recover from such disturbances, the proposed approach is self-regulated by the local feedback from neighboring TDCs, allowing continuous phase/frequency locking. In order to simulate this effect, 32 of the coupled 16 × 16 array nodes were injected with a disturbance that corresponded to 33% of the overall node charge, after 25 clock cycles, in their most sensitive phase—zero-crossing (see Figure 5.3a). The process of re-synchronization started immediately after the disturbance, taking about seven clock cycles (14 ns) to reach steady state once again (the same phase skew as before the injection).

Figure 5.3b shows similar simulation, but for a capacitive coupling of C_c = 240 fF. After steady state was reached (31 clock cycles), 32 ROs were disturbed with 33 % of the total nodal charge. The process of re-synchronization took about 20 clock cycles to return to steady state.
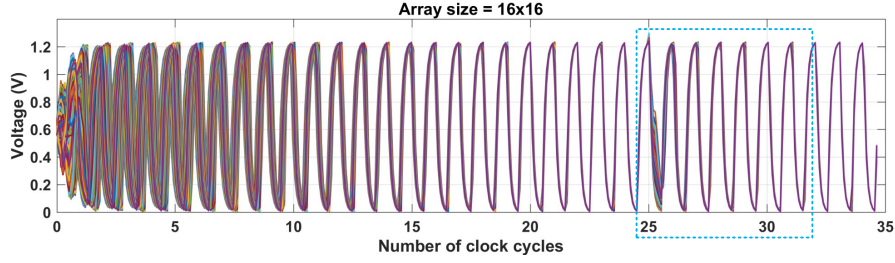
(a) Resistive coupling.
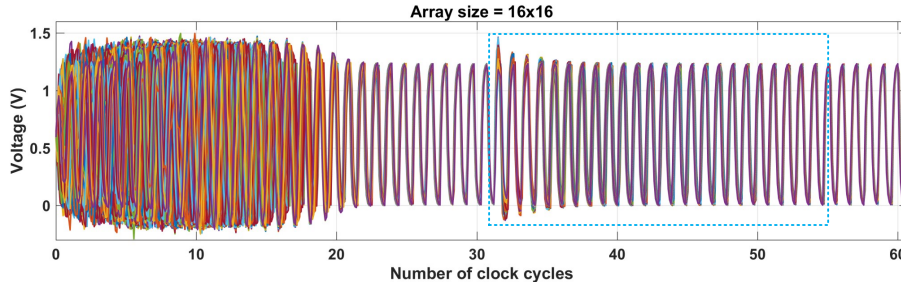
(b) Capacitive coupling.

Figure 5.3 – Voltage waveforms of a 16×16 coupled RO network under $\pm 15\%$ random initial conditions and with disturbance introduced in 32 ROs in the case of (a) resistive coupling with R_c = 250 Ω and (b) capacitive coupling with C_c = 240 fF.

The settling time can vary based on the number of ROs disturbed, the size of the array, and coupling strength. Figure 5.4 shows this dependency, over a number of disturbed oscillators for the cases of resistive and capacitive coupling.
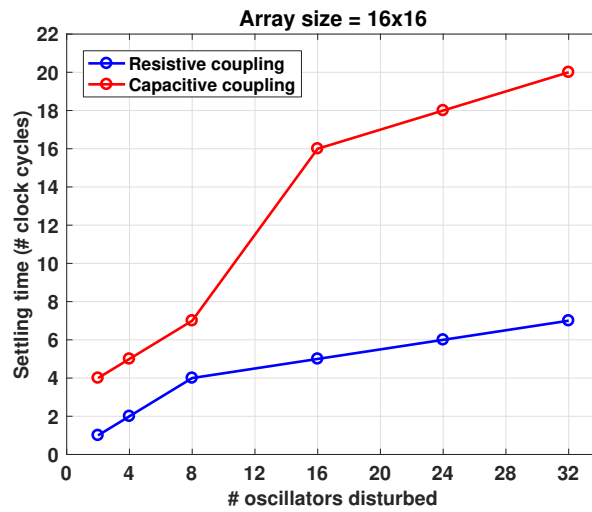
Figure 5.4 – Steady state recovery time (in cycles), after different number of ROs disturbed.

Frequency mismatches and/or PVT variation directly affect the settling time and phase skew. Variations in the coupling impedance also have an impact on the steady state. Thus, apart from ±15 % variation on the RO periods, another ±10 % on the coupling impedance was included in the simulations. Simulation results for the case of capacitive coupling are shown in Figure 5.5.

(a)　　　　　　　　　　　　　　　　(b)
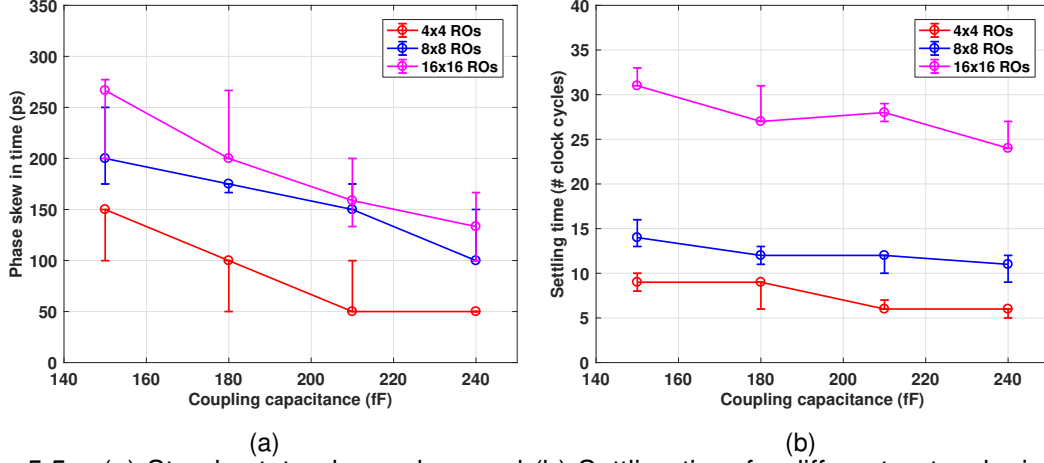
Figure 5.5 – (a) Steady state phase skew and (b) Settling time for different network sizes and coupling capacitance. Settling time is defined by the phase mismatch below 1/(67%) of value obtained in (a); vertical bars indicate variation due to ±10% mismatch in C_c.

The phase skew increased with the number of coupled ROs and for lower coupling impedances. For instance, for the capacitive coupling ($C_c = 240$ fF), it took about six clock cycles for a 4×4 array, to reach steady state, while it took 24 clock cycles for the 16×16 array with the same C_c, as can be seen in Figure 5.5b. Similarly, the same steady state parameters were obtained for the case of resistive coupling, as shown in Figure 5.6. A 600 Ω coupling resistance produced a maximal residual phase skew of 280 ps for the 16×16 array, while for the 4×4, the skew was only 60 ps. Higher coupling resistances also resulted in longer settling time, as shown in Figure 5.6b.

(a)　　　　　　　　　　　　　　　　(b)
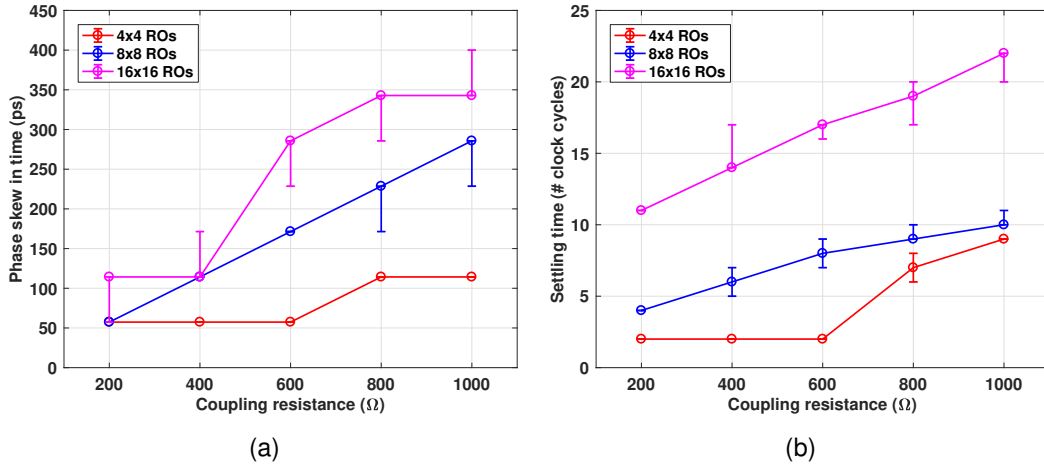
Figure 5.6 – Steady state (a) phase skew and (b) settling time, for different network sizes and coupling resistance. Settling time is defined by the phase mismatch below 1/(67%) of value obtained in (a); vertical bars indicate variation due to ±10% mismatch in R_c.

Charge injection through capacitive coupling only occurs during phase transitions, due to transient

voltage variation, which produces longer settling time. Faster coupling is possible by increasing the coupling capacitance. However, due to area constraints and excessive parasitic capacitance, it may limit the overall linearity and operating frequency. Resistive coupling, however, can provide much stronger coupling (lower impedance) at smaller areas, being more suitable for our application.

These results provide a quick insight into the dynamics of mutually coupled ROs, using different types of coupling and different strengths, thus enabling better design choices based on the target application. They also provide a qualitative and quantitative analysis of the synchronization process, allowing better planning for calibration– both foreground and background.

5.1.2 SPICE-compatible model

In addition to the macro-model developed in Section 5.1.1, a SPICE-compatible (based on Verilog-A) model was also used, since electronic circuits are normally designed and simulated in such environments and the interaction with other signals on the readout integrated circuit (ROIC) can be evaluated. The model comprises a large-signal differential transconductance, coupled to a capacitive impedance to form each stage of the oscillator [12]. The frequency is controlled by a current source (current-starved RO) and it includes noise effects (thermal and flicker) that are naturally up-converted during oscillation. Although this model can be adapted to different numbers of stages and topology, it was designed to match the RO implemented and measured in Section 4.3, which is composed of an 8-stage pseudo-differential topology, as shown in Figure 5.7.

Apart from synchronization, the uncorrelated noise between ROs is filtered out. On average, ROs have low power efficiency–FOM [13]–on the order of 145–160 dB, which is related to their noise (phase noise/jitter) and power consumption. For example, without any elaborate filtering, a 500 MHz RO, consuming 400 µW, and FOM of 150 dB, produces an integrated root mean square (RMS) jitter [14] of about 110 ps (1–100 MHz integration window), which is prohibitively large for millimetric precision measurements, requiring feedback loops for noise filtering at the expense of power, area, and complexity. However, by coupling multiple oscillators, the uncorrelated noise among them is filtered out, providing a reduction in phase noise (and jitter) at the system level by $10 \cdot \log_{10} M$ [15], where M is the number of coupled oscillators. Although the FOM of the system remains the same (overall power consumption increases and the noise reduces M fold), at each oscillator, the FOM appears to improve also by $10 \cdot \log_{10} M$, with negligible extra power consumption.

To demonstrate the described effect, multiple oscillator array sizes were coupled, and the simulation result is depicted in Figure 5.8. The phase noise reduction of the uncorrelated noise (low offset frequencies) behaved as predicted. For the correlated noise (high offset frequencies), such as the thermal noise on the coupling elements, the benefit of the coupling was reduced. A comparison between full SPICE and Verilog-A models was also evaluated. The latter took only 1.5 % of the computational power and simulation time of the former, at equivalent precision, providing an essential tool for full chip co-simulation.
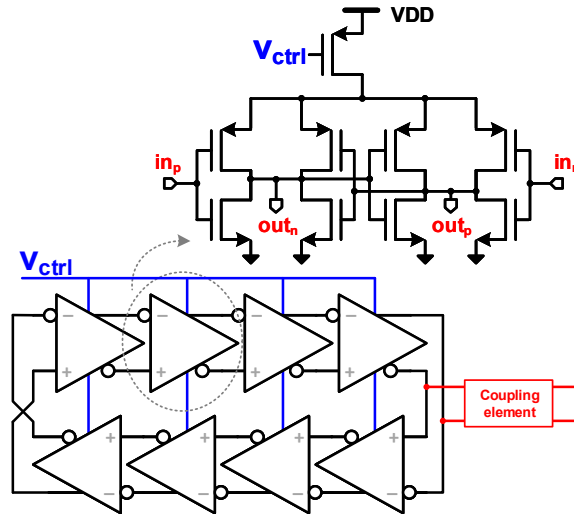
Figure 5.7 – Current-starved 8-stage pseudo-differential RO.
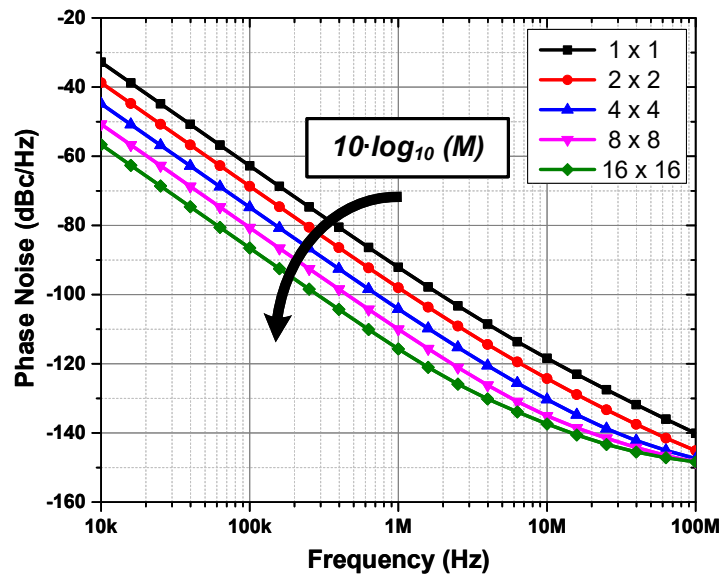
Figure 5.8 – Simulation of phase noise reduction from 1 (1×1) to 256 (16×16) mutually-coupled ROs.

