SPAD-based Light Detection and Ranging for 3D imaging

Receiver operation and in-pixel TDC design for automotive application

Dongjin Son

In NXP Semiconductors

TU Delft

SPAD-based Light Detection and Ranging for 3D imaging

Receiver operation and in-pixel TDC design for automotive application

by

Dongjin Son

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on August 23, 2022.

Student number: 4350138
Project duration: May 15, 2019 – December 15, 2020
Thesis committee: Dr. ir. M. Bolatkale, TU Delft, supervisor
Prof. Dr. P.J. French, TU Delft

This thesis is confidential and cannot be made public until August 23, 2022

TU Delft

Preface

I wish to take this opportunity to express my sincerest gratitude to a group of precious and dear people.

My foremost gratitude goes undoubtedly to Dr.Muhammed Bolatkale. His invaluable academic advices have indisputably played a significant role in this work. Additionally, the patience, care, as well as personal advices he has provided over numerous occasions have encouraged, nurtured, and comforted me. Working under his guidance have been a great pleasure and will remain unforgotten for the rest of my life.

I would like to further thank NXP's system engineer, Maxim Kulesh. His professionalism as well as interest in the subject matter of this work have always greatly inspired and engaged me. His contribution in this work, particularly in system level study, shall be noted and I am very grateful. It was delightful to work with a man of brilliance that he is.

Furthermore, I wish to express my sincere gratitude to my loving family and friends. The unconditional love, heart-warming affection, and endless support they have shown have comforted and encouraged to pursue my dream and I am forever indebted.

I thank, last but not least, NXP Semiconductors for the amazing experience and the colleagues for the insightful and intriguing conversations.

I hope the readers find this work insightful and interesting.

Dongjin Son
Düsseldorf, June 2021

Contents

1 Introduction

In recent years, automotive industry has shown great interest in technologies to enable fully autonomous driving system. In particular, a light detection and ranging (LiDAR) has attracted special attention owing to its ability to produce 3D images with milimetric precisions on the flight. The LiDAR system transmits pulses which then reflects on the object and arrives at the receiver, often an array of single photon avalanche diodes (SPADs). The system processes and calculates the distance to the object based on the travel time of the pulse.

However, LiDAR system, particularly SPAD-based system, shows limited detection range under strong background noise using the conventional receiver operations. Moreover, a wide field-of-view (FoV) with high resolution requires a large pixel array where each pixel has narrow FoV which results in high throughput and limited pixel area, respectively. Furthermore, the power consumption of the scaled system must be low. Therefore, there is a need to reconsider the system level design choices of a SPAD-based LiDAR system for automotive applications.

The goal of this work is to develop a SPAD-based LiDAR system for automotive applications. In particular, the LiDAR system employs direct time-of-flight (dTOF) using time correlated single photon counting (TCSPC) principle to reconstruct target reflected pulses. More particularly, in the LiDAR system, different receiver operational methods under strong background noise are studied. The methods target to extend an object detection range. Accordingly, Matlab simulation models are developed to compare and verify the performances of the methods.

It is further the goal of this work to develop a time-to-digital converter (TDC) for the SPAD-based LiDAR system. In particular, the TDC is located in-pixel to enable in-pixel histogramming. More particularly, a ring oscillator with various delay cell topologies are studied. Each delay cell targets to consume the least power while meeting the linearity requirement when mismatch-induced jitter dominates. Cadence Virtuoso circuit simulation tool implements delay cells in 28nm to compare the performances of each delay cell topology.

The remaining of the thesis is organised as the following:

Chapter 2 provides an overview of single-photon avalanche diode (SPAD) and achieved state-of-the-art performances. Moreover, minimum building blocks which enable photon detections using SPAD is presented. Furthermore, the recent development in SPAD array implementation is introduced and its implication in a SPAD-based LiDAR system is shortly discussed.

Chapter 3 introduces principle and assumptions of SPAD-based LiDAR system using time-correlated single photon counting method. The limitations thereof under strong background noise are identified and state-of-the-art receiver operations are presented. A new receiver operation is proposed and its mathematical derivations are explained. Accordingly, Matlab simulation models simulate realistic outdoor environment and reconstruct target-reflected pulses. The performances of simulated receiver operations are compared and concluded.

Chapter 4 presents literature review of state-of-the-art time-to-digital converters (TDC) and proposes a new normalized figure-of-merits. Then, given the external PLL, a ring oscillator-based TDC is designed with delay topologies. Cadence Virtuoso circuit simulation tool implements and simulates ring oscillators with the introduced delay cells. The performances thereof are compared and concluded.

Chapter 5 presents the original contributions, summarizes the findings of the thesis and further works are presented.

2

Overview of Single-Photon Avalanche Diode

2.1. Introduction

A pn junction diode under reverse bias voltage, V_R, disrupts an equilibrium and increases the potential barrier from built-in potential barrier, V_{bi}, to the total potential barrier, $V_{tot} = V_{bi} + V_R$. The reverse biased diode with widened depletion region width, W_D, blocks majority carriers from conducting, but allows conduction of minority carriers resulting in reverse saturation current I_S. Interestingly, as the reverse bias voltage is further increased close to and beyond breakdown voltage, V_{BD}, the junction breaks down and conducts large current. The large current is a result of tunnelling effect and avalanche effect. In general, for a lowly doped pn junction diode, tunnelling effect is negligible and avalanche effect dominates.

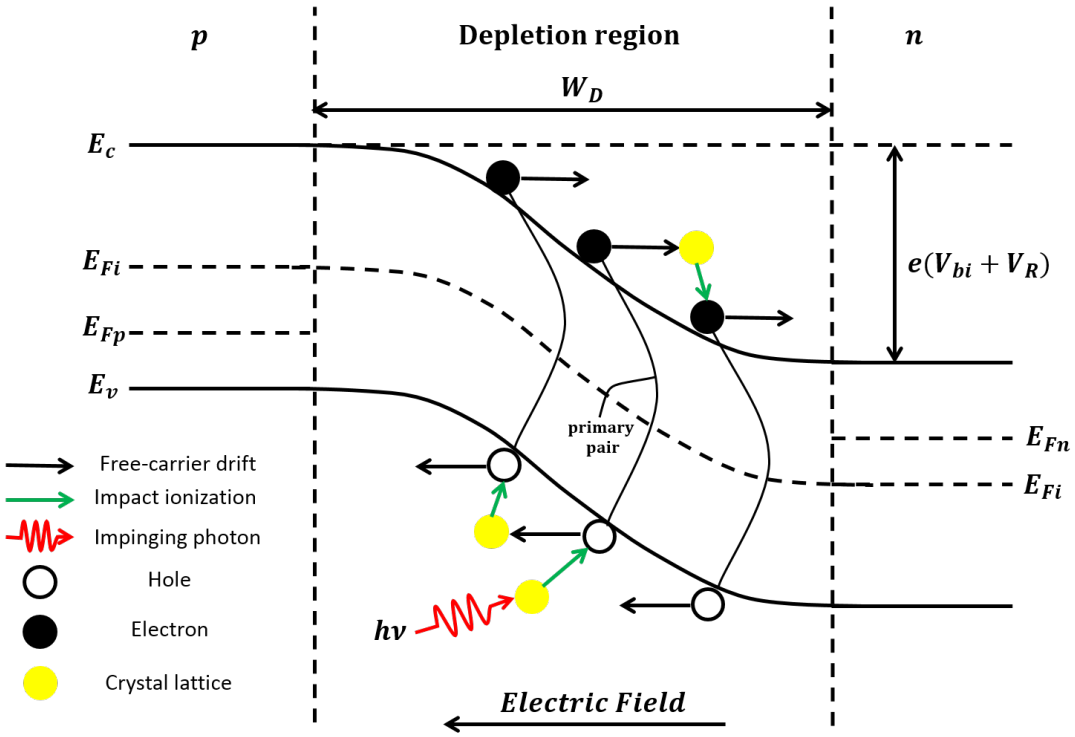


Figure 2.1: Energy-band diagram of a pn junction under reverse bias voltage beyond breakdown voltage, adapted from [1, 2]

When reverse bias voltage close to, but *below* breakdown voltage is applied to a pn junction diode, strong electric field drifts and accelerates electrons to gain enough kinetic energy to ionize crystal lattices upon

collision , generating electron-hole pairs. Then, the generated electrons may further impact-ionize lattices as a chain reaction. The resulting current depends on the applied reverse bias voltage with well defined multiplication factors, thus the operational region is referred as linear region, as shown in Figure 2.2.

When an impinging photon generates a primary electron through photoelectric effect to trigger an avalanche, the diode can be used to record photon activities. In particular, when the applied reverse bias voltage is greater than the breakdown voltage, the induced electric field drifts and accelerates both holes and electrons to cause, probabilistically, impact ionizations, as shown in energy-band diagram of Figure 2.1. The chain reaction forms positive feedback loop, thus is referred as self-sustained avalanche effect [3, 4]. The effect achieves high multiplication factor and generates a few mili-amps of current in short span of time [2]. Such characteristics allows the diode to record photon activities down to a single photon, thus is referred as single photon avalanche diode (SPAD).

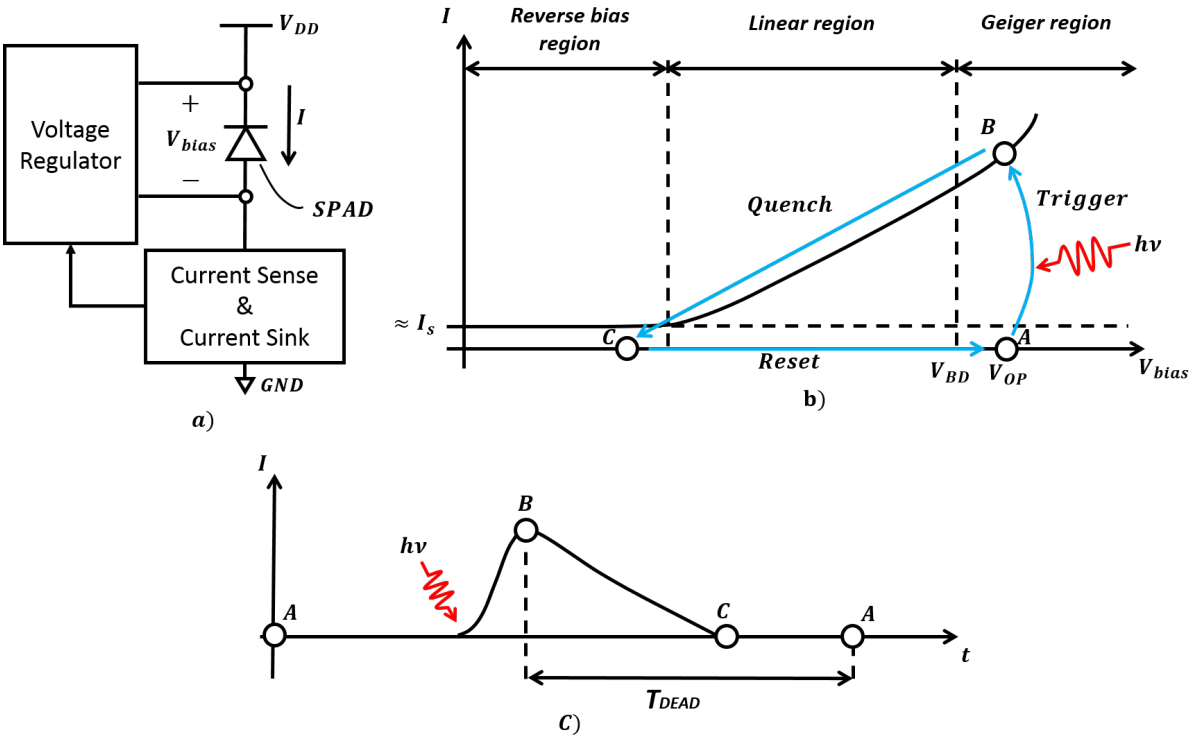


Figure 2.2: SPAD based photon detection cycle a) functional unit b) I-V curve c) SPAD current for different operating regions

The generated electron-hole pairs must be removed from the depletion region for the subsequent photon detection with functional blocks shown in Figure 2.2a. Referring to Figure 2.2b, when a SPAD is biased at an operating voltage, $V_{op} > V_{BD}$, SPAD is in the quiescent state (point A). Note that at quiescent state, SPAD does not conduct any current. At this state, minority carrier generation may lead to an avalanche, thereby causes spurious detection, thus must be kept low [5]. An impinging photon may ionize and generate a primary electron-hole pair which are drifted by electric field and may cause self-sustained avalanche. Then, the current rises quickly to ON state (point B). The generated current is sensed by the sensor block and feeds the information to the voltage regulator to reduce the bias voltage to OFF state (point C). Simultaneously, the residual free-carriers are removed from the depletion region by sinking the current, thus the process is known as quenching. A subsequent photon detection requires resetting the bias voltage to the quiescent state (point A). Note that while SPAD undergoes quenching and reset, SPAD is insensitive to the impinging photons, thus such time window is referred as dead time T_{DEAD}.

In the remaining of the chapter, the parameters denoting the probabilistic nature of photon and spurious detections are presented. Furthermore, the temporal response of avalanche trigger is further discussed. Moreover, exemplary implementations of quenching circuits and their characteristics are discussed. Lastly, state-of-the-art 3D stacked SPADs are presented and their prospects in large array LiDAR applications are discussed.

2.2. Performance parameters

2.2.1. Photon Detection Probability

Photon detection probability (PDP) refers to the probability that an impinging photon triggers self-sustained avalanche, leading to a successful detection.

- Quantum efficiency (QE) refers to the percentage of electron-hole pair generation to the impinging photons on the active area of the diode. The QE is related to the reflection coefficient and the photon absorption efficiency into the depletion region [2]. The photons absorbed in other regions of the diode such as neutral regions will generate a minority carrier which then will recombine at a certain rate with exponentially decaying lifetime. Moreover, photo-generated primary electron-hole pairs in depletion region does not guarantee the initiation of the junction breakdown.

- Avalanche breakdown probability refers to the probability of a primary electron-hole pair causing a self-sustained avalanche. The loss of kinetic energy may lead to an unsuccessful triggering of avalanche effect, thus leads to true negative detection.