The implemented block diagram can be seen in Figure 5.9. Due to resistive coupling, the phase/frequency locking operates on the array at all time, and as a result, both at startup, when the ROs have arbitrary phases (and perhaps different average frequency), or during any disturbance in one or more of the ROs, the array will always be pushed back to a locked state. This is represented by the phase diagram at the bottom of Figure 5.9. Additionally, due to the nature of the operation and the fact that all ROs are synchronized and share a common control voltage ($CTRL$), a single PLL can be implemented to define the overall frequency and to track PVT variations, using a single regional phase as reference for the feedback loop.
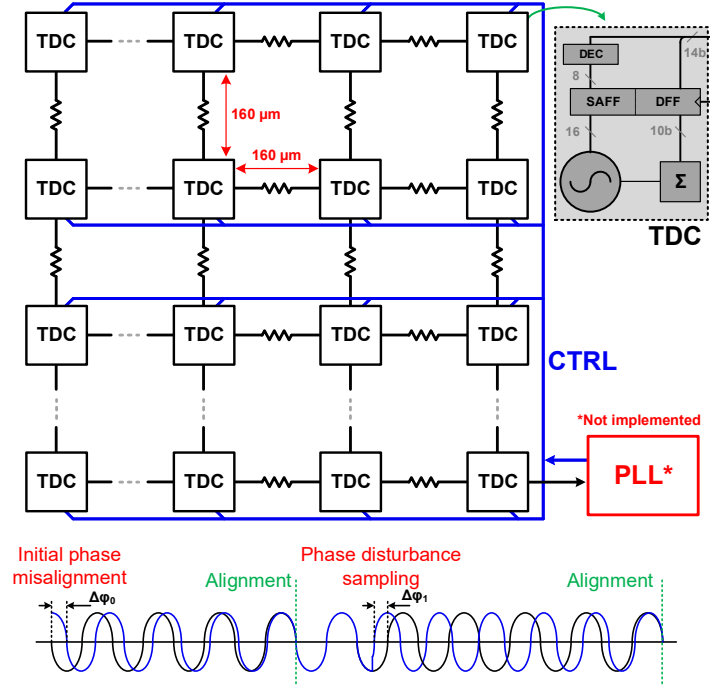
Figure 5.9 – Implemented 8×8 mutually-coupled TDC architecture and RO phase misalignment self-correction. PLL: phase-locked loop.

Thus, starting from the same 150 dB FOM RO at 0.5 GHz and coupling 64 ROs (in an 8×8 structure), the effective FOM was improved by $10 \cdot \log_{10} M \approx 18$ dB, to a moderate 168 dB FOM, which produced an integrated RMS jitter (1–100 MHz) of only 13.75 ps, instead of 110 ps as previously found.

For the final topology, an eight-stage, current-starved, pseudo-differential RO was implemented. Along with the RO, a 10-bit ripple counter and D-type and sense-amplifier flip-flops complete the TDC. Based on the conclusions from a power-efficient TDC sharing in Chapter 3, Section 3.1.1, a single TDC was expected to be shared among two independent subgroups of 8×8 pixels, as sketched in Figure 5.10. The resistive coupling used was implemented through a transmission gate, shown in Figure 5.10, so the performance in both modes could be compared. Moreover, it can be used to disable the coupling during initial calibration phase, where all ROs can be adjusted to roughly the same frequency, before coupling, thus improving INL and power efficiency.

5.2   Mutually-coupled TDC array

The prototype was fabricated using a 3D-stacked CMOS technology [16], as sketched in Figure 5.10. The 64 ROs were arranged in an 8×8 matrix, only on the bottom tier, which used low-power, 4 metal (3 thin + 1 thick) 65 nm TSMC technology, with 1.2 V core supply. The proposed technique is independent of the technology and transistor node, also suitable for monolithic implementation.

Coupled and uncoupled conditions were implemented and measured. To mimic the distribution in a real sensor, the TDCs were placed with a pitch of 160 μm, horizontally and vertically, thus achieving

Figure 5.10 – Transmission gate as resistive coupling element
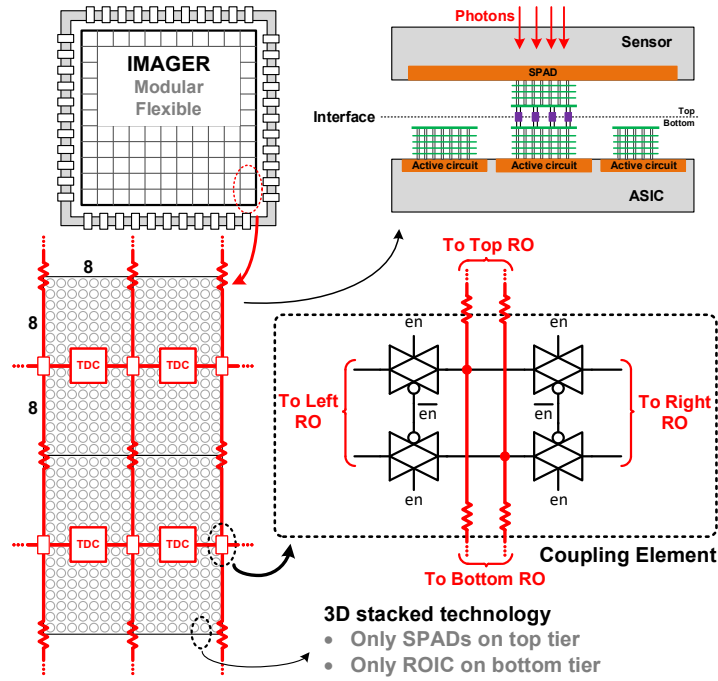
a total area of $1.3 \times 1.3 \, \text{mm}^2$. Each TDC occupied an area of $76 \times 7.2 \, \mu\text{m}^2$, including RO, a 10b counter, sampling latches, and decoupling capacitors, which occupied 60% of the TDC array, whose layout is shown in Figure 5.11.

The effects of the coupling were investigated by measuring the high-frequency clock from the ROs. All 64 ROs were combined through multiplexers and carefully routed to a single high-speed output, connected to a Rohde & Schwarz FSUP-50 signal source analyzer or a Keysight Infiniium DSOS804A real-time oscilloscope for spectrum and phase noise or jitter measurements, respectively.

Figure 5.11 – TDC layout.
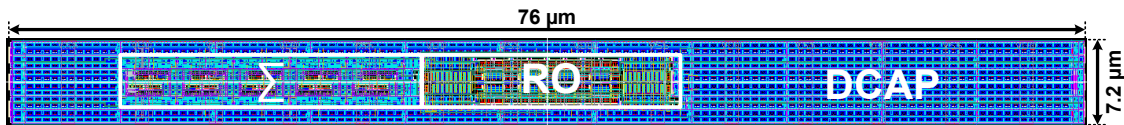
A large IR-drop was present in our fabricated chip because only a few metal layers (3 thin + 1 thick) were available. Its effects on frequency variation can be seen in Figure 5.12a. Although the intrinsic frequency of each RO varied substantially (about 24 %), the mutual coupling was very robust, reaching frequency locking as shown in Figure 5.12b. Ideally, the ROs should be independently tuned to roughly the same frequency (which can be done by foreground calibration), to ease the process of frequency correction, power consumption reduction (less charge exchange between oscillators), and local INL minimization.
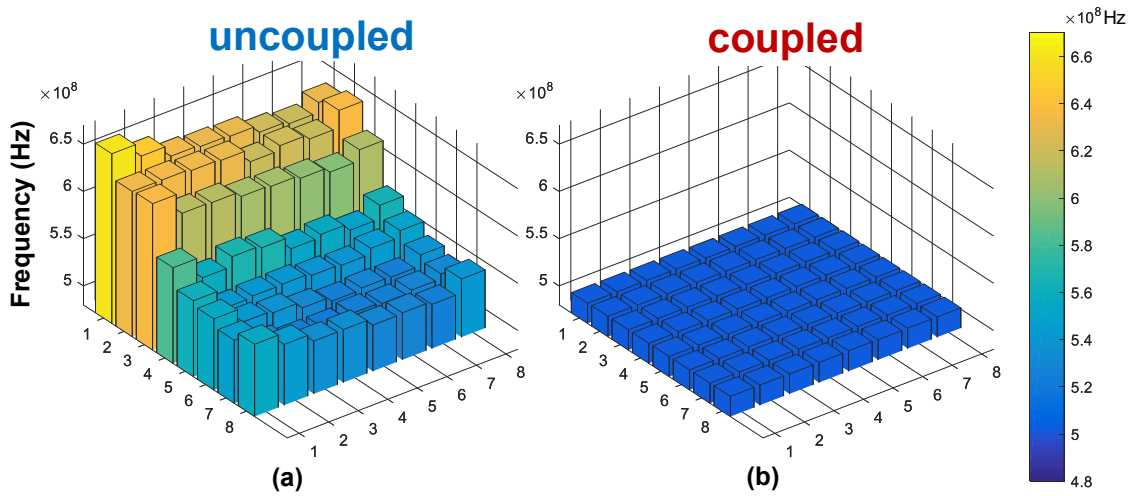
Figure 5.12 – Individual frequencies of uncoupled and coupled modes.

The array was measured in the whole range of frequencies, from 150 to 800 MHz. The mean values and variation bars, in coupled and uncoupled modes, are plotted in Figure 5.13. Before coupling, the spread in the instantaneous frequency was 22–26 %, whereas under mutual coupling, this spread reduced to less than 0.11 %. Moreover, under coupling and, consecutively, locking, all ROs operated in the same average frequency.
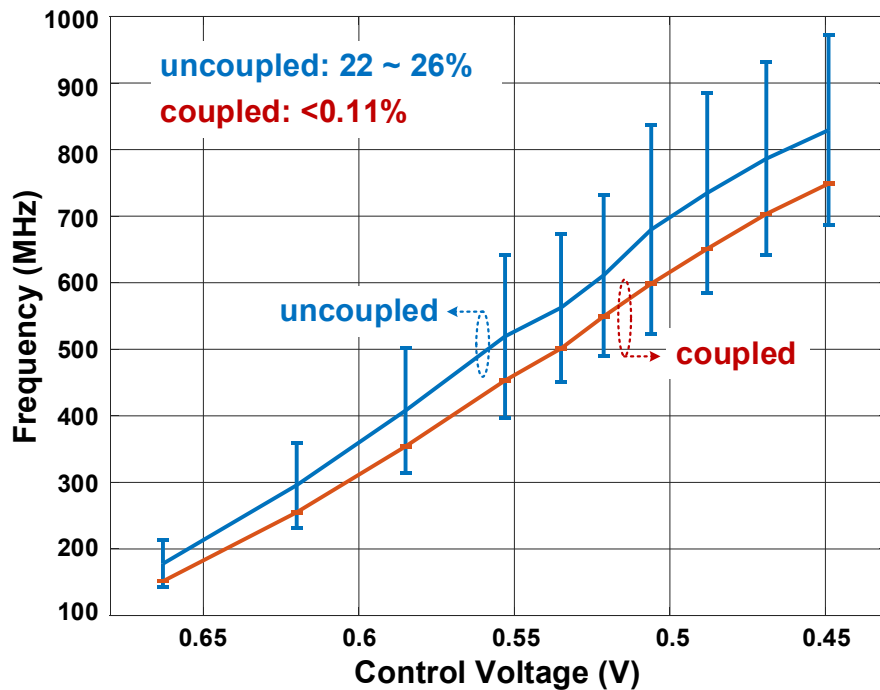
Figure 5.13 – Frequency variation of coupled and uncoupled modes, for different average frequencies.

It is evident as one observes that after coupling, the operating frequency was lower than the average of the individual oscillators, both in Figures 5.12 and 5.13. The reason is the effect of parasitic

109

capacitance from the coupling element and lines, which was only visible when coupling was enabled. For that reason, the RO was designed with asymmetric stages (stronger for the coupling phase), thus maintaining overall linearity when coupled.

Injection locking technique does not improve the linearity of the individual TDCs, and in fact trades resolution uncertainty for short-range INL. For instance, if all TDCs in the array had the same performance (the same RO frequency), by coupling them, they would present the same non-linearity as an uncoupled TDC. However, if variations were present (IR-drop, PVT variations, mismatch, etc.), they would still be locked in frequency and phase, as demonstrated in this work, but the necessary phase alignment would cause an abrupt non-linearity, increasing the overall INL. An example phase correction is presented in Figure 5.14a. For an ideal case of perfectly linear TDC, but with different speed, at every RO period the phase needs to be aligned, generating a local INL whose maximum and minimum would depend on the RO period difference to the average period ($|INL_{MAX|MIN}| = |T_{RO} - T_{AVG}|$). In the presence of intrinsic TDC non-linearity, $|INL_{MAX|MIN}|$ will be a combination of both effects. An illustration of the local INL is shown in the bottom of Figure 5.14a.

For these reasons, only the uncoupled TDC non-linearity is presented, which was evaluated using a density test method, and the results are plotted in Figure 5.14b. The maximum INL and DNL were below 3 LSB and 2 LSB, respectively, over the whole 14 bits of dynamic range, without calibration.

(a)   (b)
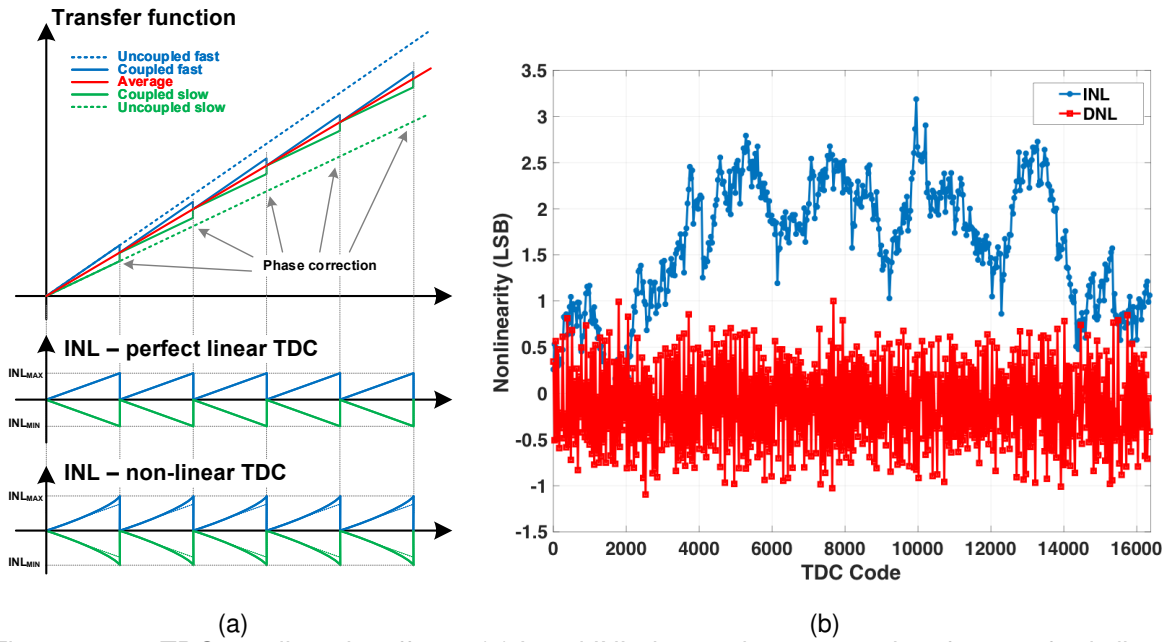
Figure 5.14 – TDC non-linearity effects: (a) Local INL due to phase correction, for a perfectly linear TDC and a non-linear TDC; (b) Uncoupled TDC INL and DNL, without calibration.

The phase noise is a key parameter to confirm the effectiveness of mutual coupling on noise filtering and synchronization. Figure 5.15 shows an 18 dB phase improvement provided by the coupling, for most of the frequency offsets, following the theory. For high-frequency offsets, the coupling elements' thermal noise dominated the phase noise, and due to its correlation within the array, the coupling was not as effective.
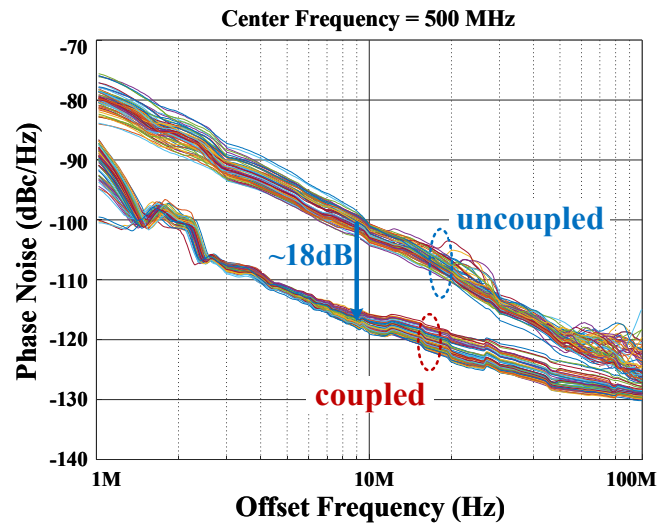
Figure 5.15 – Measured phase noise comparison, for uncoupled and coupled conditions, for all 64 ROs at 500 MHz center frequency.

The phase noise of each RO is plotted along with the integrated RMS jitter in Figure 5.16. Both measurements were performed with the ROs coupled and uncoupled, at a center frequency of 500 MHz. The phase noise at 3 MHz offset frequency showed the effectiveness of the coupling, reaching an 18 dB improvement on average. The jitter reduction reached 14 dB (instead of 18 dB), due to the presence of correlated noise from the coupling elements.
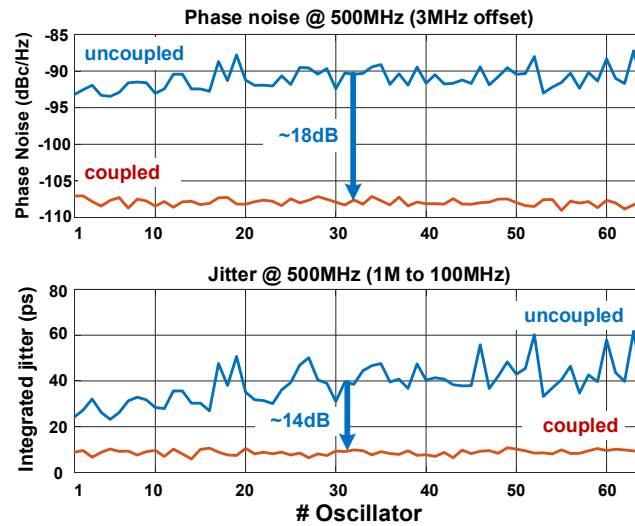
Figure 5.16 – Phase noise and integrated root mean square (RMS) jitter comparison for uncoupled and coupled modes, for all 64 ROs at 500 MHz center frequency.

Figures 5.12 and 5.16 show a variation of frequency, phase noise and jitter under "uncoupled" mode. The reason being the extreme IR-drop present in the system, where the oscillators close to the edge of the chip (lower indexes, starting from #1) had lower impedance to the supply, and their PMOS current source had higher drain–source voltage, allowing stronger inversion, and thus lower noise factor. Although such conditions existed, it did not affect the synchronization and the noise filtering technique proposed here, which was proved by the phase noise and jitter under "coupled" mode. Nevertheless, the integrated RMS jitter reduction, from about 40 ps to less than 9 ps, was enough for our application, which contained other sources of noise (e.g., SPAD timing jitter [16]) that were much higher.

5.3 *Jatayu* – A 256×128 DTOF sensor for flash LiDAR

A DTOF sensor, named, *Jatayu*[1], was designed based on the shared approach introduced in Chapter 3, Section 3.1 and the TDC arrangement and mutual coupling were implemented as described in the previous section. The high-level overview of the sensor is shown in Figure 5.17; the sensor is fabricated in a 3D-stacked technology where the top tier consists of an array of 256×128 SPADs at a pixel pitch of 7 μm, designed in a 45 nm CIS process and the bottom tier consists of dedicated pixel-electronics designed in an ultra low-power, 6-metal (4+2 thick metal layers), 22 nm CMOS technology. Since the SPADs were designed outside this thesis, the characterization of the 7 μm SPADs is not presented here. The sensor architecture is modular in nature where a module consists of an array 2×16×8 pixels (marked in Figure 5.17) and the whole sensor (256×128) can be visualized as an array of 16×8 identical modules.
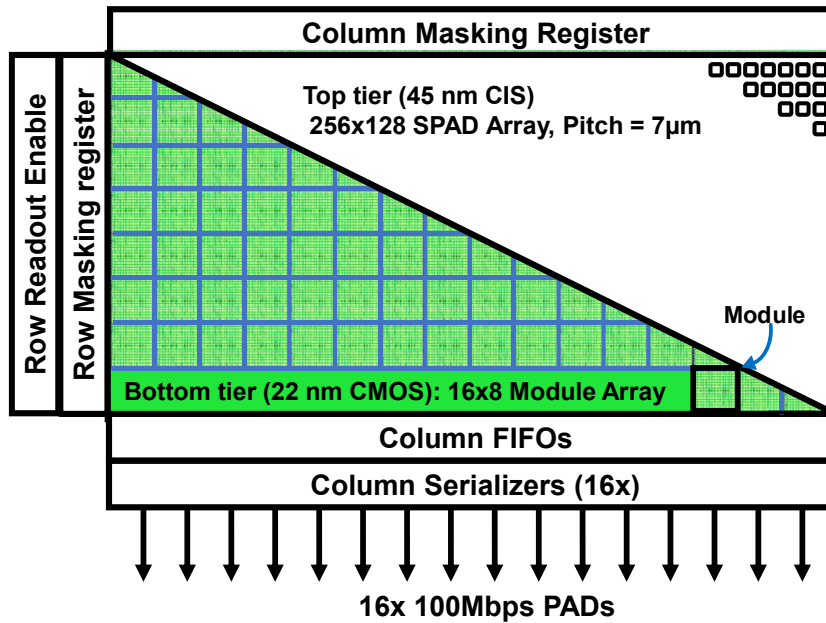
Figure 5.17 – High-level visualization of the sensor.

[1]The sensor is named as *Jatayu* which is a mythical bird, specifically a vulture. Vultures are known for their keen eyesight. Consequently, the sensor was named so, due to its deep vision abilities.

The block diagram of a module (2×16×8) is shown in 5.18, where two identical units of a sub-module (16×8) is described. The two sub-modules share an always-on TDC with autonomous access. Each pixel has its own passive quenching and recharge circuitry, already described in Figure 2.27 of Chapter 2. The pixel circuit generates an equivalent digital signal (D[127:0] in Figure 5.18) upon photon arrival which is propagated through a combination tree, referred to as the coincidence tree as seen in Figure 5.18.
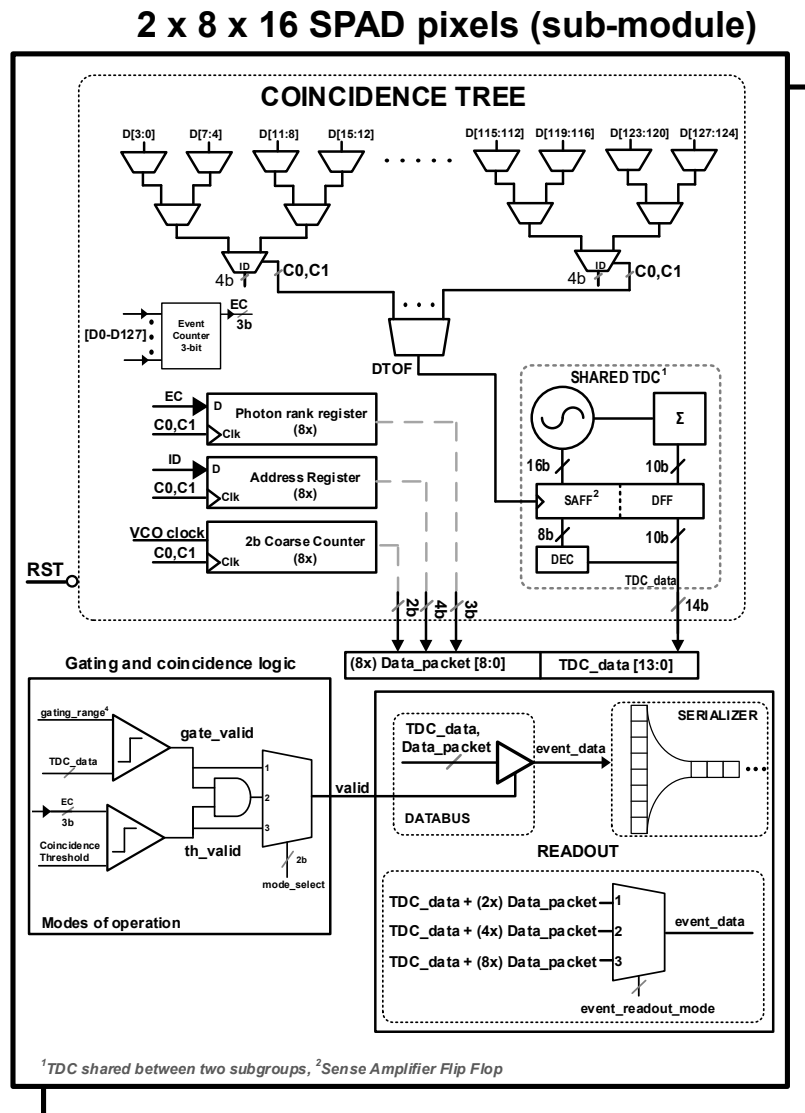
Figure 5.18 – Block-diagram of a module- 2×16×8.

The coincidence tree is responsible of managing multiple events within a sub-module while also detecting coincidence among them. When a photon bunch impinges on the sensor, the first photon detected in a given pixel triggers a pulse, the leading edge of which starts a coincidence window, t_{cw}. Simultaneously, this first-event pulse propagates through the tree, generating a signal, $DTOF$

which samples the instantaneous states of the TDC. In parallel, the corresponding pixel ID of the first event gets sampled. The window duration, t_{cw}, is tunable between 500 ps and 2.2 ns. The tree self-resets itself at the end of every detection cycle, automatically after t_{cw} has elapsed to allow subsequent detections. The dead time for detections is given by the tree dead time, as explained through Section 3.1 of Chapter 3; in this implementation, the dead time is about 700 ps, thus allowing, 1.4 Giga-conversions per second. However, it should be noted that the actual timing throughput from the chip is dictated by the readout (I/O) bandwidth.