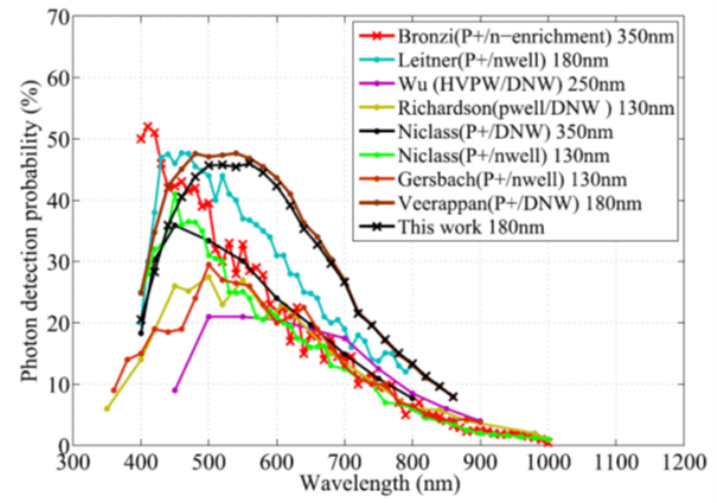


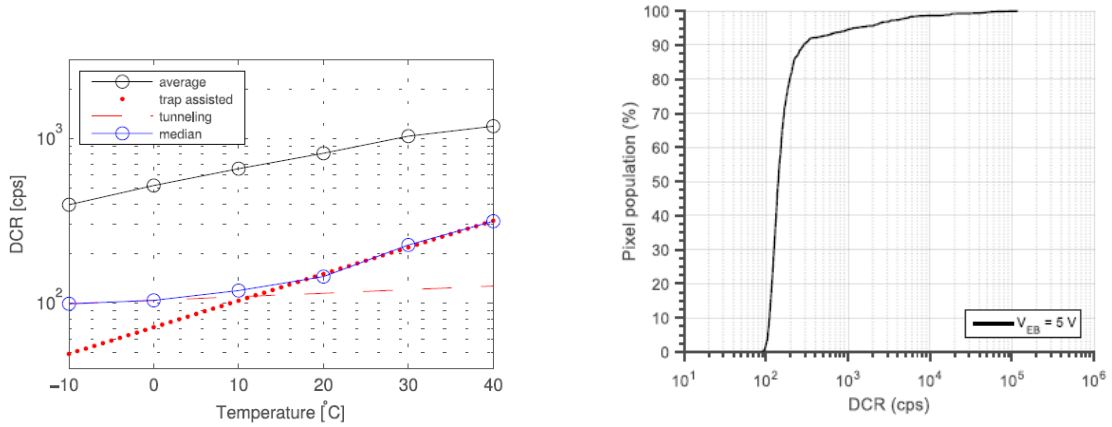
Figure 2.3: State-of-the-art PDP of SPADs, borrowed from [6]

PDP is a function of excess bias voltage and the electric field magnitude within the depletion region and the analytical model is given in [7]. Numerous researches demonstrate the measured SPAD's PDP peaks, at variable excess bias voltage, at the visible light spectrum and tails with longer wavelength, achieving <10 % around λ = 900nm as shown in Figure 2.3 [6, 8]. Peak PDP varies from 20 to 54 % in various technology nodes and standard CMOS, whereas the value converges to average of 5 % at λ = 900nm [9].

2.2.2. Noise

Uncorrelated Noise

- Dark Count rate : Dark Count Rate (DCR) is a spurious pulse frequency a SPAD generates in the absence of impinging photon, measured in count per second (cps) or in Hertz (Hz). Mainly three carrier generation mechanisms dominate DCR; trap-assisted thermal generation, trap-assisted tunnelling, and band-to-band tunnelling [2, 10, 11]. Thermal generation, described by Shockley-Read-Hall (SRH) theory, occurs when thermal equilibrium is disturbed. In case of deficit of carriers, the system generates electron-hole pairs, thereby restores the net charge density balance to achieve intrinsic charge density. The lattice defects with localized energy state close to the center of bandgap energy act as donner and acceptor to assist the process, hence the probability increases. The process is known as the trap-assisted thermal generation. The carriers are also generated due to tunnelling effect, the probabilistic nature of a carrier to overcome a potential higher than the energy of its own. Process where an electron tunnels through bandgap from a valence band to conduction band with and/or without the aids of trap is referred as trap-assisted tunnelling and band-to-band tunnelling, respectively. The tunnelling



(a) Measured DCR over temperature and average DCR of 65k-pixels, burrowed from [12]

(b) Cumulative distribution of DCR over 32×32 pixels, burrowed from [13]

Figure 2.4: Dark Count Rate over temperature and uniformity across array

probability increases with highly doped junctions resulting in a narrow depletion. Measurement results of the dark count rate over temperature, shown in Figure 2.4a [12], indicates that the dominant DCR source changes from band-to-band tunnelling to trap-assisted thermal generation over temperature, requiring SPAD temperature control to avoid false detections. The average DCR is dominated by pixels with high DCRs, referred as 'hot' pixels. The histogram of pixels over DCR can be approximated with normal distribution with standard deviation and mean increasing with temperature [12]. Cumulative distribution of DCR shown in Figure 2.4b indicates that 'hot' pixels can cause 1000 times higher dark counts. The nonuniformity of DCR poses challenge in scalability as the noisy pixels can randomly trigger at high rate, reducing the dynamic range of SPAD.

DCR also exhibits dependency with excess bias voltage. Recent researches demonstrate the influence of the geometrical doping profile on the DCR such as junction separations by means of triple-well, guard ring and shallow trench isolation [14]. The typical value of DCR/μm^2 measured at room temperature ranges from 10^0 to 10^4 cps in various process node[9, 15–19]. The comprehensive analytic model presented in [11] accurately predicts the performance without extensive simulation resources.

DCR is an important performance parameter for the system level study of 3D imagers, as the minimum noise level is calculated, from which inherent detection limit is determined. DCR is indistinguishable with the impinging photon-triggered avalanche in Geiger mode, owing to the large multiplication factor caused by self-sustained avalanche.

Correlated Noise

- After-pulsing probability : During the avalanche period of Geiger mode APD, large number of the carriers flow traverse depletion region, some of which are captured by the traps in the depletion region. The trapped carriers are re-emitted after a release time depending on the trap energy level, with the center traps having the longest release time. Once the SPAD is quenched and recharged, the re-emitted carriers have a probability to trigger an avalanche, thus is termed as afterpulsing probability. Intuitively, afterpulsing is proportional to the number of carriers generated in avalanche, n, the concentration of the carrier traps, N_t, the fraction of the unoccupied traps at the start of the avalanche $(1 - f_t)$, and the probability of carrier emission from trap, η. Furthermore, probability increases with wider high electric field region effective width, W_e, and the area of the carrier capture, A, [10]

$$P_{ap} = nN_t(1 - f_t)\eta W_e A \tag{2.1}$$

where the P_{ap} can be calculated for every trap level for both electrons and holes. The number of generated carriers n is proportional to the excess bias voltage, V_{ex}, posing the direct trade-off between PDP and spurious avalanche trigger probability. The probability decays exponentially with time. In [20], after pulsing probability of 0.1% after 100 nsec was achieved when biased at $V_{ex} = 1V$ and in another

paper [21] 0.02% was achieved after 50nsec when excess bias is 1.2V. Most recent work report negligible after pulsing probability with respect to other noise sources such as dark count rate. Moreover, exponentially decaying property allows for controlled probability with well defined SPAD's off time.

- Cross-talk : Crosstalk refers to an avalanche trigger upon photon absorption in the neighbouring SPAD and occur electrically and optically. A photo-generated carrier under PN junction may diffuse to another SPAD in vicinity to an electric crosstalk. Alternatively, a photon emitted during the avalanche process by deceleration of hot carriers can traverse and trigger an avalanche in the nearby pixel. Such electro-luminance interaction decreases with the square of the distance and can be addressed with isolation trenches [22]. As a result, successful designs report negligible crosstalk probability [23].

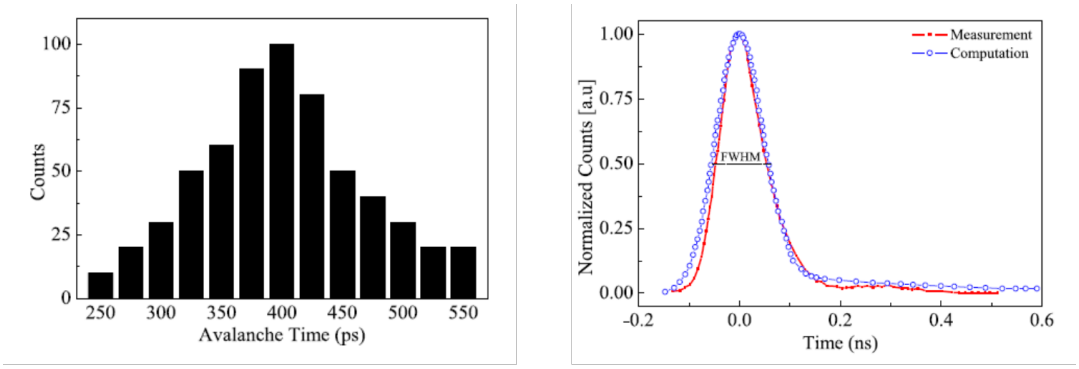
2.2.3. Dead time

Dead time is the period when SPAD is insensitive to photon absorption due to unsatisfactory conditions to initiate self-sustained avalanche. Dead time is not an intrinsic performance parameter of SPAD as the sensitivity can be controlled by the pixel logic. Nevertheless, deadtime plays a decisive role in minimizing after-pulse induced avalanche and system level optimization in 3D imagers, especially in high counting applications, as will be discussed in Chapter 3.

2.2.4. Timing jitter

The statistical spread in timing response of a SPAD is known as the temporal resolution or timing jitter. It is the time interval between the photon arrival and the detection of the edge pulse, and is typically reported in full-width at half maximum (FWHM) of the Gaussian distribution. The distribution consists of the two components; Gaussian peak followed by exponentially decaying tail [10, 11, 24].

Depending on the absorption region within the diode, the diode exhibits varying avalanche timing response as shown in Figure 2.5a). The Gaussian peak is related to the statistics of the avalanche build-up within depletion region, taking into account the carrier transit time [25, 26]. The avalanche build-up can be accelerated by the lateral propagation of the photons emitted by hot carriers, causing secondary avalanches[4]. A photon absorbed in the neutral regions generates a minority carrier which moves around randomly due to the absence of electric field. The carrier diffuses for a relatively long period before triggering an avalanche, responsible for the tail, and thereby broaden the timing response. The measured timing jitter shown in Figure 2.5b) is aligned with the computed prediction, proving the theory.



a) **b**)

Figure 2.5: SPAD temporal response a) measured avalanche buildup time distribution at $V_{ex} = 0.5V$ and b) measured and computed timing jitter curve. Figures borrowed from [24]

Timing jitter is dependent on excess bias voltage, indicating the trade-off between PDP. The simple analytic model that does not require time consuming Monte Carlo simulation is presented in [24]. The high multiplication factors of a SPAD yields a rapid transition. The recent literatures report a measured jitter with 67 ps, 27 ps and 7.8 ps in, respectively [27–29]. The stochastic nature of the timing jitter results in the uncertainty in the depth accuracy, even with the ideal jitter-free readout electronics. For instance, 67 ps of jitter corresponds to an uncertainty of 2.01 cm in distance, using the speed of light. Thus, the temporal resolution of the following blocks do not require a higher temporal resolution than of the SPAD.

2.3. Quenching and reset circuits

The sequential photon detection using SPAD requires a front-end electronics to regulate excess bias voltage which in turn controls the self-sustained avalanche. When the bias voltage is reduced to or below break down voltage, the magnitude of the electric field decreases and eventually halts the avalanche. The quenched SPAD is insensitive to photons until the voltage is reset above breakdown voltage. Quenching and reset circuits can be passive, active, or combination of both. The quantitative analysis of the quenching and reset circuits is given in [3]. In this thesis, the basic operation principles and issues are discusses qualitatively.

2.3.1. Passive quenching

Passive quenching circuit refers to a circuit where an applied operational voltage of a SPAD is self-regulated without forming a control feedback loop. A basic configuration of the passive quenching circuit (PQC) and node voltages thereof under photon detections are shown in Figure 2.6. The cathode and the anode of the SPAD are connected to a large load resistor (R_L) and a small source resistance (R_s), respectively. The supply voltage (V_{DD}) is larger than the breakdown voltage of the SPAD (V_{BD}).

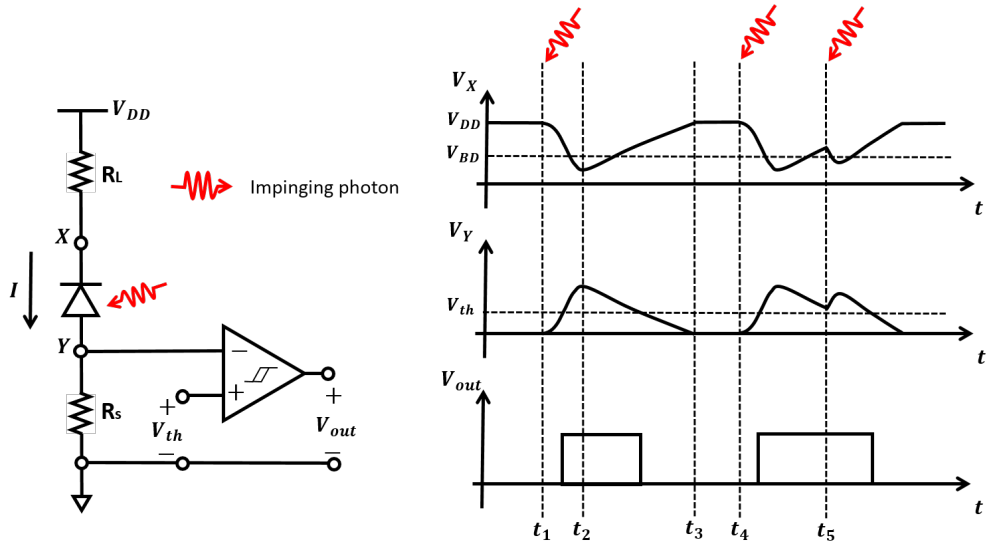


Figure 2.6: An exemplary Passive Quenching Circuit (PQC) and the corresponding node voltages under photon detections, adapted from [3]

Initially, SPAD is set to quiescent state (point A in Figure 2.2). That is, SPAD is reversed biased above breakdown voltage ($V_{XY} = V_{DD}$) and there is no current through the branch. Upon photon absorption and successful trigger of self-avalanche at $t = t_1$, the surge of current flows through the load resistor and causes the voltage drop at node X to below breakdown voltage. Simultaneously, the induced current through the small resistor causes voltage drop over the small resistor. The voltage at node Y is compared with a threshold voltage (V_{th}) and outputs a step transition to a high state, indicating a photon detection. As the current increases, the reverse biased voltage applied over the SPAD reduces, reducing the electric field strength within the depletion region, thereby quenching the avalanche ($@t = t_2$). As the SPAD quenches, the number of avalanche carriers reduces, and eventually self-sustaining condition is no longer met. Consequently, the current decreases and the reverse bias voltage recovers to the operational voltage ($V_{DD}@t = t_3$).

Similarly, a subsequent photon is detected at $t = t_4$. After a successful photon detection, the reverse bias voltage is recovering when another photon triggers self-avalanche at $t = t_5$. However, the lower reverse bias voltage than the operating voltage($V_{XY} < V_{DD}$) generates currents with a lower amplitude, thus recording a lower peak voltage at node Y. Notice that $V_Y > V_{th}$ at $t = t_5$ which prolongs the time interval that output voltage stays on the high state, instead of step transition. As the result, the photon detection at $t = t_5$ may not be registered and only causes prolonged deadtime.

The quenching and reset time is often approximated using the time constant set by the load resistance and the total capacitance associated with load and SPAD, $\tau_t = R_L C_{spad}$ [3] . The large load resistance increases the time constant which in turn reduces the maximum count rate. In a SPAD-based depth imagers using time-correlated single photon counting principle (TCSPC), the maximum count rate directly relates to SNR

and detection range, as will be shown in Chapter 3.

The above mentioned drawbacks of the passive quenching circuit renders the method impractical for high counting rate applications such as an outdoor and high frame rate system. In particular, considering the sun light as individual photons arriving at random time instances, a strong sun light has high probability that two photons are temporally close to one another, rendering the circuit inappropriate for outdoor applications.

2.3.2. Active quenching

An active quenching circuit (AQC) refers to a circuit with a control loop to manipulate the applied reverse bias voltage of SPAD. An exemplary AQC and the correspondence node voltages under photon detections are shown in Figure 2.7. The core idea of AQC is to sense the rise of avalanche, often at an early stage, and drive the bias voltage (V_{XY}) to below breakdown voltage (V_{BD}).

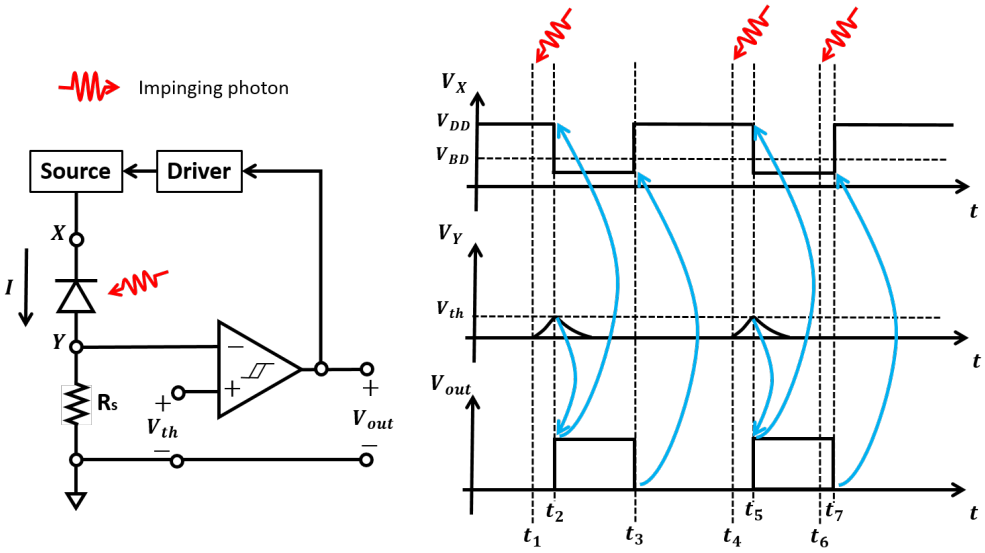


Figure 2.7: An exemplary Active Quenching Circuit (AQC) and the corresponding node voltages under photon detections, adapted from [3]

Initially, SPAD is set to quiescent state with bias voltage above breakdown voltage. Upon photon absorption and successful trigger of self-avalanche at $t = t_1$, the surge of current causes voltage drop over a resistor (R_s) and is compared to a threshold voltage (V_{th}) to determine photon detection (arrow from $V_Y \rightarrow V_{out}$ at $t = t_2$). The output transits from low to high state. Simultaneously, the driver drives the source to lower the voltage at node X below the breakdown voltage (arrow from $V_{out} \rightarrow V_X$ at $t = t_2$). Consequently, the avalanche halts and SPAD enters a predefined deadtime ($T_{DEAD} = t_3 - t_2$). After the deadtime, the driver drives the source to set the voltage to the SPAD operating voltage (arrow from $V_{out} \rightarrow V_X$ at $t = t_3$).

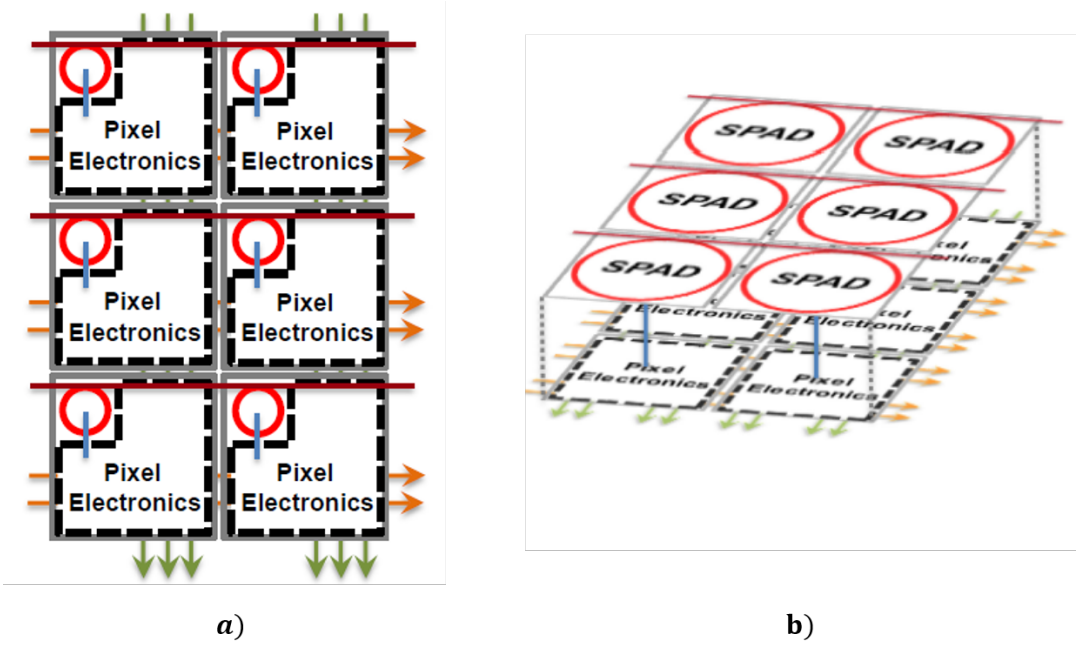
Similarly, a subsequent photon is detected at $t = t_4$. After a successful photon detection, the SPAD is rendered insensitive ($t_5 < t \leq t_7$) during which another photon is absorbed at $t = t_6$. The bias voltage at the instance is not sufficient to trigger a self-sustained avalanche, therefore the photon detection at $t = t_6$ is not registered.

As shown, AQC achieves fast transitions with well defined deadtime. The rapid reset reduces the probability of premature avalanche during recovery, thus improves maximum counting rate. In practice, the period is set to avoid false detection arising from after pulses, which sets the limit to the maximum counting rate. The deadtime is an important system parameter to be optimized in a TCSPC system, as will be discussed in Chapter 3.

2.4. 3D stacking technology

Figure 2.8a illustrates a conventional pixel composition in a multi-pixel system. Each pixel consists of an photo-sensitive area (indicated by red circles) and pixel electronics. The ratio between the photo-sensitive area and the pixel area is referred to as a fill factor which is always less than one in practice due to active region's geometry and pixel electronics such as time-to-digital converter, memory, and readout electronics as

shown in Figure 2.8a. Furthermore, physical barriers surrounding SPADs further decrease fill factor. As the result, fill factor further decreases, even down to 3.4% [30].



a) **b**)

Figure 2.8: Comparison of SPAD arrays (a) 2D fully-parallel (b) 3D-stacked, borrowed from [31]

Fill factor can be improved, effectively, by means of microlens to converge and direct incident light to the SPAD active regions. In [32–34] refractive approach was presented to achieve concentration factor upto 10 times, but the potential issues with yields and reproducibility renders the technique inappropriate for scaling. In contrast, the diffractive microlens approach presented in [30, 35] achieves 15 times of improvement in effective fill factor around the wavelength of green until NIR. However, inherently, misalignment between microlens and SPADs causes sharp drop in the concentration factor, rendering the microlens approach appropriate for relatively static environment [33, 34].

Recent works implement SPADs and digital circuit on separate wafers and bond, the technology referred as 3D-stacking. Figure 2.8b illustrates an exemplary implementation where the top tier consists of photosensitive region and pixel electronics are located on the bottom tier. Physically, such segregation allows a higher fill factor and offers more chip area for pixel electronics, in comparison to the conventional planar technology.

Pixel logic and digital blocks can be implemented on the bottom tier in an advanced CMOS, which allows SPADs to be implemented in a dedicated node such as CMOS image sensor process to optimize performances [14, 36]. However, the limitation in the excess bias voltage in an advanced node must be considered to ensure high avalanche probability. Alternatively, Pixel logic can be implemented on both tier. Pixel logic on the top tier enforces SAPDs to be implemented in CMOS node, typically high-voltage, thus high excess bias voltage can be achieved even in advanced lower node. However, the integrated pixel logic occupies the area, thereby reduces fill factor. In both cases, bottom tier can be implemented in high-voltage node, but at the cost of lower circuit density and the speed. The speed determines the minimum gate delay which relates to the minimum resolution in a 3D depth sensors, thus must be considered on a system level.

Figure 2.9 illustrates cross section of two types of stacking technology. In both cases, the top tier consists of SPAD for illumination, with digital blocks on the bottom tier. The difference lies on the orientation of the SPAD layer with respect to the illumination direction. The front-illuminated SPADs, as show in Figure 2.9a bonds the substrate with the bottom tier, thus require through-silicon-via (TSV) to propagates the detection signal to the bottom tier. The TSV reduces the fill factor and limits the interconnection between the two tiers. In contrast, back-illuminated SPAD directs the substrate towards the illumination and bonds face-to-face to the bottom tier without additional data path. However, it requires substrate thinning process, down to a few micrometers, to optimize PDP for different wavelengths.

The performances of the state-of-the-art is summarized in [31, 37]. The improved fill factor achieved

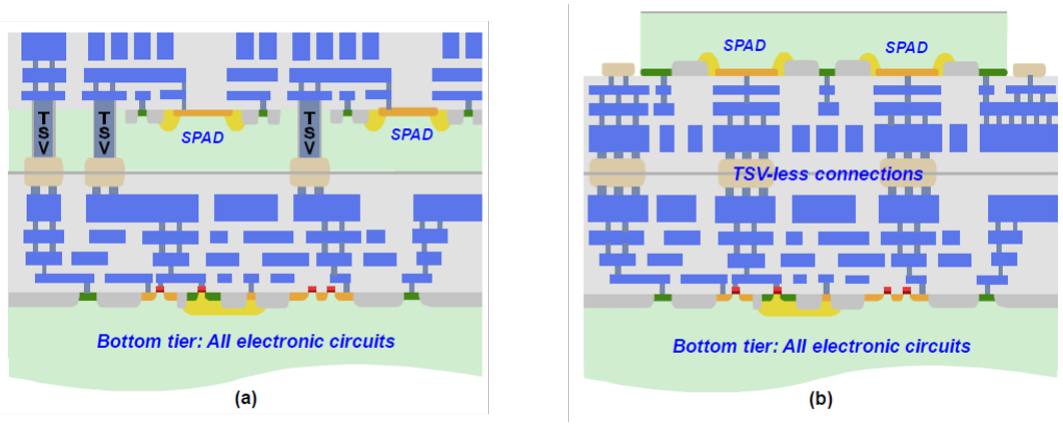


Figure 2.9: Cross section of general 3D-stacked imager with (a) front-illuminated and (b) back-illuminated SPAD, borrowed from [31]

74.4% in [38] and in [14] a peak PDP of 31.8% at $\lambda = 600nm$ was achieved with excess bias voltage of 2.5 V. The PDP at $\lambda = 900nm$ recorded consistently 5 % to 10 % when biased at the highest- PDP-achieving excess voltage, in spite different structures to manipulate the junction depth and thickness of the depletion region [14, 38–40] . In most of the works, DCR of a few kcps was achieved at room temperature with negligible after-pulsing probability. Note that, in general, photon detection is dominated by the background light for SPAD based imagers targeting outdoor application, rendering the state-of-the-art noise parameters negligible.

2.5. Conclusions

As discussed, SPAD triggers self-sustained avalanche upon photon absorption, resulting in a few mili-amps of current in a very short span of time, down to a few pico-seconds of FWHM. When connected to pixel logics and further processing units such as time-to-digital converter, the detection time with high precision is known. Analytical models of SPAD designs are well studied and recent works achieve peak PDP up to 55%, but starts decaying to achieve average of 5% around $\lambda = 900nm$ among different technology nodes. The operation at NIR wavelength is critical for outdoor application as the absorption rate of the natural light is the highest within the band. Additionally, the direct trade-off between PDP and noise parameters poses the challenge in achieving high SPAD dynamic range. Furthermore, implementation of the auxiliary circuits within a pixel in plannar technology, reduces the fill factor, posing the challenge in scalability.

Microlenses with refractive and diffractive approaches improve effective fill factor, but the non-uniformity and the skewed concentration caused by misalignment limits the area of applications. Alternatively, 3D stacking technology allows physical segregation into different tiers, thereby increases fill factor of SPADs with larger dedicated circuit area. Furthermore, SPAD can be implemented in a dedicated technology node optimized for a CMOS image sensor application to optimize SPAD performances. The back-side illuminated SPAD supports face-to-face bonding without additional TSV process step.