All SPADs in the sub-module that fire after the first event within t_{cw} are accounted for and ranked (up to the 8th); this information is latched, along with the corresponding IDs, in the Photon Rank Register (PRR) and in the Address Register (AR), respectively. There is also a Coarse Counter (CC) which stores the timestamp MSBs of these 8 events, while the full timestamp (14-bit) for the first event is available from the TDC. The full data in PRR, AR, and CC is concatenated to form a $Data\_packet$. The number of detected photons within the coincidence window is stored in the Event Counter whose output (3-bit), EC is compared with the coincidence threshold, th, representing the coincidence level. A configurable coincidence level between 2 and 7 is possible. Whenever $EC > th$, a valid signal is generated and the $Data\_packet$ along with the TDC data is committed to the serializer and eventually read out off-chip via FIFOs and digital I/Os. The timing diagram of the coincidence mechanism is shown in Figure 5.19, for valid and invalid events.
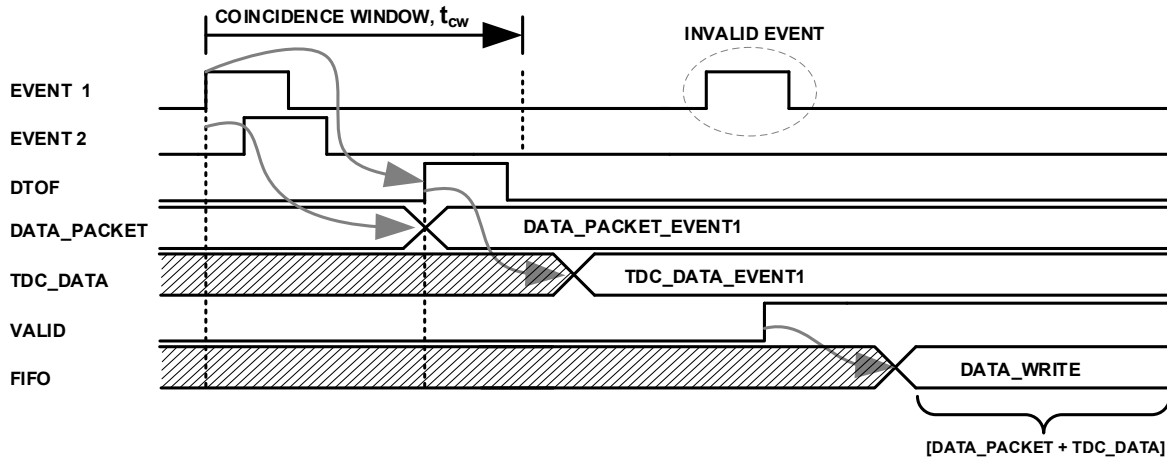
Figure 5.19 – Timing diagram of coincidence detection for an example case where threshold, $th = 2$.

Mode selection allows every sub-module to be independently configurable in three modes of operation, as seen in Figure 5.18. The 2-bit signal, $mode\_select$, when set to $'11'$, enables the coincidence mode which was just described. Another mode configures the module to operate under progressive gating when $mode\_select$ is set to $'01'$ and finally, a combination of coincidence and gating is possible when $mode\_select$ is set to $'10'$. When progressive gating is enabled, a 14-bit timestamp generated by the TDC is compared with a pre-defined gating range (5-bit) in order to validate an event. Up to 6 progressive ranges can be defined as shown in the block diagram. Please note that in order to cater to photon-starved scenarios, the sensor can also be configured to operate

in a default mode where every incoming photon is timestamped without any coincidence or gating.

Based on the event readout mode selected, a maximum of 8x $Data\_packet$ can be read out when $event\_readout\_mode = 3$, enhancing the timing throughput per sub-module by 8x, when compared to a single TDC data alone. The next sections will describe the various building blocks of a sub-module along with their implementation.

5.4 Coincidence tree cell

The coincidence tree unit cell and the address decoding logic for 4 inputs is shown in Figure 5.20, where a half adder, HA, is used to implement the tree in a binary weighted fashion to function as a binary counter. The outputs, Q, from pixel circuit (Figure 2.27 of Chapter 2) are sampled by input samplers (D-flip flops) whose outputs are then propagated into the coincidence tree. These signals are indicated as $D0 - D3$ in Figure 5.20. The tree, apart from propagating the first event to the TDC, in parallel, also decodes the pixel ID of the contributing events.
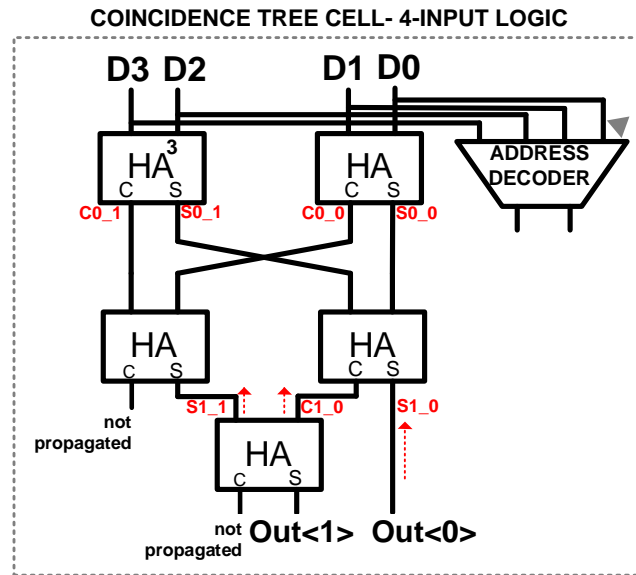
Figure 5.20 – Coincidence tree unit cell for 4-inputs.

Output signal $Out < 0 >$ and $Out < 1 >$ represent the 2-bit output with binary weights corresponding to 2^0 and 2^1 respectively. Upon arrival of the first event, $Out < 0 >$ toggles to logic-1 and this activates the address decoder block for the first event. Upon a second event, as $Out < 1 >$ toggles to logic-1 and $Out < 0 >$ toggles to logic-0, the address decoding block for the second event activates.

The sub-module (16×8 pixels) is further clustered into $8 \times 4 \times 4$ pixels where every group of 4×4 pixels has its dedicated 16-input coincidence tree, the block diagram is shown in Figure 5.21. The 16-input tree is comprised of unit cells shown in Figure 5.20. The two LSBs signal, C0 and C1 (2^0

and 2^1), from the 8x 16-input tree are combined and propagated further to generate the $DTOF$ signal which samples the TDC while the 8x 4-bit pixel address is stored in a local address register, AR. The AR typically stores the pixel ID of the first incoming event within every cluster of 4×4 pixels. However, if two events occur simultaneously or, temporally very close (< 30 ps), the AR stores the address of the second event and a flag bit is used to identify such a situation. In general, after every detection and by the end of t_{cw}, the input samplers and the ARs self-reset in order to allow next detection.
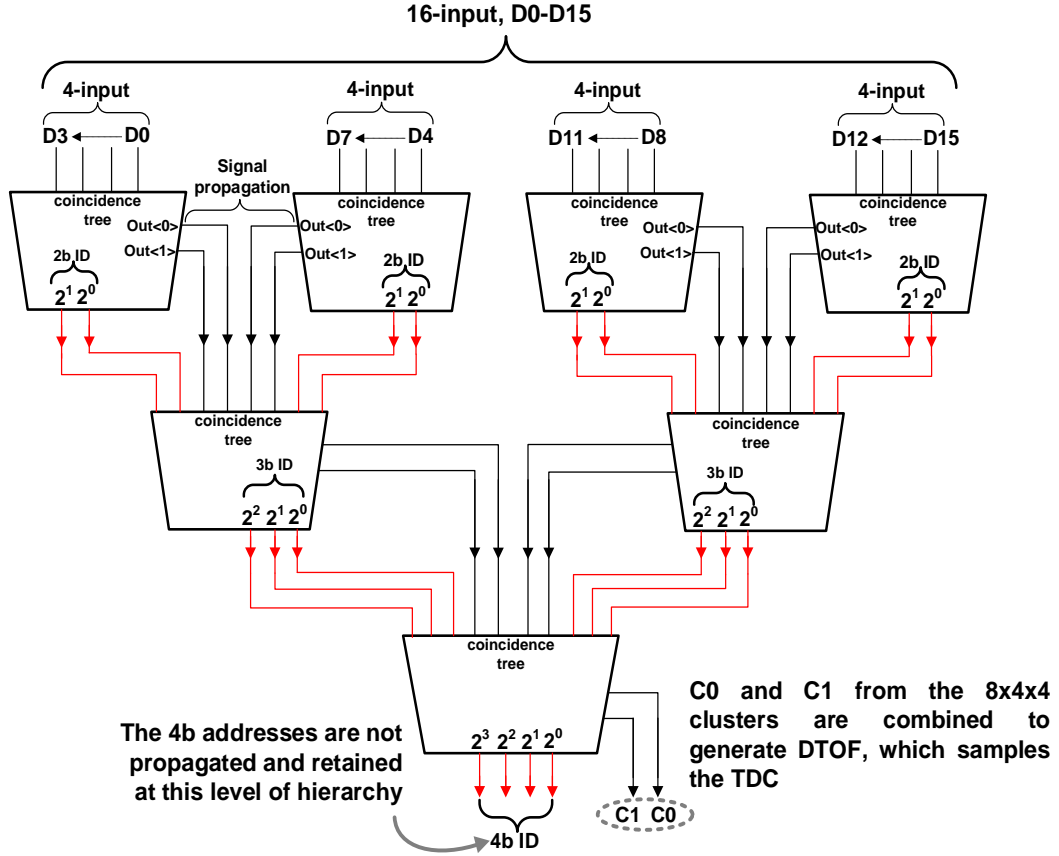
Figure 5.21 – 16-input coincidence tree.

5.5   Coarse counter (CC)

Within every sub-module, 8x coarse timestamping units are implemented for $8 \times 4 \times 4$ pixels. The block diagram of the coarse counter is shown in Figure 5.22. The output signals, C0 and C1 from the 16-input coincidence tree (see C0, C1 in Figure 5.21) are utilized to enable coarse timestamping via a 2-bit binary counter clocked by the VCO signal, VCO_clock, distributed from the VCO in the shared TDC. A buffered version is distributed to all the 8 coarse counters with minimal skew. The counter is enabled on the arrival of the first event with reference to the coincidence window (see en_ctr and $window$ signal in Figure 5.22) so that the counting happens only for the duration, t_{cw}.

The counter value is stored in a local register which is reset (rst_ctr) at the end of every detection cycle, automatically after t_{cw} has elapsed.
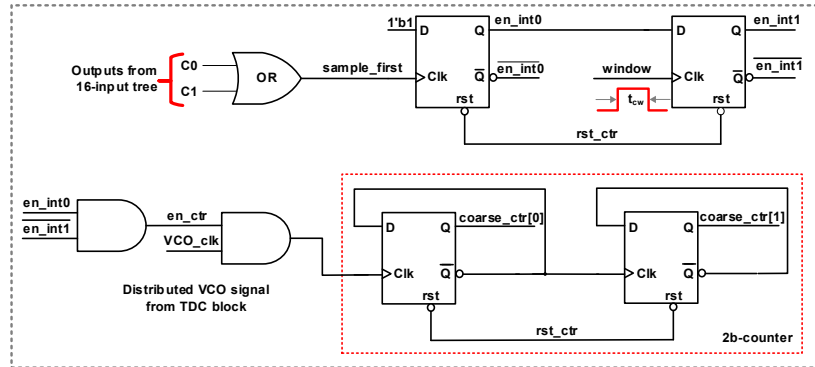
Figure 5.22 – Dedicated coarse counter for a cluster of 4×4 pixels.

5.6 Progressive gating control

Every sub-module (16×8 pixels) can optionally operate in a progressive gating mode which allows target-selective ranging as well as an improvement in the SBR. The high-level concept is shown pictorially in Figure 5.23.

Progressive gating concept – |d2| > |d1|
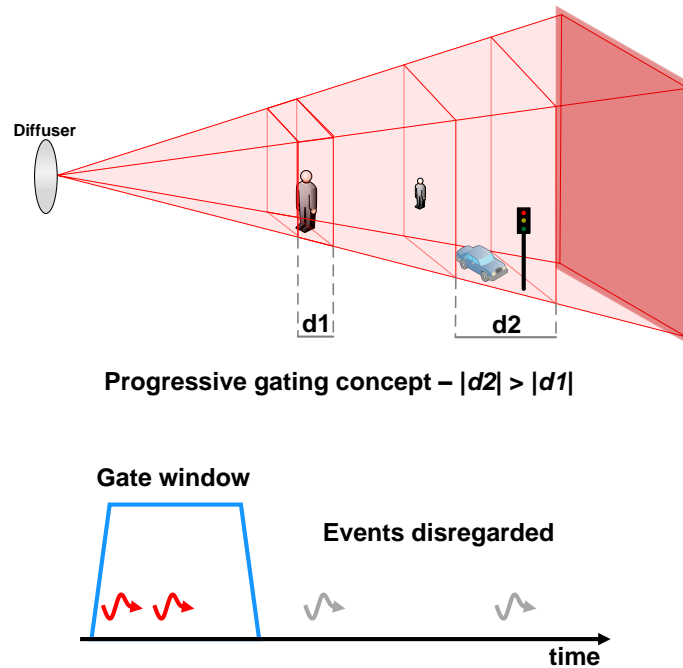
Figure 5.23 – Pictorial representation of progressive gating.

The progressive functionality is enabled to exploit the fact that a target at farther distances may not demand a fine resolution and consequently, the gate window can be coarser (longer, $d2$). As the

target approaches closer to the sensor, the gate window becomes finer (shorter, $d1$).

The gating control unit implemented in the sensor is shown in Figure 5.24. 10-bit counter data from the 14-bit TDC code is sent to the gating control unit ($VCO\_counter[9:0] = TDC\_data[13:4]$ in Figure 5.24) to be compared with a predefined comparator reference ($gating\_comparator$) to validate an event within a gating range. Any timestamp not satisfying the comparison is discarded and not read out. The $gate\_width$ is 5-bit and a maximum of 6 progressive subranges are implemented. The 6 subranges span over the entire 10 bits of the $VCO\_counter$ such that the whole TDC range is covered. Starting from $gating\_range = 1$, every subsequent subrange is coarser compared to its previous $gating\_range$. The appropriate $gating\_range$ can be chosen via a multiplexer, based on a prior (and faint) knowledge of the target.

As seen in Figure 5.24, $gating\_range = 1$ is the finest, covering [4:0] LSBs of the $VCO\_counter$. The absolute value of this range is given by

$$gating\_range_{abs} = \left[\frac{2^0}{VCO\_clock} \, , \, \frac{2^5 - 1}{VCO\_clock} \right] , \tag{5.4}$$

and the gate resolution is given by,

$$gate\_resolution_{abs} = \left[\frac{gating\_range_{abs}}{2^5} \right] . \tag{5.5}$$

For example, if the VCO oscillates at 1 GHz, the gating range is [1 , 31] ns and the gate resolution is ≈ 1 ns.

Now, on the contrary, $gating\_range = 6$ is the coarsest, covering [9:5] MSBs of the $VCO\_counter$. In this case, the absolute value of this gating range is given by,

$$gating\_range_{abs} = \left[\frac{2^5}{VCO\_clock} \, , \, \frac{2^{10} - 1}{VCO\_clock} \right] , \tag{5.6}$$

and the gate resolution is given by Equation 5.5. Therefore, for the VCO frequency of 1 GHz, the absolute value of the gate [32 , 1023] ns and the gate resolution is ≈ 32 ns.

While the gate width itself is 5-bit, the discrete nature of the implementation allows tuning of the "absolute" value of the gate relative to the VCO frequency and the chosen $gating\_range$. Every sub-module has 2 units of gating control and by designing such a progressive scheme, the entire hardware implementation is very resource optimized. The benefits of gating mode include not only "target-tracing" but SBR improvement by rejecting noise outside the selected gate. This is possible because of the amplification of the readout bandwidth by 32-fold for the events occurring within the gate. Please note that while using the gating functionality for only SBR improvement, it is possible to retain full TDC resolution of 14-bits by streaming out all the TDC bits. Alternatively, while using it progressively, the chosen $gating\_range$ can dictate the resolution within the gate.

As such, the modularity of the sensor allows every sub-module (16×8 pixels) to be configured for a unique gating range simultaneously.
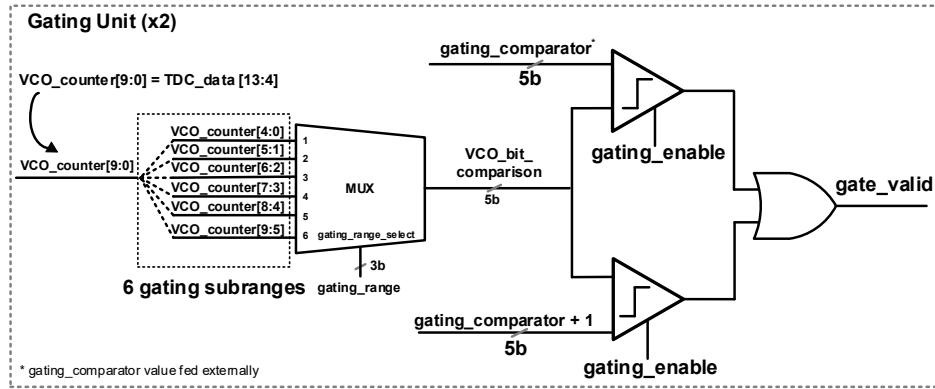
Figure 5.24 – Gating control unit per sub-module of 16×8 pixels.

5.7 Time-to-digital-converter (TDC)

The TDC core comprises an 8-stage pseudo-differential ring oscillator (RO) shown in Figure 5.25, which clocks a 10-bit gray counter (not shown in the figure). The oscillator phases are sampled using the SAFF (Figure 5.25e) and thereafter, decoded to 4-bit to combine with the 10-bit from the gray counter (binary-converted) to form a complete 14-bit TDC code. As discussed in Section 5.1, the TDC operates continuously. The resolution and range specification are maintained the same as in the previous design (see Section 3.2.2 of Chapter 3). Therefore, the RO nominally oscillates at 1 GHz, providing an LSB of about 60 ps, while dissipating 100 µW in the designed 22 nm CMOS technology.
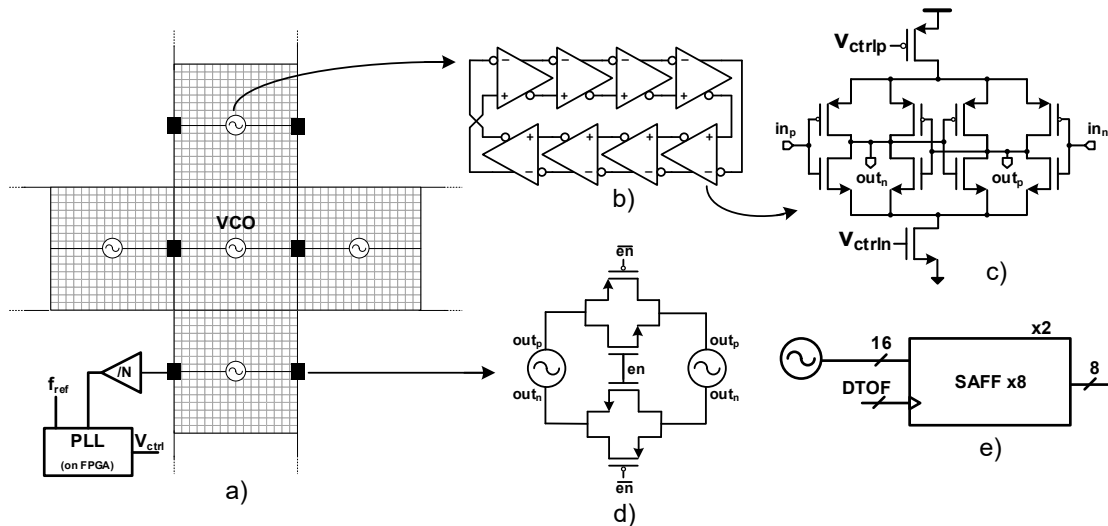
Figure 5.25 – (a) Mutual coupling of TDCs, (b) pseudo-differential ring oscillator, (c) schematic of a pseudo-differential stage, (d) Transmission gate element used for coupling and, (e) sense-amplifier flip-flop (SAFF) block.

Two sub-modules ($2 \times 16 \times 8$ pixels) share a TDC and the whole sensor (256×128 pixels) thus has

128 TDCs. Phase and frequency synchronization between multiple TDCs is obtained by coupling each VCO to its 4 neighbors, through a group of transmission gates (Figure 5.25a,d), thus reducing the overall phase noise and jitter by 21 dB with respect to a single equally-sized TDC. A $\div 1024$ version of a coupled phase of the VCO is used along with a reference signal to track for slow changes in frequency and PVT variations through an off-chip PLL (implemented on FPGA).

Figure 5.26 shows the the layout of a module ($2 \times 16 \times 8$ pixels) obtained after digital place and route where various building blocks can be identified. Except for the quenching circuit and the TDC which are custom-designed, all other blocks have been implemented in the digital flow. The overall assembly is also fully digital where analog custom cells are placed as MACROs. The whole sensor is implemented as an array of 16×8 units of the module shown in Figure 5.26.
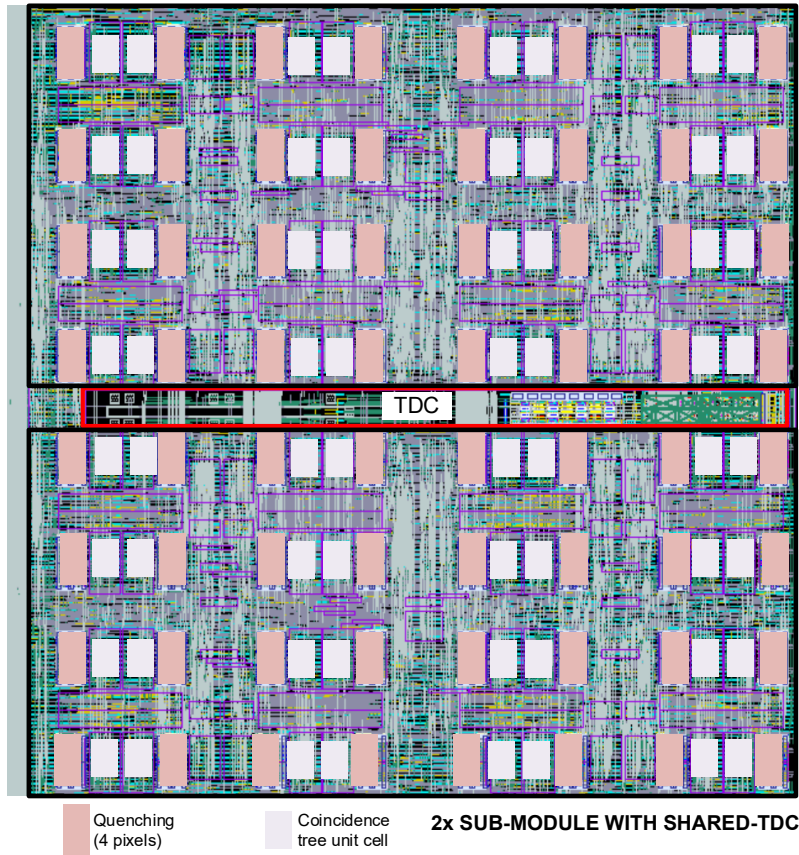
Figure 5.26 – Layout of a module consisting of two sub-modules.

5.8   Characterization results

The sensor described above was implemented in a 3D-stacked BSI technology where the SPADs were fabricated in a 45 nm CIS process on the top tier and the readout circuit in an ultra low-power 22 nm CMOS technology on the bottom tier. Figure 5.27 shows a micrograph of the sensor where,

the sensor measures 1.06×2.08 mm^2 with an active area of about 1.60 mm^2. The insets show the two 8×16-SPAD-pixel sub-modules and a detail of the BSI SPAD pixel. Please note that due to the BSI 3D-stacking, the bottom tier circuits are not visible in the micrograph.
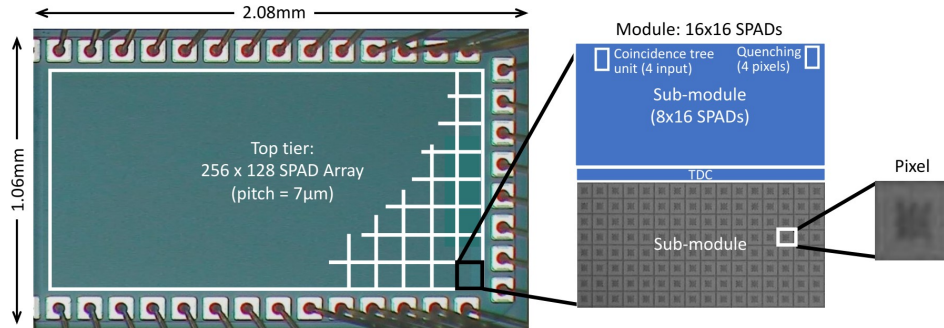
Figure 5.27 – Chip photomicrograph.

The complete system (Figure 5.28) comprises of 2 PCBs, where the mainboard hosts the sensor and some auxiliary components including LDOs and DACs for generating various biases and SMA connectors for monitoring and debug. The high voltage required for SPAD bias is drawn from an external power supply directly. The fabricated sensor die was packaged onto a ceramic QFP-120P package to be used along with a zero-insertion-force socket, as seen in Figure 5.28a.

(a) Front side of the camera board

(b) Back side of the camera board
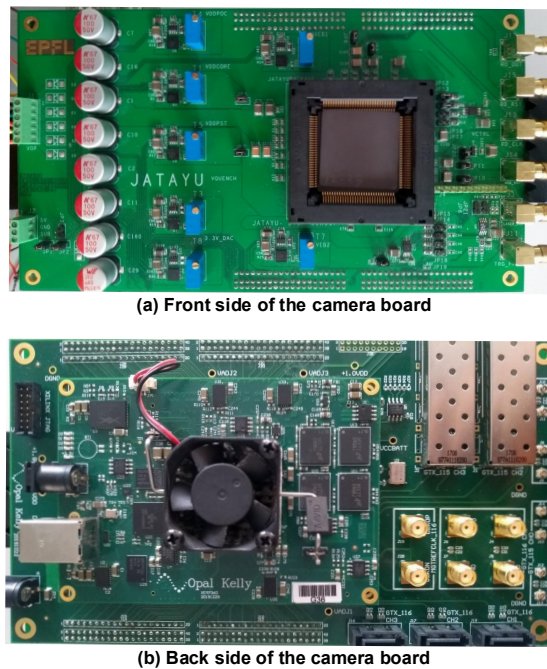
Figure 5.28 – Jatayu camera system- (a) Front side of the board hosting the sensor and (b) back side of the board hosting the FPGA.

In Figure 5.28b, the backside of the system is shown which consists of the motherboard, hosting the

FPGA board (XEM7360, Opal Kelly, USA) which includes a Kintex-7 FPGA (XC7K160T-1FFG676C, Xilinx, USA) and USB 3.0 interface. The characterization results presented through the following section are performed using a 780 nm, VisIR laser source (PicoQuant GmbH, Germany). The mean power of the laser for telemetry (ranging) measurements was 5 mW at 1 MHz repetition rate and for flash imaging experiments, 1.5 mW at 0.5 MHz repetition rate.