Quenching and reset circuits aids SPAD with serial photon detections and determine the achievable counting rate. Passive reset circuits suffer from relatively long reset time due to large time constant, during which avalanche-triggering photon absorption may not not be observed and increase the deadtime. In contrast, active reset circuit achieves fast transition to operating voltage, thereby reduce false-detection probability. The operation of SPAD, in author's opinion, is the key to achieve high system performance in 3D imagers for outdoor applications, thus is investigated in the Chapter 3 with state-of-the-art SPAD performances.

3

Time-of-Flight in SPAD-based LiDAR

3.1. Introduction

Time Correlated Single Photon Counting (TCSPC) is a method to record the time difference between a reference signal and photon detection. Figure 3.1 illustrates a generic system block diagram of a SPAD-based LiDAR system applying TCSPC to retrieve a distance to an object. The system block comprises a transmitter to repeatedly generate pulses with a set period t_r and a receiver comprising a bandpass filter to filter out undesired wavelengths of an incident rays, a lens to direct the filtered rays onto a SPAD array, a SPAD array, gating circuits, and a time-to-digital converter (TDC) array. Upon reflection of the transmitted pulses on the object, the pulse travels back through a medium and arrives at the receiver R_x.

A successful detection at the SPAR array triggers the TDC array which generates the time difference between the emission and the detection. The digitized time difference is stored and the process repeats at the next period. Iteration yields a histogram, which ideally reconstructs the received signal.

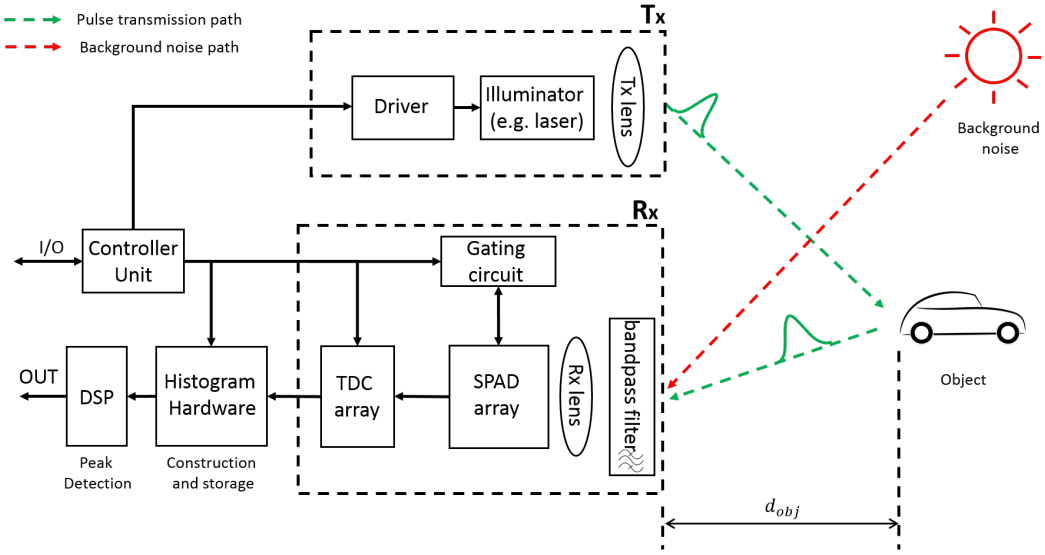


Figure 3.1: Generic SPAD-based Time Correlated Single Photon Counting (TCSPC) system block diagram for direct-time-of-flight (dToF) method

Direct time-of-flight (dToF) considers the bin with the highest counting and calculates the distance to the object according to the speed of the light as

$$\widehat{d_{Obj}}[\tau] = \frac{1}{2} c \times \tau_{i,peak} \times \Delta t \tag{3.1}$$

where $\tau_{i,peak} \in \{1, 2, \cdots, n_b\}$ and Δt denote the histogram bin index with the highest counting and the histogram bin size, respectively. The bin size is limited by the resolution of TDC.

Figure 3.2a depicts two constructed histograms of object-reflected pulses of a system implemented as illustrated in Figure 3.1, but utilizing a single SPAD. Figure 3.2aa illustrates an ideally reconstructed histogram of the object-reflected (or received) signal $\lambda(t)$ using n_b bins. The ideal histogram records constant noise counts and signal counts superimposed on the noise counts and the object distance d_{obj} is calculated using the bin with the highest counts τ_{peak} according to eq.(3.1). The ideal reconstruction assumes maximally one photon detection per repetition period. That is, photon arrival rate is low, thereby the detection probability is also low such that the detector is always available for detection at any bin within the repetition period t_r. Such assumption does not impose any requirement on the receiver operation in terms of synchronicity to the reference clock.

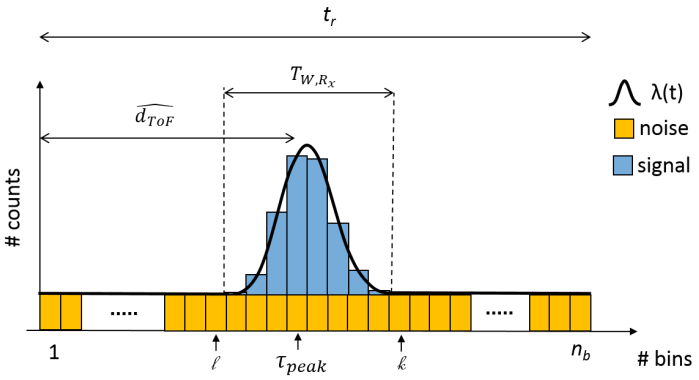
In contrast, Figure 3.2a illustrates a poorly reconstructed histogram of the same system in a scenario where low photon rate assumption no longer holds. Suppose that the receiver is synchronized to the repetition period t_r such that SPAD is available for detection at the beginning of each repetition period. For a constant photon arrival rate λ, the detection probability of the first bin follows the Poisson distribution

$$Pr_s = Pr(X \leq 1) = 1 - Pr(X = 0) = 1 - e^{(-\lambda \Delta t)} \tag{3.2}$$

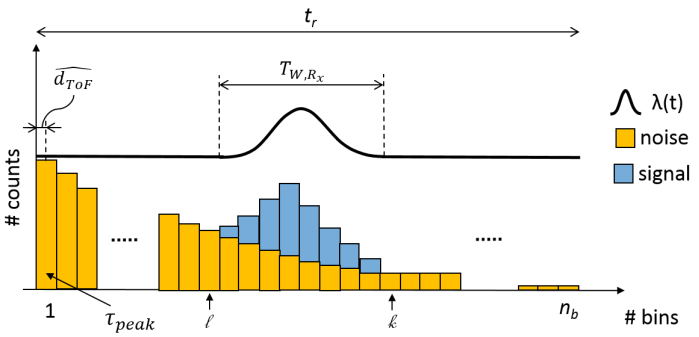
where X denotes a random variable representing the number of detected photons. Intuitively, following the restriction of a single detection per repetition period, the detection during the next bin is only possible when the photon is not detection in the first bin. Thus, detection probability at the n^{th} bin is the probability that no photon detection occurs from the first bin to $(n-1)^{th}$ bin and the detection probability can be expressed as:

$$Pr_s^{(n)} = Pr_{av}^{(n)} \times Pr_{r,det}^{(n)} = \exp(-\lambda \Delta t)^{n-1}(1 - \exp(-\lambda \Delta t)) \tag{3.3}$$

where the super script denotes the bin number. The resulting histogram shown in Figure 3.2b visualizes the detections at earlier bins blocking the detections at the later bins within the time period t_r, the phenomenon referred as pile-up effect. Consequently, the peak detection yields largely inaccurate estimation $\widehat{d_{obj}}$.



(a) Ideal reconstruction



(b) Poor reconstruction due to pile-up effect

Figure 3.2: Histogram reconstruction of the received signal $\lambda(t)$ in TCSPC using single SPAD

In summary, the conventional TCPSC successfully reconstructs the received signal in environments where less than one photon detection is expected per repetition period. Such condition is referred as photon starved

condition and is often found in controlled environments with low background noise. The method may be used in a photon abundant environment, but at the cost of detection distance. However, the automotive application requires long detection range (preferably >100m) in the presence of strong background noise level (tens of kilo-flux). Thus, the LiDAR receiver system must be designed according to the strong background noise requirements.

Beside the aforementioned methodological issues in implementation of TCSPC for automotive applications, there exists practical limitations with scalability of the system. Firstly, high sensor activity and long range detectability with milimetric resolution requirement translates to a large data size to be processed either internally or externally. Internal or on-chip processing enables parallel processing, but is limited by the physical chip area allocated to memory. On the other hand, external or off-chip processing implies potential saturation in data transfer, a bottleneck in bandwidth posed by the chip interface such as I/O pad.

The remaining of the chapter is organized as the following: Section 3.2 presents the state-of-the-art system level solutions towards the aforementioned methodological and and practical limitations, Section 3.4 discusses three possible solutions. In particular, Section 3.4.1 presents the synchronous SPAD operation with time window shifting technique based on [41]. Then, Section 3.4.2 presents asynchronous SPAD operation with multi-photon systems. Section 3.4.4 discusses the proposed data compression technique based on the proposed asynchronous SPAD operation system. The statistical model of the presented asynchronous SPAD operation is explained in Section 3.4.3. Finally, Section 3.5 presents the simulation results of the presented techniques and their performances are compared.

3.2. State-of-the-art

Coincidence detection

In the interest of data compression and TDC power efficiency, it is essential to reject noise-related photons, and detect and register only the signal.[42, 43] achieves such filtering of uncorrelated noise using a spatio-temporal correlation principle. The temporally concentrated target reflected pulses have higher probability

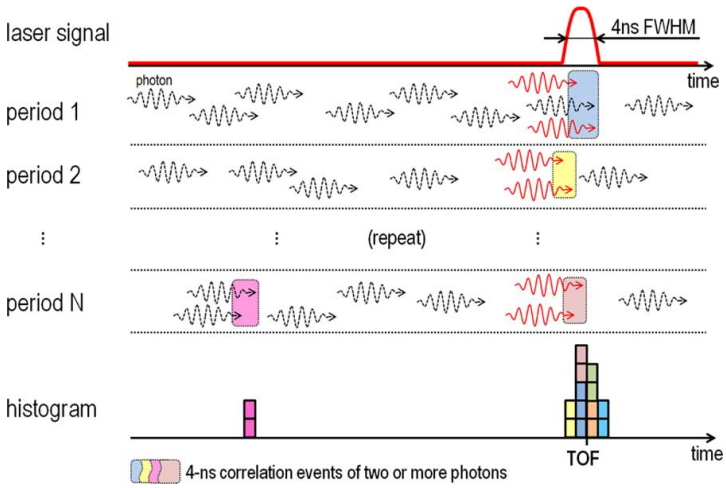


Figure 3.3: Diagram of the Spatiotemporal principle borrowed from [43]. The waves represent photons where the tip thereof indicates photon detections on each SPAD. When more than two tips coincide within a time frame indicated as coloured boxes, the timings are stored in the histogram.

to trigger SPADs upon absorption, leading to temporal correlation within a signal window. Moreover, neighbouring SPADs covering FOVs in vicinity are likely to detect photons at the same time for the distance. The combination of both features are referred as spatio-temporal correlation principle. As shown in Figure 3.3, once the detection number exceeds a predetermined threshold, time stamp is generated and stored. In contrast to temporally correlated photons arising from reflected pulse, noise-related photons arrive at random time instances. Thus, uncorrelated triggers are rejected. However, maximally resolvable detection range is shortened due to exponentially decreased signal power with respect to the range [44]. More importantly, macro-pixel based filtering reduces final image resolution. Alternatively, an adaptive approach can be applied to set the threshold for assessing the spatio-temporal correlation according to the ambient noise level, thus enhancing robustness against dynamically changing environment[45]. Nevertheless, such adaptive sys-

tem suffers from the above-mentioned short-comings of the spatio-temporal correlation detection such as reduced detection range and lower image resolution.

Partial Histogramming readout
In attempt to reduce in-pixel memory area for histogramming while retaining a long detection range with high temporal and image resolutions, a zoom-in approach can be applied. The zoom-in approach determines peak time bin for each iteration and records a window of neighbouring bins for the succeeding iteration with an improved time resolution. Then the fine bins within a portion of the full detection range are saved on-chip [13].

The histogram counts on the remaining bins are readout off-chip to construct full-range histogram, thereby reducing the data transfer rate. The method is referred as partial-histogramming. The biggest drawback of partial histogramming is the loss of frames. The zoom-in approach reduces the measurement time given the constant frame-rate. Furthermore, incorrect peak detection at an early stage leads to significant error in retrieved ToF. Nevertheless, zoom-in approach shortens the range of interest, relaxing timing requirement of ring-oscillators as base TDC block, a useful insight for long-range detection scenario.

Reconfigurable Flash LiDAR
A reconfigurable system presented in [46] configures the system depending on the application to offer flexibility. In particular, 3D depth image sensing configures 16 SPADs as a pixel and connects to 16 TDCs, counting maximally 16 counts per laser pulse. Furthermore, different timing sources are chosen depending on the detection range. As a result, histogram time resolution is varied and time conversion power is saved. However, the system trades spatial resolution to enable multi-photon counts per laser cycle. Furthermore, temporal resolution is traded to meet timing and power requirements, leading to increased uncertainty range of time of flight. However, timing source selection based on detection range offers power reduction and relaxes jitter requirements which are crucial system considerations for a large TDC array measuring long range.

3.3. Link budget modeling

Figure 3.4 illustrates a propagation and reflection of a transmitted pulse in a LiDAR system. The pulse transmitted by a transmitter TX travels a distance to the object d_{obj} through a medium with an attenuation factor α, reflects on the Lambertian surface, and returns to the receiver RX. It is assumed that the illuminated object covers the entire area of field of view covered by the optical system, such that there is no energy loss. Furthermore, the transmitter is aligned parallel to the normal of the object surface, such that an angle of incidence is zero.