5.8.1 Single-point ranging and linearity

A module of 16×16 pixels in the sensor was used to measure telemetry of a target, where the TDC was operating at about 1 GHz. Two flat targets with reflectivity ranging from 10% to 50% were used under a 10 klux ambient light. Figure 5.29 shows the optical setup of the telemetry experiment conducted outdoor where a maximum range of 100 m was covered. An optical bandpass filter centered at 780 nm, with a passband of 20 nm was used in front of the sensor. The measurements were conducted with the coincidence window set at 2.2 ns.
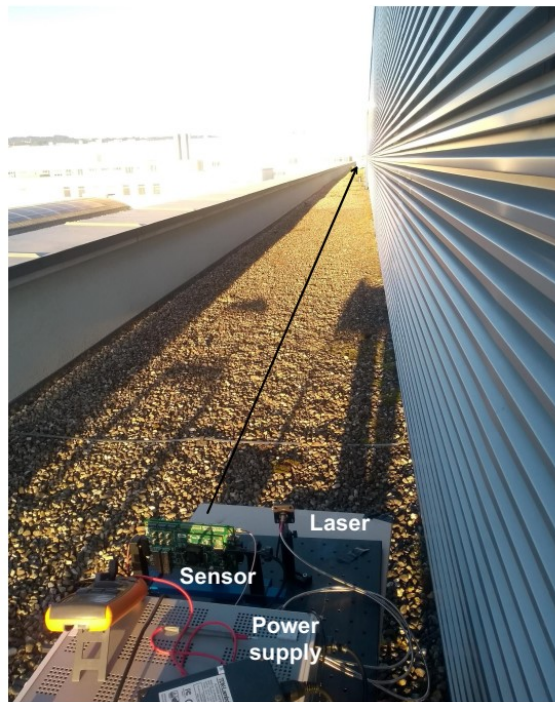
Figure 5.29 – Outdoor setup for telemetry measurement.

Figure 5.30 shows the measured distance vs. ground truth, along with the accuracy for 10 % and 50 % reflectivity targets. Maximum ranges of 100 m and 50 m were achieved for 50 % and 10 % reflectivity targets respectively. On comparing results in Figure 5.30c,d, telemetry was obtained without much loss of accuracy between 10% (fewer reflected photons) and 50% reflective targets. Maximum accuracy errors of 0.05 m and 0.07 m (0.05 % and 0.07 % non-linearity over 100 m) were achieved for 50% and 10 % reflective targets respectively.
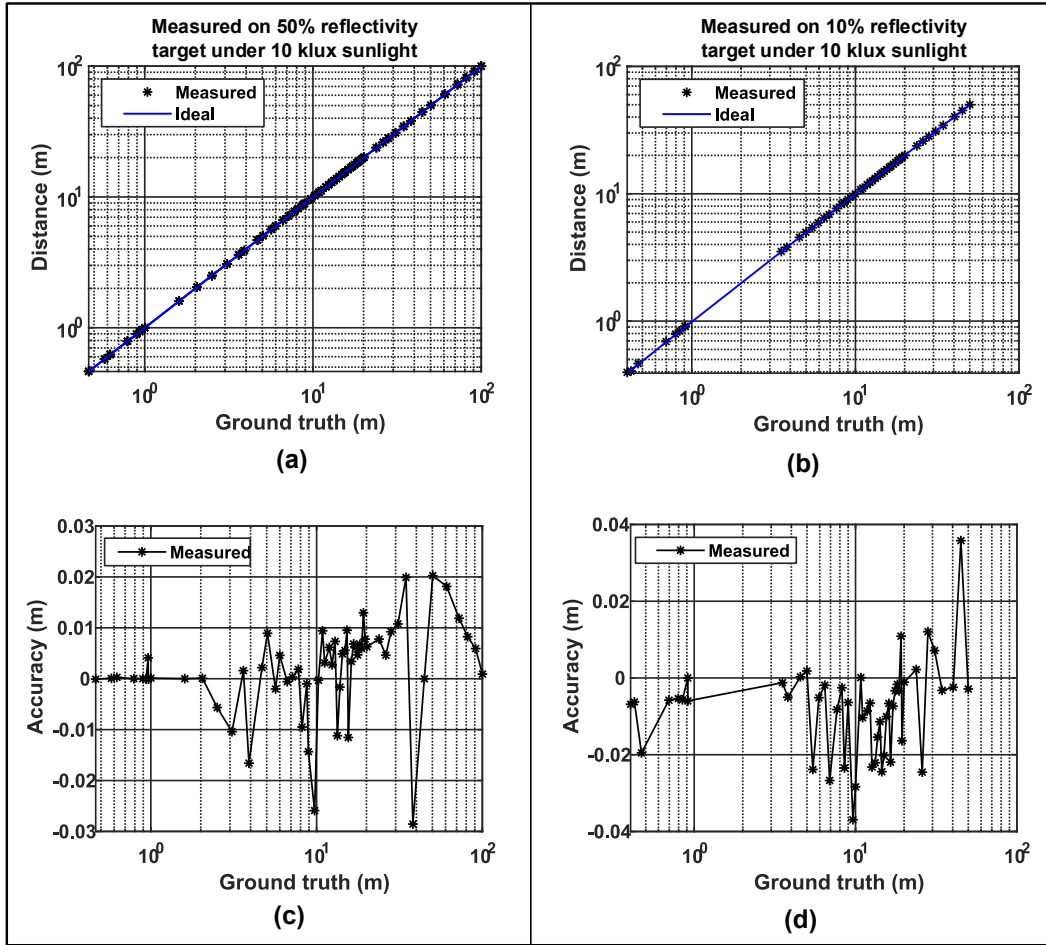
Figure 5.30 – Telemetry measurements- (a) and (c) Measured distance and accuracy vs ground truth on 50 % reflectivity target, (b) and (d) Measured distance and accuracy vs ground truth on 10 % reflectivity target.

5.8.2 Gating measurement

Gating was experimented with a 50 % reflective flat target situated at about 10 m. With the VCO frequency set at 976 MHz ($t_{bin,TDC} \approx 64$) ps, a target at 10 m corresponds to a TDC code of 1041. Figure 5.31a shows the histogram acquired without any gating where the target peak is distributed around 1041. Of course the histogram has uniform noise distributed across all the TDC codes as expected in the absence of any gating. In the above conditions, a SPAD detected on average 13,000 photons, of which roughly 300 were photons reflected by the target, in a 1.3 ms of exposure. This corresponds to a SBR of -32 dB. Following this, progressive gating was performed. In order to apply gating around the target situated at TDC code $\approx$ 1041 (14-bit), $gating\_range = 3$ is selected where, the bits, $VCO\_counter[6:2]$ are used for comparison (see Figure 5.24). Figure 5.31b shows the results after applying gating, where 12,494 valid events were registered, all of which belonging to the gating range. This is possible because the TDC bandwidth is amplified 32-fold

to the events within the gate and discarding events outside of the range. As a result, the SBR improved to -0.34 dB, an increase of 31.7 dB. One can also notice that gating behaves differently from coincidence detection, in that, any event irrespective of whether it is signal or noise is read out as long as it belongs to the valid gate. Gating is therefore useful when used in combination with coincidence detection, as it helps discriminate coincident photons more efficiently.
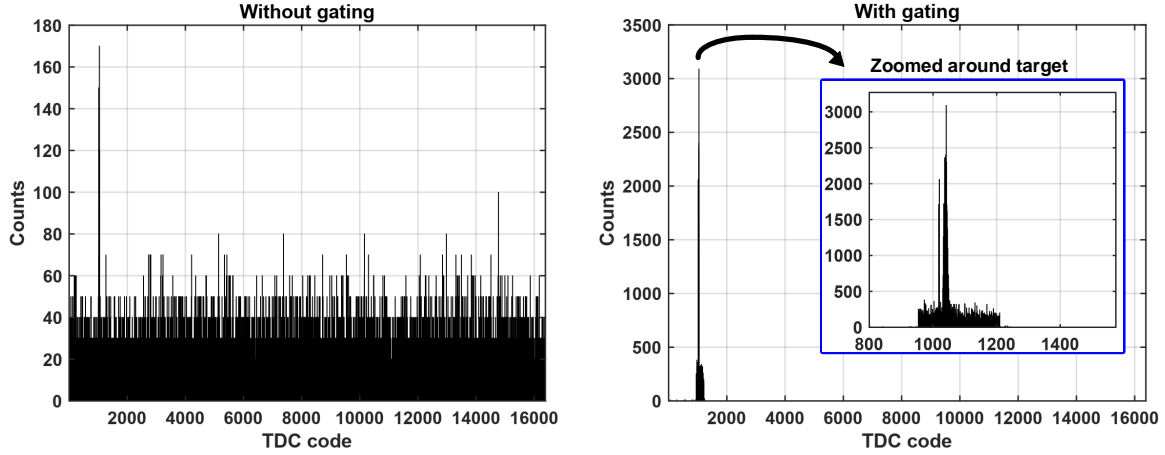
Figure 5.31 – Histograms of a target (a) without and (b) with gating.

5.8.3   Flash LiDAR measurements

The spatial resolution of 256×128 pixels can be exploited to perform flash LiDAR measurements. However, due to current firmware limitations (and, no particular sensor limitation), the image resolution is limited to 128×128. A full 3D reconstruction of various objects placed at short and medium range was conducted over 1 ms exposure. For this experiment, the laser was set to a mean power of 1.5 mW and 0.5 MHz repetition rate which was uniformly diffused onto a scene spanning up to 10m with a FOV of 2°, both horizontally and vertically. A F/2 lens was used to focus the scene onto the sensor. The whole sensor is operated with TDCs running continuously and mutually-coupled. Although, this guarantees synchronization, there were still small offsets for which calibration was performed. Calibration was performed against a flat target and the initial phases (TDC codes) were stored in a look-up table for all the 128 TDCs. After a measurement, the table was used to correct for the minimal offsets between multiple TDCs to acquire correct depth data. The depth map and a 3D/2D image of the targets are displayed in Figure 5.32. As can be seen, irrespective of the wide range in target reflectivities, the 3D image is reliably reconstructed. A cross-section of the scene across row 71 confirming the telemetry is also shown in 5.32d. Note that a few artifacts can be seen in the depth image, due to a combination of localized activity and cross-talk among certain SPADs. A new tapeout is currently being done in order to improve the performance.

While the sensor provides a maximum of 111 Mtimestamps/s, the current firmware is limited and therefore, most processing was done offline on MATLAB software which limited the achievable

frame rate. However, an improved firmware with histogram and processing on FPGA in future, should allow acquiring video rate 3D images without any issue.
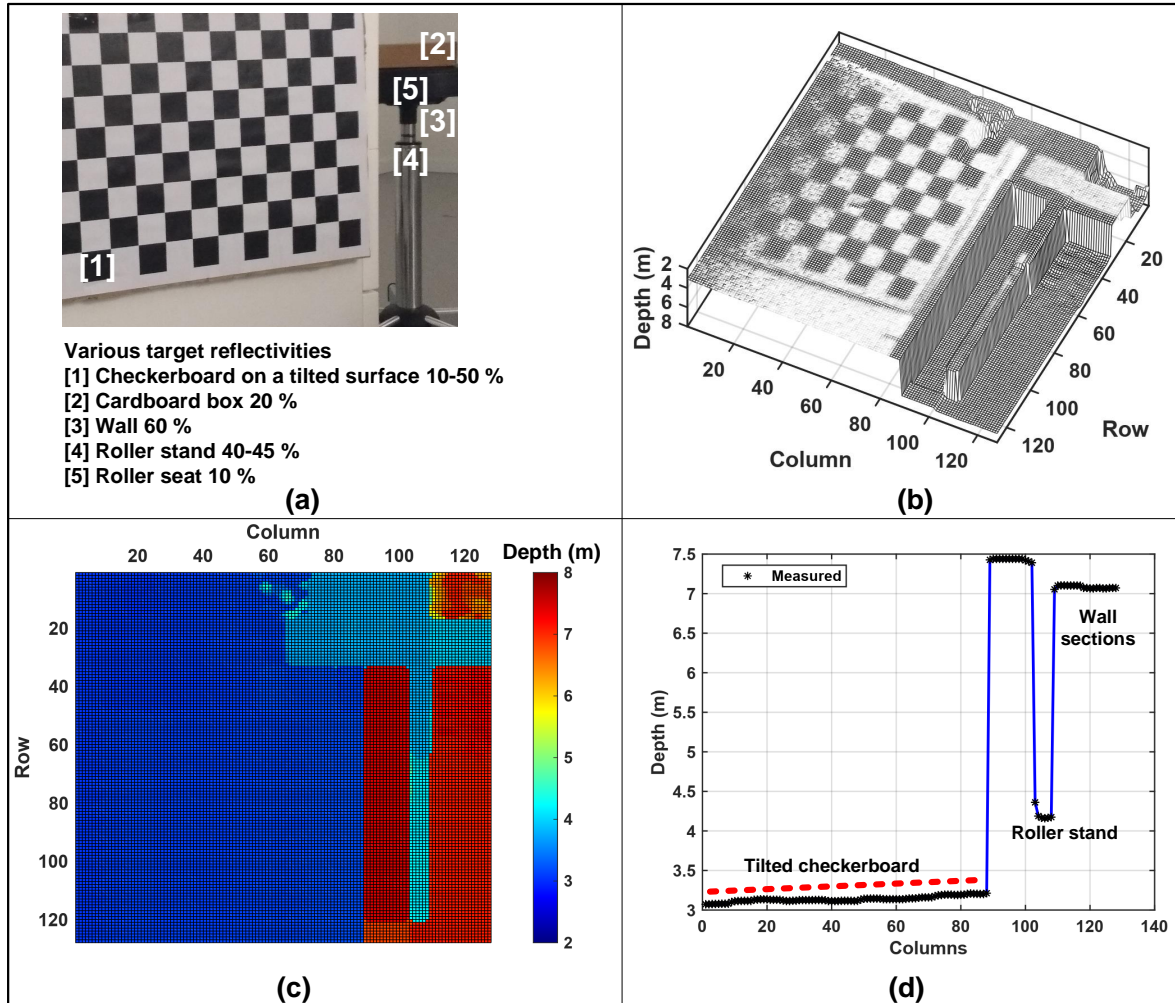
Figure 5.32 – 3D imaging of multiple targets with different reflectivity- (a) Photograph of the scene, (b) Superimposed depth and intensity image, (c) color-coded depth data and, (d) Cross-section of depth across row 71.

5.9 Power consumption

Figure 5.33 shows the distribution of power consumption from various blocks. Under a timing throughput of 111 Mtimestamps/s, the total power consumption stands at 51.9 mW (excluding the FPGA board) where the dominant contribution is from the always-on, mutually-coupled TDCs, which contribute to 25 % of the total power where every TDC consumes about 0.1 mW at an oscillation frequency of 1 GHz. However, this power is mostly constant and uniform over varying photon activity, as expected. Readout and consequently, IO power consumption follow with a contribution around 20 % of the total power.
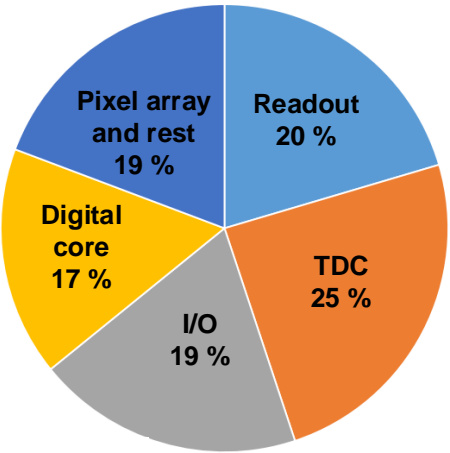
Figure 5.33 – Pie chart indicating power consumption of various blocks

5.10   State-of-the-art comparison

Table 5.1 shows the performance comparison of the proposed sensor with other state-of-the-art DTOF sensors where it can be seen that the presented sensor is the first ever reported SPAD-based DTOF sensor in a 45/22 nm 3D-stacked technology with a pixel pitch of 7 µm. The presented sensor measures ranges up to 100 m with accuracy under 0.07 m on 10-50 % target reflectivity under a background light of 10 klux. The sensor offers the highest image resolution (128×128) when compared to other DTOF, TDC-based sensors under flash laser projection. With a timing throughput of 111 Mtimestamps/s, the total power consumption measures under 52 mW, thus, being the lowest among the compared state-of-the-art DTOF sensors (see Table 5.1).

Table 5.1 – Performance comparison of state-of-the-art DTOF sensors (2020)

Parameter	Unit	This Work	[17]	[18]	[19]	[20]	[21]
Technology	–	45/22 nm SPAD CMOS	40/90 nm SPAD CMOS	45/65 nm SPAD CMOS	150 nm SPAD CMOS	180 nm SPAD CMOS	65 nm VAPD CMOS
Architecture	–	Coincidence-detection, shared TDC	Event-driven, shared multi-event TDC	Always-on, shared TDC	Event-driven, single TDC per pixel	Column-wise shared TDC	ADC & subrange syntheses
Format (HxV)		256×128	$256 \times 256/64 \times 64$	16×8^a	64×64	340×96	1200×900
Sensor characteristics							
Pixel pitch	μm	7	9.2/38.4	19.8	60	25	6
Pixel fill factor	%	N/A^e	51	31.3	26.5	70	N/A
SPAD DCR@V_E	cps{μm^2}	N/A^e	20 @ 1.5 V	5.3k @ 2.5 V	6.8k at 3 V	2.65k @ 3.3 V	N/A
PDP @ V_E	%	N/A^e	23 @ 3 V	21 @ 2.5 V	20 @ 3V	N/A	N/A
TDC depth	bit	14	14/4	14	16/15	12	7-b ADC
TDC resolution	ps	60	35/560	60/320	250/20000	208	-
TDC linearity — DNL [LSB]		+0.05/-0.05	+0.05/-0.05	+0.8/-0.7	+0.3/-0.25^b	+0.52/-0.52	-
TDC linearity — INL [LSB]		+1.1/-1.1	+0.1/-0.08	+3.6/-0.4	+1.2/-0.8^b	+0.73/-0.49	-
Laser projection	-	Flash	Flash	Scanning	Flash	Scanning	Flash
Image resolution	-	128×128	64×64	256×256^a	64×64	202×96	N/A
FOV	Deg	2×2	1.2×1.2	N/A	N/A	170×4.5	N/A
Repetition rate	MHz	1/0.5^d	1.9	1	N/A	0.133	0.05
Illumination power (Mean)	mW	5/1.5^d	1.8	6	N/A	21	N/A
Illumination wavelength	nm	780	671	532	470	870	N/A
Distance range	m	100	50	$150 - 300$	$367 - 5862^c$	128	250
Accuracy	m	0.07	0.17	$0.07 - 0.8$	$1.5 - 35$	0.37	1.5
Accuracy	%	0.07	0.34	$0.3 - 0.4$	$0.37 - 1.9$	0.37	0.6
Background light	-	$8 - 10$ klux	1 klux	N/A	100 MPh/pixel/s	80 klux	N/A
Target reflectivity	-	10 % – white	white	white	50 %	9 %	N/A
Power consumption — TDC (mW)		0.1	N/A	0.5-0.1	N/A	N/A	-
Power consumption — Total (mW)		51.9	77.6	N/A	93.5	530	2500

^{a} Up to 256×256 resolution achieved by flexible scanning system. ^{b} Measured over 5% of the total range. ^{c} Emulated results with optical fiber.
^{d} Telemetry/Imaging. ^{e} Foundry-confidential data.

5.11 Conclusions

In this chapter, a 256×128 Flash LiDAR sensor was presented. The concepts modeled and analyzed in the previous chapter (Chapter 4) were successfully implemented and tested. The design was implemented in a 3-D-stacked technology, featuring a BSI SPAD array on the top tier in 45 nm CIS process, connected to a readout and processing circuit on the bottom tier in 22 nm CMOS technology, with an overall power consumption under 52 mW. A robust synchronization method was also proposed and implemented where TDCs were mutually-coupled and thereby, offered a better timing jitter (and phase noise). 7-level coincidence detection along with progressive gating were proposed for the first time, as ways to mitigate low SBR due to high background noise, where up to 31.7 dB SBR improvement was achieved with progressive gating. The sensor measured ranges up to 100 m with accuracy error lower than 0.07 % for targets between 10–50 % reflectivity. Flash LiDAR experiments were conducted over a scene spanning from short to medium range with targets of various reflectivity and 3-D images were successfully reconstructed. Throughout the experiments, a near-IR 780 nm laser source was used, migrating from the previously used 532 nm (Chapter 3), towards a LiDAR-friendly choice [22].

REFERENCES

[1] P. Padmanabhan, C. Zhang, M. Cazzaniga, B. C. Efe, A. Ximenes, M. J. Lee, and E. Charbon, "A 256×128 3D-stacked (45nm) SPAD FLASH LiDAR with 7-level coincidence detection and progressive gating for 100m range and 10klux background light," *To appear in 2021 IEEE International Solid-State Circuits Conference-(ISSCC)*, 2021.

[2] A. Ronchini Ximenes, P. Padmanabhan, and E. Charbon, "Mutually coupled time-to-digital converters (tdcs) for direct time-of-flight (dtof) image sensors," *Sensors*, vol. 18, no. 10, p. 3413, 2018.

[3] R. C. van de Beek, E. A. Klumperink, C. S. Vaucher, and B. Nauta, "Low-jitter clock multiplication: A comparison between plls and dlls," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 49, no. 8, pp. 555–566, 2002.

[4] H. Mizuno and K. Ishibashi, "A noise-immune ghz-clock distribution scheme using synchronous distributed oscillators," in *1998 IEEE International Solid-State Circuits Conference. Digest of Technical Papers, ISSCC. First Edition (Cat. No. 98CH36156)*, pp. 404–405, IEEE, 1998.

[5] F. O'Mahony, C. P. Yue, M. A. Horowitz, and S. S. Wong, "A 10-ghz global clock distribution using coupled standing-wave oscillators," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 11, pp. 1813–1820, 2003.

[6] J.-C. Chien and L.-H. Lu, "Analysis and design of wideband injection-locked ring oscillators with multiple-input injection," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 9, pp. 1906–1915, 2007.

[7] C. Verhoeven, "A high-frequency electronically tunable quadrature oscillator," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 7, pp. 1097–1100, 1992.

[8] R. Adler, "A study of locking phenomena in oscillators," *Proceedings of the IRE*, vol. 34, no. 6, pp. 351–357, 1946.

[9] B. Razavi, "A study of injection locking and pulling in oscillators," *IEEE journal of solid-state circuits*, vol. 39, no. 9, pp. 1415–1424, 2004.

[10] A. Demir, A. Mehrotra, and J. Roychowdhury, "Phase noise in oscillators: A unifying theory and numerical methods for characterization," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, no. 5, pp. 655–674, 2000.

[11] A. Demir, "Floquet theory and non-linear perturbation analysis for oscillators with differential-algebraic equations," *International journal of circuit theory and applications*, vol. 28, no. 2, pp. 163–185, 2000.

[12] P. R. Gray, P. Hurst, R. G. Meyer, and S. Lewis, *Analysis and design of analog integrated circuits*. Wiley, 2001.

REFERENCES

[13] P. Kinget, "Integrated ghz voltage controlled oscillators," in *Analog circuit design*, pp. 353–381, Springer, 1999.

[14] A. Hajimiri, S. Limotyrakis, and T. H. Lee, "Jitter and phase noise in ring oscillators," *IEEE Journal of Solid-state circuits*, vol. 34, no. 6, pp. 790–804, 1999.

[15] H.-C. Chang, X. Cao, U. K. Mishra, and R. A. York, "Phase noise in coupled oscillators: Theory and experiment," *IEEE Transactions on Microwave Theory and Techniques*, vol. 45, no. 5, pp. 604–615, 1997.

[16] M.-J. Lee, A. R. Ximenes, P. Padmanabhan, T.-J. Wang, K.-C. Huang, Y. Yamashita, D.-N. Yaung, and E. Charbon, "High-performance back-illuminated three-dimensional stacked single-photon avalanche diode implemented in 45-nm cmos technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, pp. 1–9, 2018.

[17] R. K. Henderson, N. Johnston, S. W. Hutchings, I. Gyongy, T. Al Abbas, N. Dutton, M. Tyler, S. Chan, and J. Leach, "5.7 a 256 × 256 40nm/90nm cmos 3d-stacked 120db dynamic-range reconfigurable time-resolved spad imager," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, pp. 106–108, IEEE, 2019.

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

[19] M. Perenzoni, D. Perenzoni, and D. Stoppa, "A 64 × 64-pixels digital silicon photomultiplier direct tof sensor with 100-mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6 km for spacecraft navigation and landing," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 151–160, 2016.

[20] C. Niclass, M. Soga, H. Matsubara, S. Kato, and M. Kagami, "A 100-m Range 10-Frame/s 340×96-Pixel Time-of-Flight Depth Sensor in 0.18-μm CMOS," vol. 48, no. 2, pp. 559–572, 2013.

[21] Y. Hirose, S. Koyama, T. Okino, A. Inoue, S. Saito, Y. Nose, M. Ishii, S. Yamahira, S. Kasuga, M. Mori, *et al.*, "5.6 a 400× 400-pixel 6μm-pitch vertical avalanche photodiodes cmos image sensor based on 150ps-fast capacitive relaxation quenching in geiger mode for synthesis of arbitrary gain images," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, pp. 104–106, IEEE, 2019.

[22] ISO-9845-1, "Solar energy—reflectance solar spectral irradiance at the ground at different receiving conditions—part 1: direct normal and hemispherical solar irradiance for air mass 1.5," 1992.

6 Conclusions and future work

6.1 Conclusions

The main goal of this thesis was to develop direct time-of-flight sensors for LiDAR applications while advancing the art from previous sensor solutions but also keeping the application requirements at the forefront. To this end, two DTOF sensors were designed within the scope of this thesis focusing on two main application challenges- , i.e. the suppression of intereference and of background noise.

The initial phase of this thesis involved exploration of hybrid detectors from III-V family, due to their inherent out-of-band solar rejection, beneficial for higher signal-background-noise-ratio (SBR). Gallium nitride based APDs developed at the Jet Propulsion Laboratory were experimented as primary candidates where a CMOS front-end readout circuit was designed for their combined use (Chapter 2). A simple, yet effective solution based on capacitive transimpedance amplifiers (CTIAs) was proposed and implemented, achieving optical gains on the order of 10^3 with UV sensitivity. Soon after extensive characterization, it was realized that GaN APDs needed further process development so that they can operate in the Geiger mode, enabling single-photon detection. As a result of this, the GaN-based research was not pursued further and Si-based SPADs were opted for their single-photon detection as well as compatibility with mass-produced CMOS technology. In particular, 3D-stacked back-illuminated SPADs are discussed in Chapter 2, as promising candidates for DTOF sensors. In addition to improving the fill-factor compared to monolithic implementations, the 45 nm BSI SPADs achieved a NIR PDP reaching up to 10 % with a timing jitter of 107.7 ps FWHM, favoring them for the targeted LiDAR applications.

Benefiting from the 3D-stacking process, a first DTOF sensor was designed with the aforementioned SPADs on the top tier in the 45 nm CIS process, hybrid-bonded to the readout electronics on the bottom tier, which was fabricated in a 65 nm CMOS technology. At the sensor level, resource sharing between pixels was extensively evaluated, as a result of which, a modular, power-efficient and shared-TDC based architecture was presented in Chapter 3. A decision-tree was implemented to manage multiple pixel events while also preserving the origin of the event by generating a pixel ID.

A quality timing reference is a key element in DTOF imagers which rely on the timestamps generated from TDCs to measure depth. The next major contribution in the sensor design was in this direction where a robust timing solution based on multiple shared TDCs was proposed. When mutually-coupled, the proposed solution significantly improved timing jitter (and the overall phase noise). A phase macro model was developed on MATLAB to thoroughly analyze this phenomenon, following which the concept was successfully verified in silicon. In Chapter 5, the performance of mutual coupling was presented on 64 TDCs. The same concept was extended in the second sensor, *Jatatyu*, where all the 128 TDCs were mutually coupled, easing the overall calibration process during the TOF measurement.