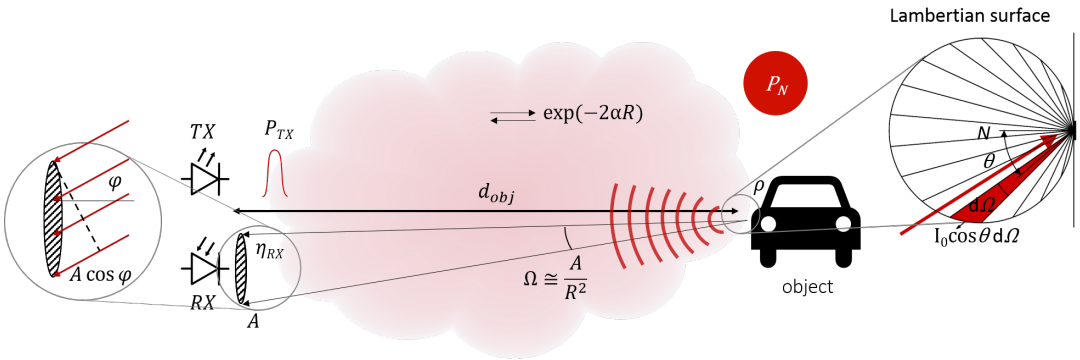


Figure 3.4: Propagation and reflection of the transmitted pulse in a LiDAR system assuming the target surface is Lambertian surface.

The received optical power P_{R_x} at the receiver of a LiDAR system is given by:

$$P_{R_x}(t) = \eta_{R_x}\left(\int_{r=0}^{ct/2} P_{T_x}\left(t - \frac{2r}{c}\right)H(r) + P_N\right) \tag{3.4}$$

where η_{R_x} denotes receiver efficiency, r and t respectively denote distance and time variables, c denotes speed of light, and $P_{T_x}(t)$, $H(r)$, and P_N denote transmitted power over time, channel impulse response, and

noise power, respectively [23, 44]. The integral sign indicates the convolution between the transmitted pulse and the medium.

Figure 3.5 shows a spectral solar irradiance provided by the American Society for Testing and Materials (ASTM). The shown ASTM G173-03 refers to the terrestrial solar spectral irradiance $[W/m^2/nm]$ on a 37deg sun-facing tilted surface in a cloudless condition. Accordingly, the noise power P_N given in eq.(3.4) is derived. The average irradiance around central wavelength of $\lambda = 905nm$ over the bandwidth of $10nm$ is used as the

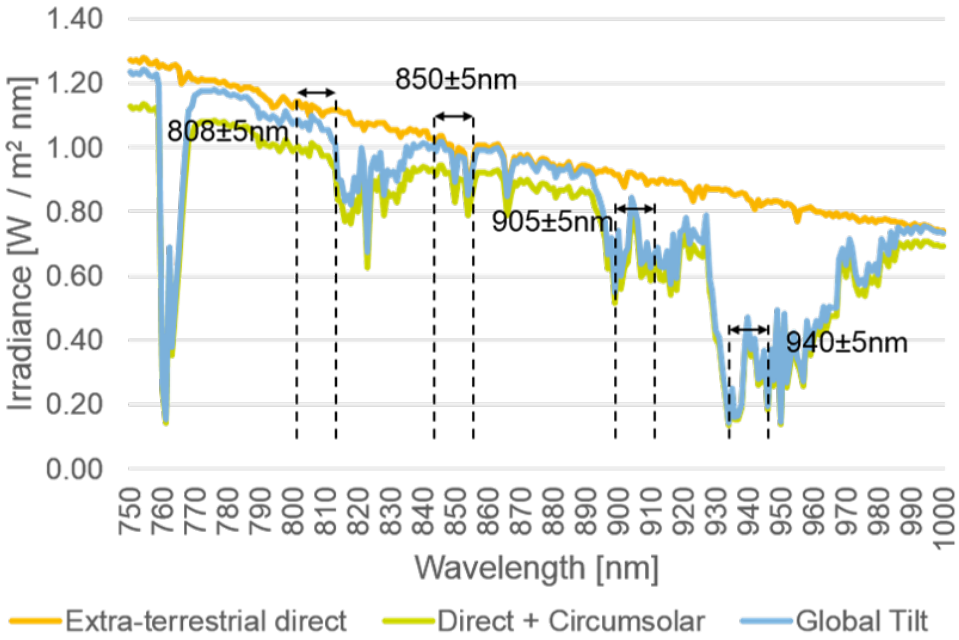


Figure 3.5: ASTM G173-03 spectral solar irradiance provided by the American Society for Testing and Materials

sun irradiance I_{sun}. Assuming that the target surface is tilted at an angle the power noise is given by:

$$P_N = I_{sun}\Big(\frac{fov_{lens}}{2}\Big)^2 A_{lens}\rho BW_{opt} \tag{3.5}$$

where fov_{lens}, A_{lens}, ρ, and BW_{opt} denote lens field of view, lens aperture, object reflectance, and lens optical bandwidth, respectively.

The further calculation of the received optical power is given in the following. Assuming that the target's surface exhibits Lambertian reflectance, that is the surface diffuses the incident ray isotropically such that the same radiance $[Wsr^{-1}m^{-2}]$ is observed regardless of the direction of the observer. The reflected radiance is given by

$$I = I_o \cos\theta \cdot \partial\Omega \tag{3.6}$$

where I_o, θ, and $\partial\Omega$ denote initial radiance incident normal to the surface, angle between the normal and the incidence vector, and unit solid angle, respectively. The unit solid angle is given by

$$\Omega = \frac{A_{Rx,sr}}{R^2} \tag{3.7}$$

where $A_{Rx,sr}$ and d_{obj} denote receiver area within the unit solid angle perceived by Lambertian surface and target distance, respectively. The receiver area is angled at ϕ with respect to the reflected ray incident to $A_{pix,sr}$, thus is given by

$$A_{R_x} = A_{R_x,sr} \cdot \cos\phi. \tag{3.8}$$

Substituting eq.(3.7) and eq.(3.8) into eq.(3.4) yields the received power as

$$P_{R_x}(t) = \eta_{R_x}\Big(P_{T_x}\big(t - \frac{2d_{obj}}{c}\big) \cdot \rho_{target} \cdot \cos\theta \cdot \frac{A_{R_x}}{R^2} \cdot \cos\phi \cdot exp(-2\alpha R) + P_N\Big) \tag{3.9}$$

where ρ_{target} denotes the target reflectivity and α denotes the attenuation factor of the medium and is the function of the distance.

The received power is focused onto the image sensor array area A_{array}, by a lens with a given optical efficiency η_{lens} which takes into account energy loss during the transmission through the lens. The power received per pixel for an array with N_{pixels} pixels is given by [44]

$$P_{pixel} = P_{R_x} \cdot \eta_{lens} \cdot \frac{2}{N_{pixels}\pi} \tag{3.10}$$

The receiver area A_{R_x} in terms of optical components is given by

$$A_{R_x} = \pi\left(\frac{D}{2}\right)^2 \tag{3.11}$$

where D denotes diameter of the receiver. Assuming the receiver area is equivalent to the lens area, the diameter is given by

$$D = \frac{f_o}{f\#} \tag{3.12}$$

where f_o denotes focal length and is calculated as

$$f_o = \frac{h \cdot R}{HFoV} \tag{3.13}$$

where h denotes height of the sensor array and $HFoV$ denotes horizontal field of view. The above relations between the received pixel power and the optical components form the basis of the system simulation on Matlab using the simulated parameters given the appendix.

3.4. Receiver operation for photon-abundant application

3.4.1. Synchronous interleaved SPAD gating

Figure 3.6 shows alternating SPAD gating waveforms with phase shifts and the received pulse $\lambda(t)$. The gating windows purposely discharge SPADs to be insensitive to incoming photons and subsequently recharges to allow detection at later time instances with respect to the reference timings, for example the beginning of every laser transmission period T_r. Repeating such measurements with phase shifts consequently increases detection uniformity over the detection range [41]. The technique is known as interleaved SPAD gating. Intuitively,

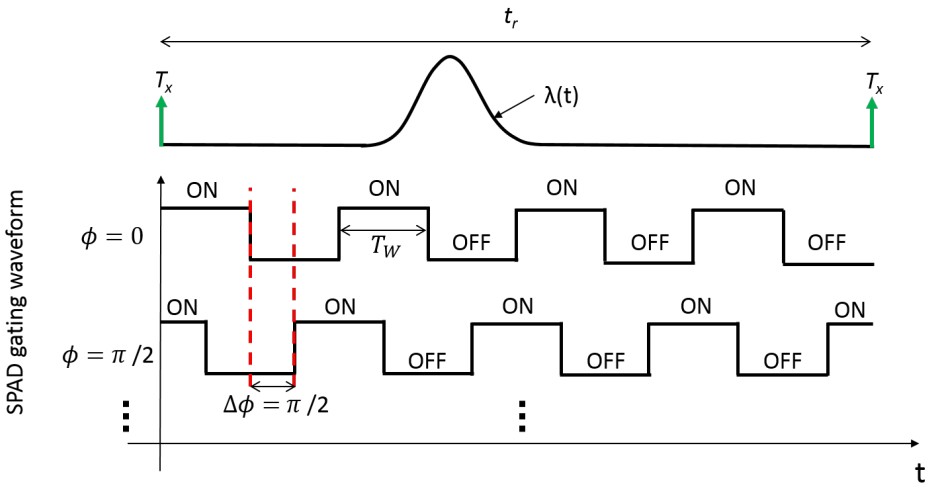


Figure 3.6: SPAD gating waveforms of the interleaved SPAD gating scheme

pile-up effect exhibits peak counts at the start of each gating window and decreasing in an exponential manner. In other words, the amplitude of the peak counts decreases, but instead spreads over the time period. Such phenomenon is referred as the cluster-edge effect. The cluster-edge effect is a function of the gating

window and the coarseness of the phase-shift, thus the detection uniformity normalized to the maximum detection can be written as:

$$\frac{\delta D}{D_{max}} = 1 - \exp\left(-\frac{T_W}{q\tau}\right) \tag{3.14}$$

where T_W denotes gating window period, q denotes a phase division factor, and τ denotes noise photon detection meantime [41]. Intuitively, the narrow gating window and fine-phase shift covers time bins relatively uniformly, thereby decreases the cluster-edge effect which is in agreement with the equation.

3.4.2. Asynchronous SPAD operation

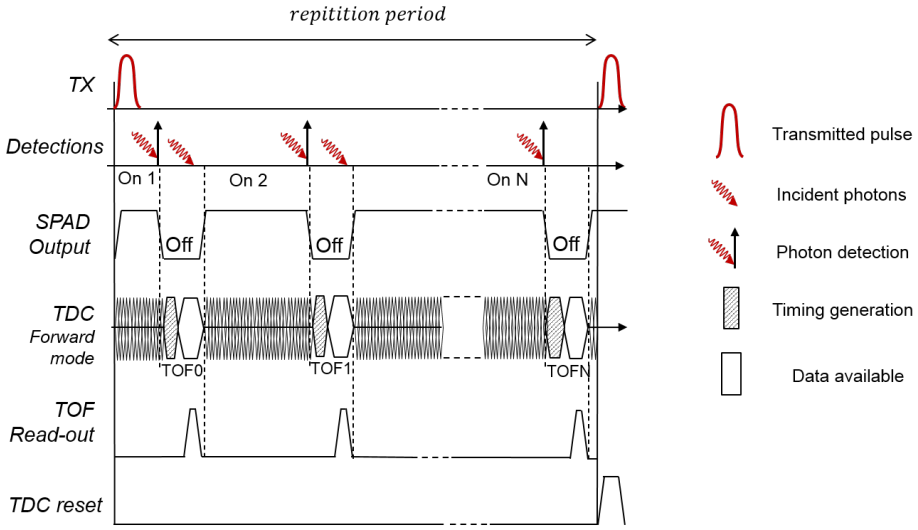


Figure 3.7: Asynchronous SPAD operation timing diagram with non-paralizable receiver in a photon abundant environment.

Alternative method to increase detection uniformity is to operate SPAD in an event-driven manner, also referred as asynchronous SPAD operation. In such scheme, photon detection renders SPAD insensitive for a predetermined deadtime, after which recharges for succeeding detections. In this work, a non-paralizable receiver is assumed. A non-paralizable receiver refers to a receiver which during a predefined deadtime t_d, SPAD is insensitive regardless of the photons impinging on the detector as shown in Figure. 3.7. Consequently, the current SPAD availability is determined by the time instance of the last photon detection of the SPAD. Note that, the lack of synchronicity to a reference signal means that detections towards the end of repetition periods result in an extended deadtime to the succeeding periods. As the result, SPAD recharging instance is uniformly distributed for a constant photon arrival rate λ, hence the detection probability is uniform for a constant photon arrival rate. As discussed, TCSPC system in a photon-abundant environment will most likely result in multiple photon detections per repetition period (assuming $t_d < t_r$). Thus, corresponding TDC with capability to produce multiple timestamps per repetition period is considered in this work.

3.4.3. Statistical model of Asynchronous operation

In the following, the detection probability of SPAD is statistically modelled based on [47] for constant λ and time-varying $\lambda(t)$.

Consider small observation time window δt, during which SPAD is either available for photon detection or insensitive. For simplicity, SPAD is assumed to recharge at the beginning of the available time window and instantly quenches upon detection for maximally one photon detection. Also, suppose a constant SPAD deadtime $t_d = m\delta t$ where $m \in \mathbb{Z}$. Let $\{X_n; n \geq 0\}$ be a random process $\{X_n, n = 0, 1, 2, ...\}$ with random variable X_n denotes the occupying state at time n within the countable state space $S = \{0, 1, 2, ..., m\}$. Then, the stochastic process of asynchronous SPAD receiver can be modelled as Markov chain as shown in Figure 3.8 State 0 represents available state where no photon is detected during δt. Thus, self transition probability follows Poisson distribution with discrete random variable K taking zero

$$Pr_{fail} = Pr(X_{n+1} = 0 | X_n = 0) = Pr(K = 0) = exp(-\lambda\delta t) \tag{3.15}$$

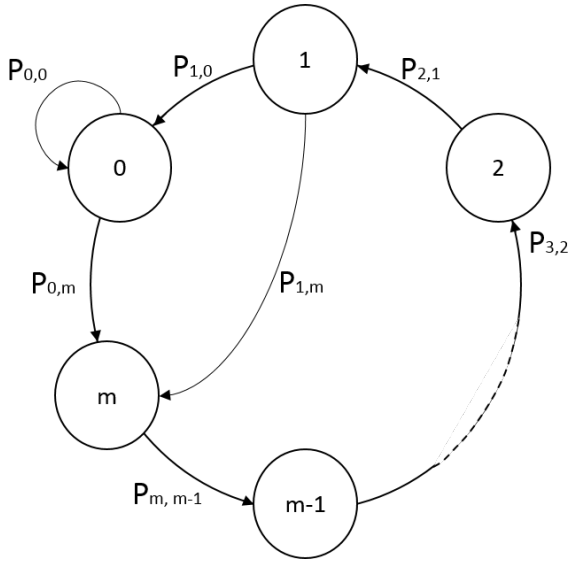


Figure 3.8: State diagram with possible stationary transitional probabilities

where K denotes the number of photon detections within δt. Upon photon detection, the state transit to the dead state n with conditional probability $1 - Pr_{fail}$. Then, SPAD is insensitive for the next $m-2$ states until $X_n = 2$, advancing to the succeeding state for every δt with transition probability of 1. State 1 represents an intermediate state, where zero photon detection leads to the available state $X_n = 0$ or the detection leads to the beginning of deadtime $X_n = n$. State 1 can be considered as an arbitrary state to represent the memory property of deadtime in a memoryless Markovian process. Thus, the transition probability at the n^{th} interval can be written as:

$$P_{1,0}^{(n)} = Pr(X_{n+1} = 0 | X_n = 1) = \frac{\sum_{i=1}^{k} Pr(X_{n-mi} = 0) \quad \times \quad P_{Pois}(K = i, \lambda\delta t)}{\sum_{i=1}^{k} Pr(X_{n-mi} = 0) \quad \times \quad (1 - P_{Pois}(K \leq i-1, \lambda\delta t)} \tag{3.16}$$

where $k \equiv \lfloor \frac{n-1}{m} \rfloor$, and P_{Pois} denotes Poisson distribution [47]. Summation in both numerator and denominator indicates that the transitional probability is dependent on the probability that SPAD was in available state at multiple integers proceeding it. The ratio between Poisson PDF and the inverted CDF limits the transitional probability to maximally 1. Obviously, the transition probability varies with n, thus is referred as non-homogeneous Markov chain. The dependence on n is removed by taking the limit $n \to \infty$

$$P_{1,0} = \lim_{n\to\infty} P_{i,j}^{(n)} = \frac{1 - \exp(-\lambda\delta t)}{\lambda\delta t}. \tag{3.17}$$

where $P_{1,0}$ denotes the steady state transitional probability of state transition from 1 to 0. The steady state implies that the long term behaviour of the chain does not change from one interval to another. Similarly, the long term probability of state occupancy, referred as stationary distribution, can be obtained. Let $1 \times (m+1)$ non-zero row vector $\pi = [\pi_0 \;\; \pi_1 \;\; \pi_2 \;\; \cdots \;\; \pi_m]$ denote the stationary distribution, then the following holds,

$$\pi = \pi P \tag{3.18}$$

where P is a $(m+1) \times (m+1)$ transition probability matrix (long term behaviour). The equation can be rewritten as $\pi(I - P) = 0$, where I denotes $(m+1) \times (m+1)$ identity matrix. The set of equations to be solved are

$$\begin{aligned} (1 - P_{0,0})\pi_0 - P_{1,0}\pi_1 &= 0 \\ \pi_1 - \pi_2 &= 0 \\ \pi_2 - \pi_3 &= 0 \\ &\vdots \\ (P_{0,0} - 1)\pi_0 + (P_{1,0} - 1)\pi_1 &= \pi_m. \end{aligned} \tag{3.19}$$

Additionally, considering Markovian property, $\sum_{i=0}^{m} \pi_i = 1$, the stationary distributions are given as:

$$\begin{aligned} \pi_0 &= \frac{1}{1+m\lambda\delta t} \\ \pi_i &= \frac{\lambda\delta t}{1+m\lambda\delta t}, \quad for \quad i = 1,2,...,m \end{aligned} \tag{3.20}$$

The histogram can be reconstructed using the probability that SPAD occupies state m, which follows binomial distribution. Let random variable Y represent the number of counts per histogram bin and follows binomial distribution with the number of measurement cycles n_c, the mean is given by,

$$E[Y] = n_c \frac{\lambda\delta t}{1+m\lambda\delta t}. \tag{3.21}$$

Alternatively, the detection probability for homogeneous Poisson process can also be approximated with Geometric distribution. Let random variable Z represents the number of failures before obtaining one success in a series of Bernoullii trials and $E[Z]$ denotes the mean value. Then, the detection probability is given by,

$$P_{det,\delta t} = \frac{1}{m + E[Z]} \tag{3.22}$$

Let photon detection probability follows Poisson process with zero photon detection according to eq.(3.15), then mean can be written as,

$$E[Z] = \frac{1}{1 - exp(-\lambda\delta t)} \quad \approx \quad \frac{1}{\lambda\delta t} \tag{3.23}$$

where mean is approximated using the first order Taylor series. Substituting eq.(3.23) into eq.(3.22) and multiplying the number of measurement cycles yields the expected histogram counts per bin,

$$\overline{C_{hist}} \approx n_c \frac{\lambda\delta t}{1+m\lambda\delta t} \tag{3.24}$$

which is equivalent to eq.(3.21).

The above derivation is only applicable for homogeneous Poisson process where photon arrival rate is constant over time, thus not applicable for TCSPC system. Nevertheless, it provides a useful insight into inhomogeneous Poisson process because the process reduces to Markovian process, as will be shown in the following derivation. Ultimately, we are interested in the detection probability of the individual histogram bin given time-variant photon arrival rate. That is, the stationary distribution for each histogram bin $\pi = [\pi_1 \;\; \pi_2 \;\; \cdots \;\; \pi_{b_{max}}]$ where $b_{max} \equiv T_{meas}/\delta t$. For simplicity, assume b_{max} is a natural number. The derivation follows the similar approach as above; process is reduced to Markovian process, steady state transitional probability is proposed, and stationary distribution is calculated.

Let $\{N(t), t \in [o,\infty)\}$ be a random counting process, where $N(t)$ denotes the number of event occurrences from 0 upto t, for all t within right half infinity. $N(t)$ increments at each occurrence time upto time t, denoted as $t \in J$ for $J = \{T_1, T_2, ..., T_{N(t)}\}$. Let $\mu(t, N(t); T_1, T_2, ..., T_{N(t)})$ define the conditional infinitesimal occurrence probabilities as:

$$\mu(t, N(t); T_1, ..., T_{N(t)}) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} Pr(N(t, t+\Delta t) = 1 | N(t); T_1, ..., T_{N(t)}) \quad for \quad N(t) \geq 1 \tag{3.25}$$

that is, the probability of single point occurrence in the infinitesimal interval $[t, t+\Delta t)$ given the number of events with all the occurrence times marked upto time t [48]. Intuitively, it is the momentary detection probability considering the detection history upto the time of consideration. Similarly, define the conditional infinitesimal probability that $(n+1)^{th}$ point occurs after time t as:

$$\mathscr{P}_{t_{n+1}|t_1,...,t_n}(t|T_1, ..., T_n) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} Pr(t_{n+1} \geq t | T_1, ..., T_{N(t)}) \quad for \quad N(t) \geq 1 \tag{3.26}$$

to represent the probability that there is no detection beyond time t before t_{n+1} given all the occurrence times marked upto time t. Thus, the probability is also referred as the conditional survival probability, which can be written with respect to eq.(3.25) as:

$$\mathscr{P}_{t_{n+1}|t_1,...,t_n}(t|T_1, ..., T_n) = exp\Big[-\int_{w_n}^{t} \mu(\tau, n; T_1, ..., T_n) d\tau \Big] \tag{3.27}$$

which simply reduces to Poisson pmf for $\mu(\cdot) = \lambda\delta t$ with integral replaced by sum ranging from initial point to the number of observing intervals [48]. As the result, the conditional probability density function for the $(n+1)^{th}$ occurrence time is given by multiplying the survival probability (eq.(3.27)) and the detection probability (eq.(3.25)) at time t,

$$\lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} Pr(t \le t_{n+1} \le t+\Delta t | t_1 = T_1, ..., t_n = T_n) = \mu(t, N(t); T_1, ..., T_{N(t)}) \times exp\Big[-\int_{t_n}^{t} \mu(\tau, n; T_1, ..., T_n) d\tau\Big]. \quad (3.28)$$

For non-paralyzable SPAD receiver, the infinitesimal detection probability described in eq.(3.25) is zero during dead time and thereafter follows photon intensity $I \equiv \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t}\lambda$. Mathematically,

$$\mu(t, N(t); T_1, ..., T_{N(t)}) = \begin{cases} \mathscr{I}(t) & for \quad t \ge T_n + t_d \\ 0 & for \quad T_n \le t < T_n + t_d. \end{cases} \quad (3.29)$$

Substituting eq.(3.29) into eq.(3.28) reduces the process to Markovian process as :

$$\begin{aligned} \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} Pr(t \le t_{n+1} \le t + \Delta t | t_1 = T_1, ..., t_n = T_n) &= \mathscr{I}(t) \times \exp\Big(-\int_{T_n + t_d}^{t} \mathscr{I}(\tau) d\tau\Big) \\ &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} Pr(t \le t_{n+1} \le t + \Delta t | t_n = T_n) \end{aligned} \quad (3.30)$$

where the equation holds for $t > T_n + t_d$.

As TCSPC repetitively transmits pulses with set interval t_r, the detection time can be written as a linear sum of multiple intervals and the remainder, $T_i = k_i t_r + x_i$ for $x_i \in [0, t_r)$. Let X_i be random variable denoting i^{th} detection time relative to the repetition interval, then eq.(3.30) can be written as:

$$\lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} Pr(t \le x_{i+1} \le t + \Delta t | x_i = X_i) = \frac{\mathscr{I}(x_{i+1})}{1 - \exp(-\Lambda)} \exp\Big(-\int_{X_i + t_d}^{c_{upper}} \mathscr{I}(\tau) d\tau\Big) \quad (3.31)$$

where $c_{upper} \equiv \lceil \frac{x_i + x_d - x_{i+1}}{t_r} \rceil t_r + x_{i+1}$ and $\Lambda \equiv \int_0^{t_r} \lambda(\tau) d\tau$ [49]. Intuitively, $c_{upper} = t_r + x_{i+1}$ for $t_i \le t < t_i + t_d$, that is for considered time t to be in the deadtime, then $(i+1)^{th}$ detection probability exists in the next repetition interval.

Using the presented continuous form, steady state transitional matrix can be approximated. Let Δt and n_b denote histogram bin width and equally spaced number of bins within the repetition period t_r, respectively. Let $P_{l,m}$ denotes steady state transitional probability of transition from l^{th} bin to m^{th} bin. Then, $P_{l,k} \equiv Pr(X_{i+1} = l | X_i = k)$ is found using eq.(3.31), yielding $n_b \times n_b$ transitional probability matrix P. According to the Markovian property,

$$\sum_{k \in \mathbf{S}_{n_b}} P_{l,k} = 1 \quad (3.32)$$

sum of all possible transitional probabilities from a single state, l, must be equal to 1, thus P is row-normalized, denoted as $\bar{P}$. Then, the stationary distribution is found by computing the eigenvector as described in eq.(3.18-3.19) [49].

3.4.4. Adaptive TDC gating scheme

As discussed earlier, the increase in photon throughput due to scaling and change of receiver operation poses challenges in both on-chip and off-chip processing, limited by chip area and data transfer bandwidth, respectively. Thus, this thesis proposes an adaptive TDC gating scheme to reduce photon throughput.

The method operates TDC only during the time interval where the reflected signal is expected to trigger at least one SPAD. Figure 3.9 presents an exemplary receiver block of such method where 4 SPADs are grouped to determine the expected signal time interval. Note that the size of the group pixel may be at least one SPAD.

Photon detections of each SPAD are registered individually in a coarse counter over multiple detection periods t_r, to yield coarse histograms as shown in Figure 3.9b. The individual coarse histogram of the SPADs are integrated and compared with an ambient noise count to determine the signal window. The ambient noise count is provided by another SPAD operating in the same receiver operation and covering FoV where target-reflected signal is absent. In the succeeding periods, TDC array may operate only during the determined signal windows to produce timestaps with high temporal resolutions. Note that, the coarse counter is simultaneously running in the background and may be subject to filtering to provide update target bins

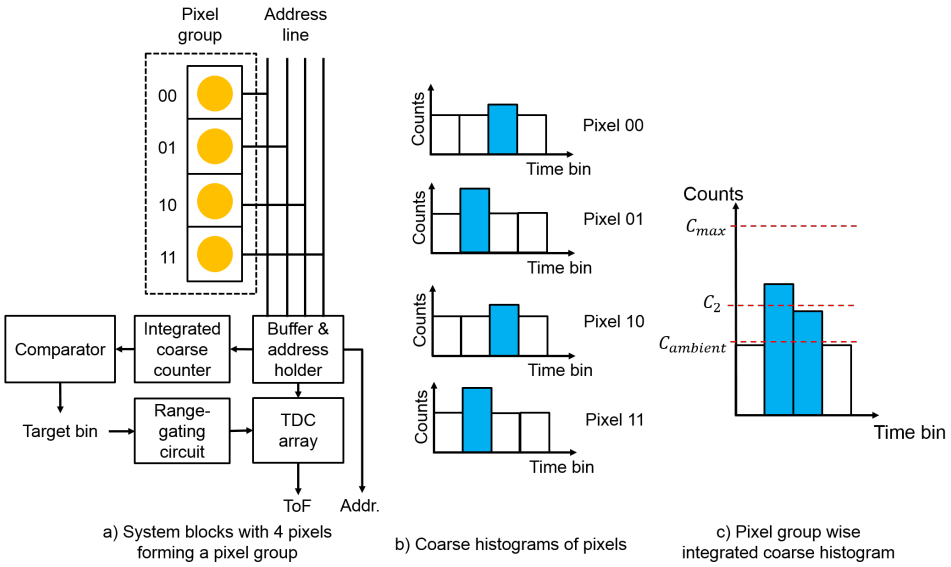


Figure 3.9: Exemplary receiver block of adaptive TDC gating scheme and target bin identification.