The first DTOF sensor design was tested successfully in a LiDAR scenario, where up to 300 m telemetry was measured. However, as discussed in Chapter 3, 4, the sensor was not prepared to cope with high background noise due to the lack of any on-chip noise filtering. Furthermore, due to smaller spatial resolution (8×16), the sensor was only tested in a scanning LiDAR setup. With an aim to mitigate the fundamental challenge of ambient light in LiDARs while retaining the modularity of the first design, a second DTOF sensor was proposed.

The actual IC design of the sensor evolved through systematic analysis and simulation carried on a probabilistic model developed on MATLAB, the next major contribution of this thesis. This was presented in Chapter 4. A new architecture was modeled with primary focus on noise suppression and wide dynamic range 3D imaging, required in Flash LiDAR. Simulations on the model allowed definition of key target specifications for the actual chip design.

The second DTOF sensor, *Jatayu*, was designed based on the above model, with a coincidence-based architecture, achieving up to 10 klux background noise suppression. Resource sharing between pixels was similar to the first sensor, however, a new combination tree, based on coincidence was proposed. Up to 7-level coincidence was enabled where every module in the sensor could be configured independently. The proposed coincidence tree had a non-blocking nature, different from the winner-take-all approach in the decision tree based design. Up to 8 events could be processed where their pixel ID, coarse time stamp as well as photon rank were locally stored. Multiple coarse timestamping also increased the timing throughput per module by 8x.

Another modular feature, called, progressive gating was also implemented which allowed range-selective imaging in addition to improving the SBR. Gating-based SBR improvement is particularly useful in long distance ranging where returning signal events from the target may also be low in counts and coincidence detection may not add significant value. Instead, by acting on the TDC bandwidth and restricting it to a particular (chosen) gate around the target, SBR is enhanced electrically by approximately 32x. This was demonstrated in Chapter 5.

It can be concluded that the second sensor, *Jatayu*, has significantly advanced the previous sensor in terms of background noise and flash LiDAR operation, however, it still has a lot of scope for performance improvement. There were some important observations made during measurement requiring further work at both process and sensor level. While it is not within the timeline of this

thesis to examine this further, however, a number of useful recommendations for future work is provided in the next section.

Also, as seen in Chapter 1, there are numerous ongoing challenges as well as scope for alternative solutions for LiDAR.

6.2 Recommendations for future work

Sensor-level improvements in *Jatayu* – *Jatayu*, the most recent of both the DTOF sensors in this thesis, had new concepts implemented for mitigating high background illumination challenge. However, as seen in Chapter 5, the sensor could only be tested up to 10 klux ambient light while the initial target 100 klux, a typical outdoor scenario of bright sunlight.

The primary reason of not being able to test at higher incoming photon activity was due to the performance of the SPAD. The 7 µm N+/P-well SPADs, differently from the 19.8 µm P+/N-well SPADs in the first sensor, showed an unexpectedly poor recharge time on the order of a few µs. Consequently, the photon throughput per pixel was significantly lower due to a lower dynamic range. Even if multiple pixels could be combined to operate in an SiPM-like fashion, it is still desirable to have a lower dead time so as to foster more detections (of course, accounting for afterpulsing which was considered negligible in this work). While this issue of higher dead time did not prohibit the testing of sensor functionality, it however limited the achievable performance from the overall sensor.

Addressing this, a new tape-out of the top tier (SPAD layer) is currently being planned with improvements at the process-level. With the improved version of the existing chip, it should be possible to reach higher throughput while also opening up new pathways for various experiments. With support from an improved firmware design, some of the planned experiments include:

- u progressive gating feature for dynamic target tracing as well as imaging in the presence of scattering medium such as fog, cloud etc.,
- performing histogramming on FPGA to support reconstruction of video rate 3D images in a flash LiDAR setup,
- utilizing 7-level coincidence to perform photon-activity dependent imaging, concept elaborated in Chapter 4 (see Figure 4.13).

In addition to the above recommendations to the current work, a parallel future work is planned at the modeling level. The current analytical model is limited to Lambertian targets. A significant addition to this model would include modeling the nature of targets (diffusive, scattering or alike) to more accurately emulate the photon interaction. Similarly, light propagation path should also be modeled to include fog or cloud like phenomenon. Finally, analytical models should eventually transition towards ray-tracing models, which will help understand a LiDAR system more realistically and consequently, enable more accurate definition for the IC design of the sensor.

Next-generation sensor development – As seen in Chapter 1, high data rate is a perpetual challenge in SPAD-based sensors capable of producing large volume of data with a bottleneck of limited I/O bandwidth. Therefore, the future generation of sensors should include on-chip processing including minimum of a histogramming functionality. Combined with noise-filtering such as coincidence detection and gating, this can significantly reduce the burden on I/O bandwidth while still maintaining high frame rate imaging under high ambient light.

In general, as imagers scale in array size, the addition of any additional on-chip processing comes at the cost of increased silicon area (and power). To this end, technology wise, 3D-stacked implementations will continue to favor packing of complex electronics required in high resolution imagers.

Artificial intelligence (AI) with LiDAR – The vast amount of data being generated in SPAD-based imagers should be exploited. Apart from on-chip processing mentioned above, machine learning algorithms should become active part of sensing systems. AI-based algorithms cannot only help mitigate the data challenge but also boost the overall detection process by performing object classification (and identification) and feature extraction in 3D depth maps. AI-based methods can significantly improve automation of the detection process, directly impacting various consumer and automotive LiDAR applications. The future will demand LiDAR systems with high fidelity where AI will play a crucial role in target-aware detection.

In summary, a LiDAR system development should be comprehensive and inclusive of other important elements in the chain including illuminator and optics in addition to sensor design, which has been the focus of this thesis. Together, they can then provide a realistic understanding of the system.
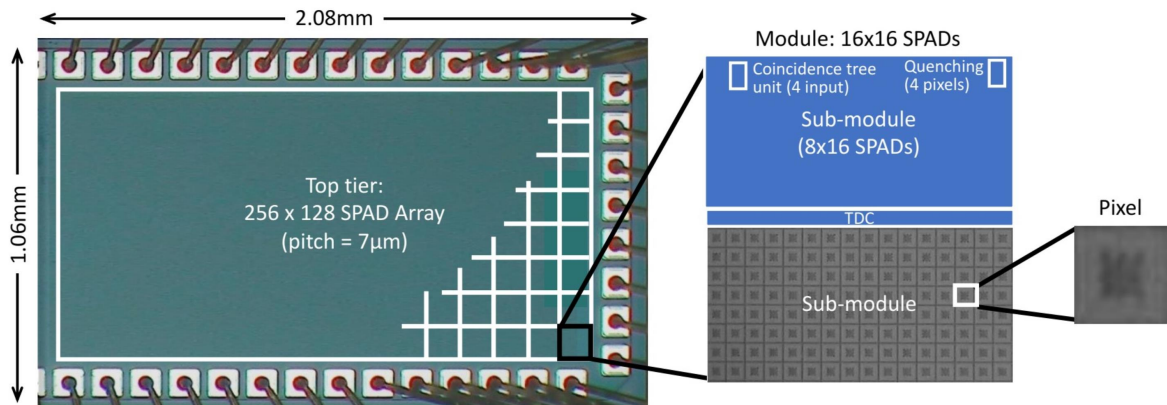
Chip gallery
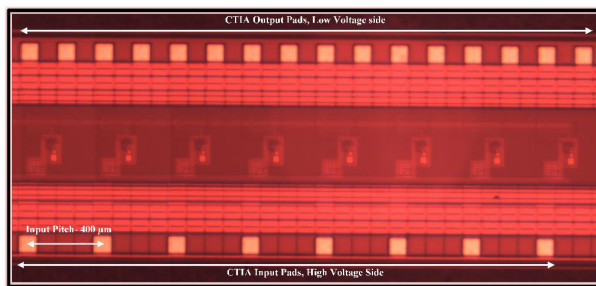
(a) CMOS front-end circuit – capacitive transimpedance amplifier (8x), Chapter 2
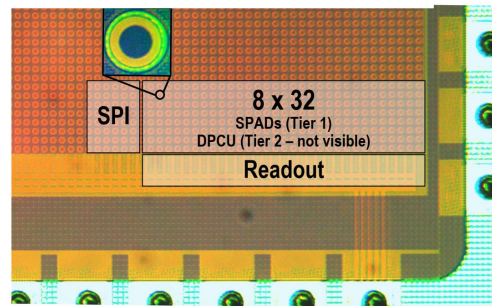
(b) 45 nm/ 65 nm 3D stacked DTOF sensor, Chapter 3

(c) 45 nm/ 22 nm 3D stacked DTOF sensor, Chapter 5

Photomicrographs of various chips in this thesis.

List of publications and awards

JOURNAL ARTICLES

1. **P. Padmanabhan**, B. Hancock, S. Nikzad, L. Bell, K. Kroep and E. Charbon : "A Hybrid Readout Solution for GaN-Based Detectors Using CMOS Technology." Sensors. 2018-02-03.

2. A. R. Ximenes, **P. Padmanabhan** and E. Charbon. "Mutually Coupled Time-to-Digital Converters (TDCs) for Direct Time-of-Flight (dTOF) Image Sensors ." Sensors 2018, 18, 3413. [*Shared first authorship*].

3. A. R. Ximenes, **P. Padmanabhan**, M.-J. Lee, Y. Yamashita, D. N. Yaung and E. Charbon. (2019). "A Modular, Direct Time-of-Flight Depth Sensor in 45/65-nm 3-D-Stacked CMOS Technology." IEEE Journal of Solid-State Circuits, 54(11), 3203-3214. [*Shared first authorship*].

4. **P. Padmanabhan**, C. Zhang, and E. Charbon. (2019). "Modeling and analysis of a direct time-of-flight sensor architecture for LiDAR applications." Sensors, 19(24), 5464.

5. M.-J. Lee, A. R. Ximenes, **P. Padmanabhan**, T.-J. Wang, K.-C. Huang, Y. Yamashita, D. N. Yaung, and E. Charbon. "High-Performance Back-Illuminated Three-Dimensional Stacked Single-Photon Avalanche Diode Implemented in 45-nm CMOS Technology." IEEE Journal of Selected Topics in Quantum Electronics. 2018-04-16.

CONFERENCES

1. **P. Padmanabhan**; B. Hancock; S. Nikzad; L. Bell; K. Kroep and E. Charbon : "A CMOS Front-end for GaN-based UV Imaging." 2017 International Image Sensor Workshop, Hiroshima, Japan, May 30- June 2, 2017.

2. Lee, M-J., A. R. Ximenes, **P. Padmanabhan**, T. J. Wang, K. C. Huang, Y. Yamashita, D. N. Yaung, and E. Charbon. "A back-illuminated 3D-stacked single-photon avalanche diode in 45nm CMOS technology." In 2017 IEEE International Electron Devices Meeting (IEDM), pp. 16-6. IEEE, 2017.

3. A. R. Ximenes, **P. Padmanabhan**, M.-J. Lee, Y. Yamashita, D. N. Yaung and E. Charbon."A 256×256 45/65nm 3D-stacked SPAD-based direct TOF image sensor for LiDAR applications

with optical polar modulation for up to 18.6dB interference suppression." 2018 IEEE International Solid - State Circuits Conference - (ISSCC), San Francisco, CA, USA, February 11-15, 2018. p. 96-98. [*Shared first authorship*].

4. **P. Padmanabhan**, C. Zhang and E. Charbon, "Analysis of a modular SPAD-based direct time-of-flight depth sensor architecture for wide dynamic range scenes in a LiDAR system." 2019 International Image Sensor Workshop, Snowbird, Utah, USA, June 23-27, 2019.

5. **P. Padmanabhan**, C. Zhang, M. Cazzaniga, B. Efe, A. R. Ximenes, M. J. Lee and E. Charbon. "A 256×128 3D-Stacked (45nm) SPAD FLASH LiDAR with 7-Level Coincidence Detection and Progressive Gating for 100m Range and 10klux Background Light." To appear in 2021 IEEE International Solid - State Circuits Conference - (ISSCC), San Francisco, CA, USA, February 14-18, 2021.

PATENTS

1. A. R. Ximenes, **P. Padmanabhan**, E. Charbon, Photon detecting 3d imaging sensor device. WO2019154513. 2019.

2. A. R. Ximenes, **P. Padmanabhan**, E. Charbon, Oscillator arrangement for time-to-digital converter for large array of time-of-flight image sensor, U.S. Patent Application No. 15/941,411.

3. **P. Padmanabhan**, C. Zhang and E. Charbon; Direct time-of-flight depth sensor architecture and method for operating of such a sensor (Application PCT/EP2019/066478, 21 June 2019).

WORKSHOP

1. C. Bruschini, **P. Padmanabhan** and E. Charbon: "LiDAR and 3D-stacked technologies for consumer, automotive and biomedical applications." In 2019 Image Sensors Europe, London UK.

AWARDS

1. **Best Student Paper Award**, International Image Sensor Workshop, June 23-27, 2019, Snowbird, USA.

2. **ISSCC Student Travel Grant Award (STGA)**, International Solid-State Circuits Conference (ISSCC), 2018, San Francisco, USA.

3. **Best Poster Award**, International Image Sensor Workshop, May 30- June 2, 2017, Hiroshima, Japan.

About the author

Preethi Padmanabhan was born in 1992, in Chennai, India. She received her MSc degree (*cum laude* and Honors) in Electrical Engineering from TU Delft in the Netherlands, in August 2016. From July to September 2015, she was a student researcher at NASA's Jet Propulsion Laboratory (JPL) in Pasadena, USA, where she designed a CMOS readout circuit for UV avalanche photodiodes. From October 2015 to August 2016, she worked on her MSc thesis designing a CMOS time-to-digital converter array for application in time-of-flight (TOF) sensors. From November 2016 to October 2020, she was working towards her PhD degree at the Advanced Quantum Architecture (AQUA) Lab at EPFL, Switzerland. Her current research interests include modeling and design of integrated circuits (analog and digital) for TOF image sensors in light detection and ranging (LiDAR) applications.