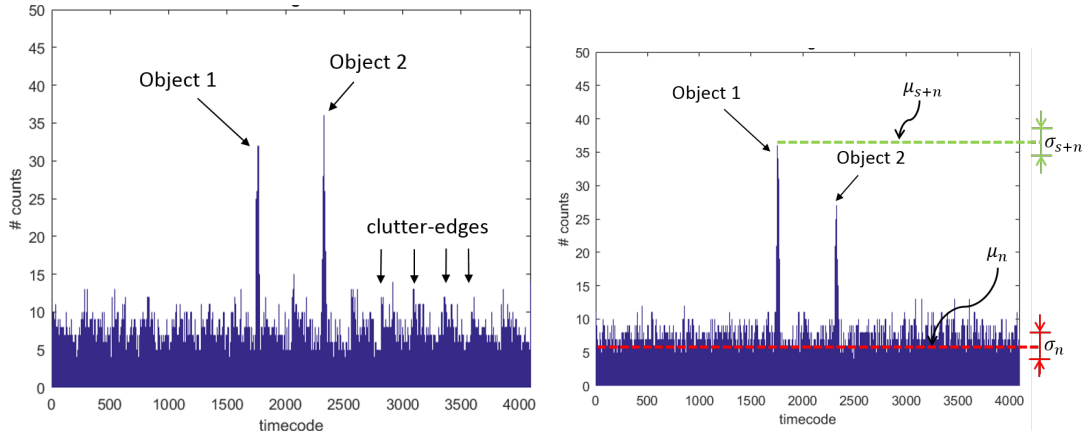
for the next cycle. The reduced TDC operation intervals achieves data compression by registering less timestamps with high temporal resolution.

In ocntrast to the zoom-in approach presented in [13], adaptive TDC gating retains the frame rate. In other words, each histogram is constructed with more counts from longer measurements which translates to an improved detection range. Furthermore, the proposed method reduces the data throughput significantly at the cost of full-range histogram reconstruction. Thus, the proposed method is advantageous for a system where reconstruction of the received signal is not the goal. In case the constant noise level is of interest, the ambient noise count may provide the information.

3.5. Simulated results

3.5.1. Histogram and estimation

Figure 3.10 shows the simulated histogram of TCSPC system using the interleaved gating scheme with square waveforms as shown in Figure 3.6. Using the simulation parameters found in appendix A.1, the histogram was constructed over 5000 laser repetition period over the range of 150m with histogram bin period of 200 psec. Each laser pulse period was divided into 16 ON/OFF SPAD gating windows with varying phases. Figure



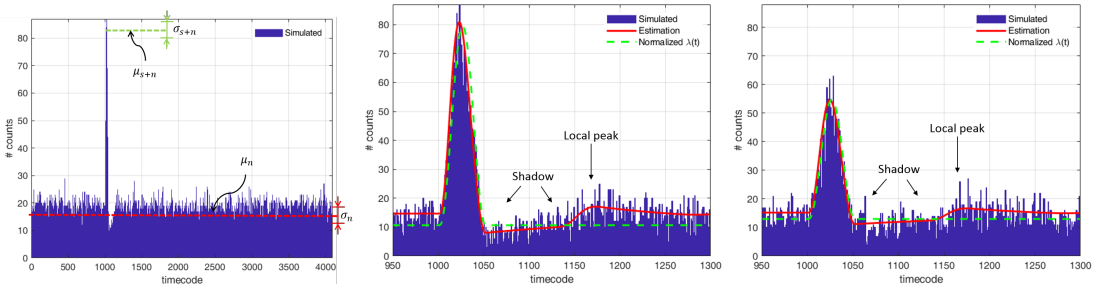
(a) Coarse phase-shifting : $\phi = \pi$ (b) Fine phase-shifting : $\phi = \pi/4$

Figure 3.10: Simulated histogram of interleaved gating operation with 16 gating windows, $T_{meas} = 5000 \times t_r$, and $T_{bin} = 200$ psec

3.10a shows the simulated histogram with phase-shift with $\phi = \pi$. The presence of cluster-edges are clearly for each time window with exponentially decreasing photon counts over distance according to eq.(3.3). As the result, the received white noise with constant photon rate is reconstructed as a sawtooth wave with peaking counts towards the beginning of each window, implying non-uniform photon detection probability. Such non-uniformity causes the photon counts of the target-reflected pulses to depend on the temporal position relative to the windows position. The simulated object 1 and 2 shown in Figure 3.10a are located $52m$ and $69m$ away from the transmitter, respectively. For the described simulation set-up, object 1 is located towards the end of a gating window, whereas object 2 is located towards the beginning of another gating window. Consequently, despite the higher received photon rates of object 1 according to inverse square law, object 1 records lower photon counts than object 2. The combination of the cluster edges and reduced photon counts of the target-reflected pulses due to localized pile-up effect significantly reduces the detection range.

The skewed reconstruction is improved by increasing the uniformity of detection probability. Figure 3.10b depicts histogram with finer phase-shift of $\phi = \pi/4$. The finer phase-shift reconstructs white noise more uniformly and the target-reflected received signals more accurately, in comparison to the coarser gating-window shifting. The resulting more accurate representation of the target-reflected received signal is a strong evidence of improved detection probability.

Figure 3.11 shows the constructed histogram of TCSPC system using the asynchronous gating scheme as explained in section 3.4.2. The histogram is constructed with 5000 laser repetition periods to detect the object 30m afar with simulation parameters found in appendix A.1.3. Figure 3.11a depicts randomly varying



(a) Over the whole range with $d_{object} = 30$m

(b) Zoomed view of (a) around received pulse

(c) Zoomed view around received pulse with $P'_{T_x} = \frac{P_{T_x,0}}{2}$

Figure 3.11: Simulated histogram of asynchronous SPAD operation with $T_{meas} = 5000 \times t_r$, $T_{dead} = 25.6$ nsec, and $T_{bin} = 200$ psec

noise counts due to random white noise from the background source, thus proving the detection uniformity of asynchronous operation as formulated in eq.(3.21) and eq.(3.24). The signal counts and the following noise counts within the time windows of received signal T_{W,R_x} and the deadtime $T_{W,dead}$ follows eq.(3.31). Accordingly, the estimations are numerically calculated and are in good alignment with the constructed histograms as shown with solid red lines in Figure 3.11b and 3.11c. The received photon rates $\lambda(t)$ is normalized to the peak of the estimated histogram count as a reference, and is not to scale.

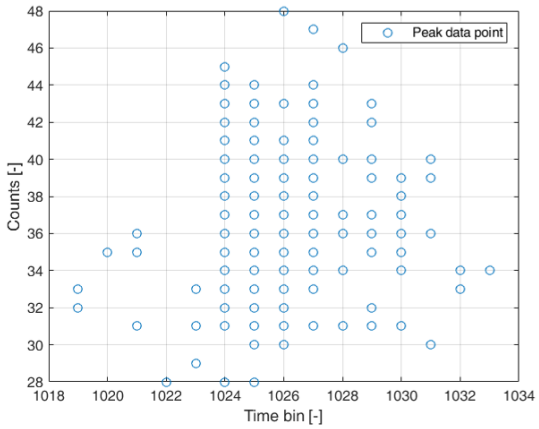
The bins succeeding the signal window count low time stamps with respect to mean noise counts μ_n as SPADs at the suceeding bins are likely to be dead from the detections in the signal window, the phenomenon referred here as shadowing effect. The shadow effect is more pronounced in Figure 3.11b than Figure 3.11c due to higher received photon rates. Intuitively, the dip recovers and achieves local peak at the bin approximately SAPD deadtime later the estimated signal peak.

3.5.2. Reliability of retrieved ToF

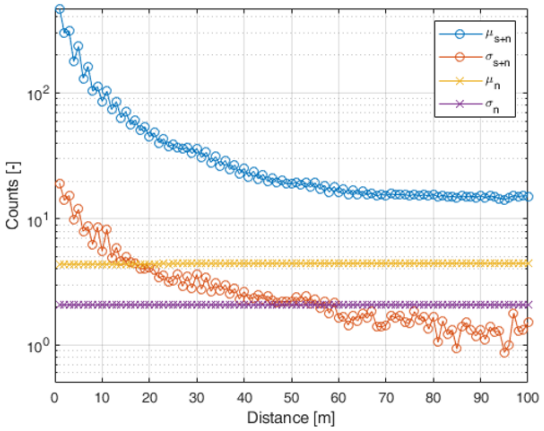
Although the presented estimation accurately estimates the trend of the constructed histogram, the probabilistic nature of photon detection requires Monte Carlo method to evaluate the confidence level of the retrieved ToF over distance. The distribution of peak counts in terms of mean μ_{s+n} and standard deviation σ_{s+n} and of noise counts in terms of mean μ_n and standard deviation σ_n are recorded, as shown in Figure 3.10b & 3.11a. Accordingly, SNR is defined as

$$SNR = \frac{\mu_{s+n} - \mu_n}{\sqrt{\sigma_{s+n}^2 + \sigma_n^2}} \tag{3.33}$$

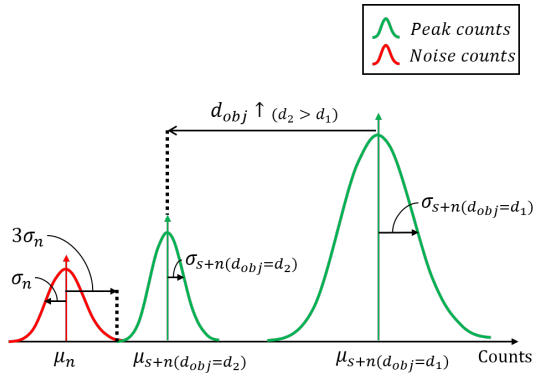
and the retrieved ToF is defined reliable when the retrieved ToF falls within the received signal window T_{W,R_x}. Figures 3.12-3.13 show the simulation results of introduced interleaved operation and asynchronous SPAD operation, collected over 300 histograms, where exemplary histograms are shown in Figure 3.10b and Figure 3.11a.



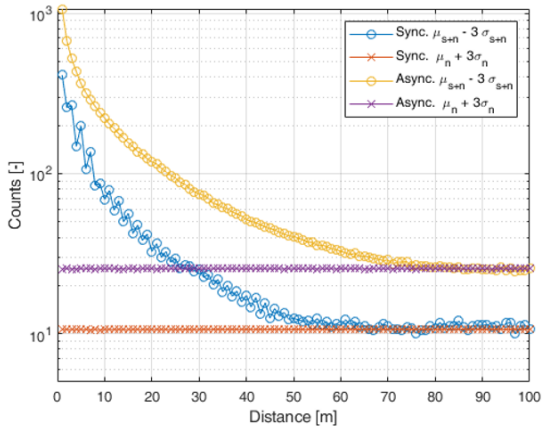
(a) Peak counts and the corresponding time bins for $d_{object} = 30m$ in interleaved gating scheme



(b) Mean and standard deviation of the peak counts and noise counts in interleaved gating scheme



(c) Peak and noise counts distribution over distance

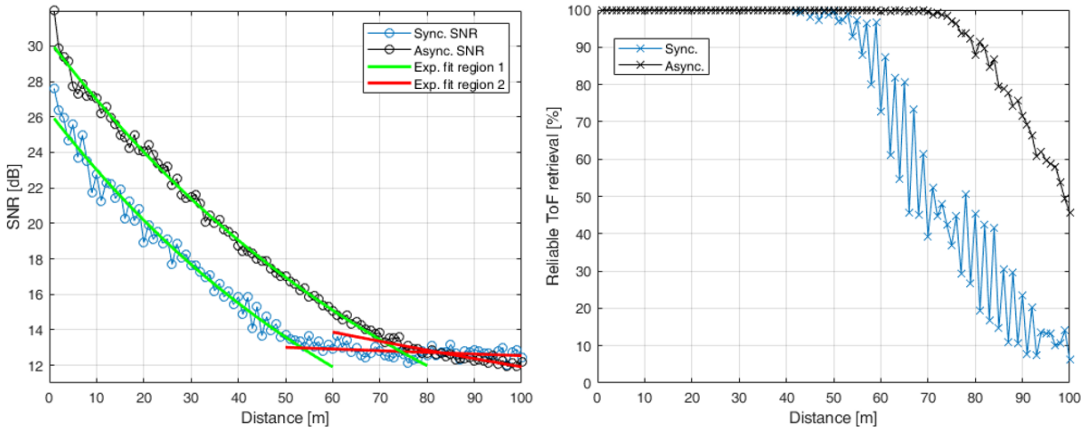


(d) Converging peak and noise counts in interleaved gating scheme (Sync.) and asynchronous scheme (Async.)

Figure 3.12: Distribution of simulated histogram counts

Figure 3.12a shows simulation results of Monte Carlo simulation where the peak counts and the corresponding time bins are recorded for object distance $d_{object} = 30m$, equivalently the bin number $n_b = 1024$. Each peak count consists of counts arising from both target-reflected signal and white noise. The distribution of peak counts is represented in terms of mean and standard deviation, denoted as μ_{s+n} and σ_{s+n}, respectively.

Figure 3.12b depicts the distribution characteristics for both peak counts and noise counts over object distance. As shown, μ_{s+n} and σ_{s+n} demonstrates a saw-tooth pattern on inverse proportional envelope with distance following decrease in received power. Such pattern is the result of varying detection probability due to relative disposition between the object and the beginning of shifted gating windows. The mean and standard deviation of noise counts are constant, indicating the achieved detection uniformity by fine phase gating. The granularity of the phase can be traded-off with signal peak counts. μ_{s+n} does not converge to μ_n for negligible received signal power in comparison to white noise, in which case peak counts are dominated by the standard deviation of noise σ_n. That is, μ_{s+n} and σ_{s+n} decrease inverse proportionally upto a distance where $\mu_{s+n} - 3\sigma_{s+n} \approx \mu_n + 3\sigma_n$ holds, from which μ_{s+n} represents the noise peaks, thus does not converge to μ_n for $\sigma_n \neq 0$, as shown in Figure 3.12c. In Figure 3.12d, such transitional distance of convergence for interleaved and asynchronous operations are recorded as 54m and 71m, respectively.



(a) SNR over distance

(b) Successful retrieval rate over distance

Figure 3.13: SNR and Reliability of the retrieved ToF over distance in interleaved gating scheme (Sync.) and asynchronous scheme (Async.)

Fitting exponential functions through SNR data as shown in Figure 3.13a, where the data is divided into two regions depending on the dominant source the peak counts arise from, clearly visualizes the transition points by means of intersections, thereby assists the estimation. Comparatively, asynchronous operation improves SNR by 4dB in region 1 with respect to the interleaved gating scheme. SNR and the transitional distance measures 12.98 dB and 53.5m in interleaved gating scheme and 70.9m and 13.31 dB in asynchronous operation. The found transitional distances are in good alignment with the converging distances shown in Figure 3.12d, confirming the transition of dominant source of counts.

Transitional distances provide useful information in evaluating the reliability of the retrieved ToF. Referring to Figure 3.13b, the reliability rate drops below 96% at the measured transitional distances and begins to deteriorate. Interleaved gating scheme exhibits saw-tooth pattern while deteriorating as the result of the varying peak counts caused by the relative disposition between the object and the beginning of the gating windows. A finer phase shifting would reduce the variation, but the confidence level begins to deteriorate at an earlier distance due to the reduced peak counts. Comparatively, asynchronous operation measures the same reliability rate at 74m, extending the detection range of the system by 20m.

3.5.3. Photon count throughput

The improved SNR achieved by asynchronous SPAD gating scheme, thereby extending the detection range, comes at the cost of increased photon throughput. Figure 3.14 and Figure 3.15 illustrate simulated mean photon counts per pixel per detection range over object distance for different gating schemes. Note that, photon counts of each Monte-Carlo simulation are recorded over 5000 laser repetition period over the range of 100m with histogram bin period of 200 ps and using the parameters found in appendix A.1.

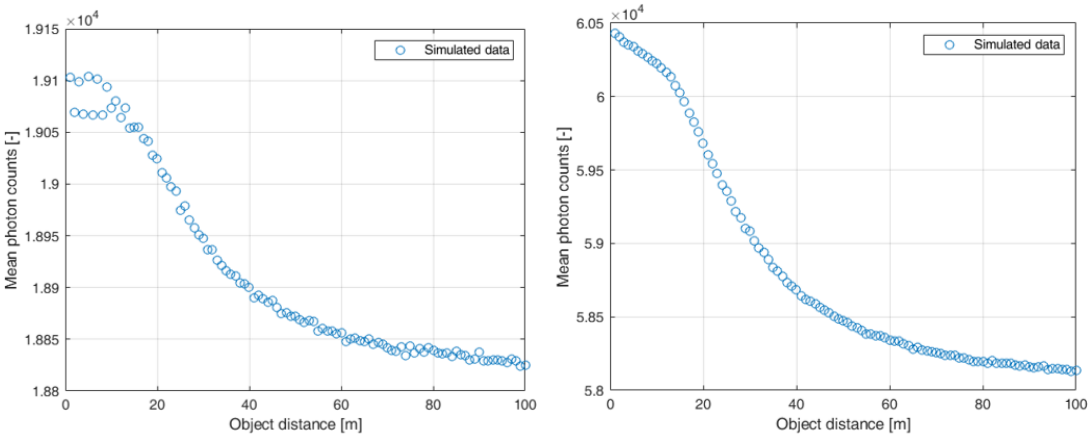
Figure 3.14a and Figure 3.14b show the simulated mean photon counts obtained using interleaved SPAD gating scheme and asynchronous SPAD gating scheme, respectively. Comparatively, asynchronous scheme records about 3 times more photons than interleaved scheme, achieving maximally 13 photons per detection range for $T_{SPAD,dead} = 2^7 \times 200$ps. Such high photon counts pose a serious challenge in both off-chip and on-chip processing, when SPAD array is scaled.

For off-chip processing, the required readout bandwidth to transfer timestamps to an external device is calculated as:

$$BW_{req} = N_{pixels}\overline{C_{ph,pixel}}N_{meas,frame}N_{frames,sec}N_{bits,ph} \tag{3.34}$$

where $N_{pixels}, \overline{C_{ph,pixel}}, N_{meas,frame}, N_{frames,sec}$, and $N_{bits,ph}$ denote number of pixels in SPAD array, mean photon counts per pixel per detection range, number of measurement cycles per frame, number of frames per second, and number of bits per photon count, respectively. Substituting the obtained simulated results $\overline{C_{ph,pixel}} = 13$ into eq.(3.34), the required bandwidth is approximately 112 GB/s. Note that the computed required bandwidth is a minimal bandwidth and increases with additional information such as flags and pixel address information, rendering off-chip processing impractical.

Alternatively, on-chip histogramming can be considered, in which case the required memory is calculated



(a) Interleaved SPAD gating scheme

(b) Asynchronous SPAD gating scheme

Figure 3.14: Simulated mean photon counts per pixel over object distance

as:

$$M_{req} = N_{pixels} N_{bins,pixel} N_{bits,bin} \tag{3.35}$$

where $N_{bins,pixel}$ and $N_{bits,bin}$ denote number of time bins per pixel and number of bits to represent photon counts per bin, respectively. Considering a memory representing the number of photon counts must be able to store maximal photon counts per bin, the required minimal bits per bin is computed as: $N_(bits, bin) = log_2(Max(C_{bin}))$. Using the above-mentioned simulation parameters, the required memory is approximately 235 MB.

Figure 3.15 shows the simulated mean photon counts obtained using asynchronous SPAD gating scheme in conjunction with adaptive TDC gating scheme as presented in Figure 3.9. The simulation set-up to obtain data shown in Figure 3.14 is used. Intuitively, adaptive TDC gating scheme reduces total photon counts over a

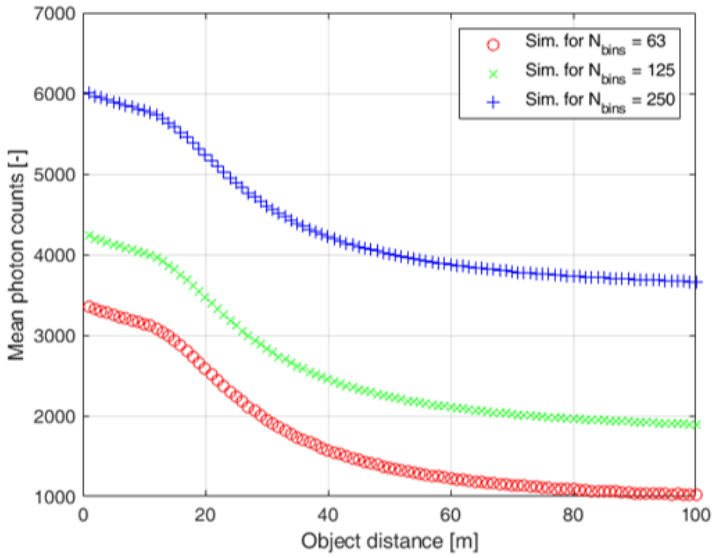


Figure 3.15: Simulated mean photon counts per pixel over object distance for asynchronous SPAD gating scheme in conjunction with adaptive TDC gating scheme for varying TDC gating window sizes ($N_{bins,W}$)

dynamic range by recording only a fraction of time bins. In comparison to the system operating asynchronous SPAD gating without adaptive TDC gating scheme, the proposed adaptive TDC gating scheme compresses data by a factor of α given by,

$$\alpha = \beta \frac{N_{bins,W}}{N_{bins,DR}} \tag{3.36}$$

where β, $N_{bins,window}$, and $N_{bins,DR}$ denote a compensation factor, number of bins recording photon counts, and number of bins per dynamic range, respectively. The compensation factor is dependent on the received pulse amplitude and asynchronous SPAD gating window size, and $1 < \beta < 3.5$ is obtained in this experiment. The corresponding required memory of the proposed adaptive TDC gating system is calculated as:

$$M_{req} = N_{pixels}\Big(N_{bins,W}N_{bits,bin} + N_{coarsebin}N_{bits,coarsebin}\Big) \tag{3.37}$$

where, $N_{coarsebin} := \frac{N_{bins,DR}}{N_{bins,W}}$ denotes number of coarse bins defined by the ratio between the number of bins within the dynamic range and the number of bins recording photon counts. For various TDC gating window sizes $N_{bins,W}$ = 63, 125, and 250 bins, the required memory is approximately 8, 10, and 16 MB. That is, the required memory of the proposed adaptive TDC gating system reduces required memory size by a factor of at least 14 with respect to the system operating asynchronous SPAD gating without adaptive TDC gating scheme.

3.6. Conclusion

Basic principles of TCSPC system is presented and the limitations thereof are discussed. The conventional receiver operational methods assume and successfully reconstruct the returned pulses in a photon starved condition. Otherwise, pile-up effect renders the LiDAR system impractical for long range detection.

Time-shifting the interleaved SPAD gating waveform extends the detection range, but the relative temporal disposition between the object and the beginning of gating windows yields cluster-edges due to local pile-up effect. Moreover, such relative dependency results in disproportionate peak counts which reduces reliability of peak detection method. In contrast, asynchronous SPAD operation achieves uniform detection probability with the exception of bins succeeding the received pulse signal. The presented statistical model of asynchronous operation is in good alignment and estimates the general trendline accurately as shown in Figure 3.11b. The reliability of retrieved target distance is studied using Monte Carlo method due to the probabilistic nature of the photon detection.

Considering the peak counts within the signal window, Figure 3.12d depicts converging functions, indicating the transition of dominating sources of peak counts. SNR and the transitional distance measures 12.98 dB and 53.5m in interleaved gating scheme and 70.9m and 13.31 dB in asynchronous operation. Comparatively, asynchronous operation improves SNR by 4dB in region 1, where peak counts consist of both signal and noise, with respect to the interleaved gating scheme.

Referring to Figure 3.13b, the reliability rate drops below 96% at the measured transitional distances and begins to deteriorate. Furthermore, interleaved gating scheme follows a saw-tooth pattern while deteriorating due to relative dependency between the object distance and gating window disposition. A finer phase shifting reduces the variation according to eq.(3.14), but the confidence level begins to deteriorate at an earlier distance due to the reduced peak counts. Comparatively, asynchronous operation measures the same reliability rate at 74m, extending the detection range of the system by 20m. The simulation result presents the potential of the asynchronous SPAD gating as a receiver operation of a TCSPC LiDAR system in a photon abundant environment.

The improved SNR achieved by asynchronous SPAD gating scheme, thereby extending the detection range, comes at the cost of increased photon throughput. Comparatively, asynchronous scheme records about 3 times more photons than interleaved scheme, achieving maximally 13 photons per detection range for $T_{SPAD,dead} = 2^7 \times 200$ps. Such photon throughput requires bandwidth is approximately 112 GB/s for off-chip processing or 235 MB of on-chip memory, rendering the system highly impractical. The proposed adaptive TDC gating scheme utilizes coarse-fine structure, thus achieves reduction in the required memory size by a factor of at least 14 with respect to the system operating asynchronous SPAD gating without adaptive TDC gating scheme. In light of the reduced throughput, consequently reduced storage, adaptive TDC gating scheme is further considered in the following chapters.

4

Time-to-Digital Converter

4.1. Introduction

This chapter presents a timing unit design for the LiDAR system of Chapter 3 which implements asynchronous SPAD gating in conjunction with TDC gating methods. As discussed, outdoor application of such scalable system imposes several challenges, in particular on-chip data processing, thus rendering a coarse-fine timing structure to be suitable. That is, an external PLL source provides global coarse timing references and per-pixel TDCs provide the respective local fine timing references for each pixel. Therefore, this chapter focuses on designing per-pixel TDC, in particular a ring oscillator based TDC, for a coarse-fine timing structure. Table 4.1 lists the corresponding target specification studied.

Parameter	Value	Unit	Condition
LSB	50	$psec$	Propagation delay of a cell ($t_{LH} = t_{HL}$)
Dynamic range	14	bits	123m ToF range (in total using a coarse-fine structure)
f_{EXT}	80	MHz	External TDC control signal (single phase)
TDC range	7	bits	Covers 1 PLL period
Conversion rate	1	MSamples/sec	-
DNL	1	LSB	-
INL	1	LSB	Full range
Power	100	μW	Saturation (<3.2W in total of 320×100 array)
Area	2500	μm^2	Incl. memory in case of on-chip histogramming

Table 4.1: Target specification of a ring oscillator based TDC with an external PLL

LSB size of 50 ps translates to minimally distinguishable distance of 1.5 cm assuming that transmitted pulse travels at the speed of light. In conjunction with the LSB size, dynamic range of 14 bits approximately translates to approximately 123m of maximally resolvable distance taking into account the pulse reflection on the object. Conversion rate of 1 Msamples/sec takes into account auxiliary circuit delays for every laser pulse period. Integral and differential non-linearities (INL and DNL, respectively) requirement is 1 LSB size for the full range. Power consumption requires below $100\mu W$ assuming the TDC generates a single time stamp for every detection period. A pixel size of $50 \times 50\mu m^2$ is assumed which should comprises at least a memory block and a TDC unit.

Section 4.2 presents an overview and implementations of the state-of-the-art TDCs in various LiDAR systems. In section 4.3 the ring oscillator based TDC design is explored with various delay cell topologies. Furthermore, timing error sources are discussed and clock distribution power is estimated. In section 4.4, the simulation results correlate the decisions from circuit to system level through evaluation of performance parameters. The performance parameters of the delay cells are compared and the optimal topology is suggested.

4.2. State-of-the art

4.2.1. Overview of Time-to-Digital Converters

A fair comparison of any data converter requires a careful consideration of performance evaluation quantity known as Figure-of-merits (FoM). FoM for a TDC is not agreed upon, thus varies for different architecture and applications. Nevertheless, similar to the standardized FoM for an ADC such as Walden FoM, all proposed FoMs for a TDC take into account conversion power, frequency, and linearity. However, specific to a TDC, the minimum achievable gate delay benefits from lower technology node, which conventional FoM fails to capture. Arguably, the merits of reduced gate delay is represented by the increased sampling frequency. However, TDC architectures such as Vernier are implemented specifically to achieve sub-gate delay for given technology node trading the power as well as the area. Therefore, in this thesis, a new normalized FoM is proposed based on the FoM model presented in [50] as the following:

$$\overline{FoM_{TDC}} = \frac{P}{2^{N_{lin}} \times f_s} \times \frac{LSB}{1e^{-12}} \quad [\frac{pJ}{Conv.-step}] \tag{4.1}$$

where P and f_s refer to dissipated power and sampling frequency, respectively. $N_{lin} := NOB - log_2(INL+1)$ refers to the number of linear bits which takes linearity into account. Note that N_{lin} is not equal to the effective number of bits ($ENOB$) as proposed in [51] which relies on the spectral power components of signal, noise, and distortion. The difference between N_{lin} and $ENOB$ arise from the measurement methods. Further note that in case of an event-driven TDC, dissipated power is normalized to half of the TDC dynamic range to represent the averaged power over a uniformly distributed input.

$ENOB$ of an ADC is calculated using signal-to-noise-and-distortion ratio ($SNDR$), where power of each component is measured using power spectral density in frequency domain. The power spectral density is the transformed spectrum of the transient waveform of the converted output waveform using fast Fourier transformation (FFT), given the sinusoidal input to the converter in time domain. Similarly, one characterizes dynamic behaviour of a TDC with an offset dc level and sinusoidal signal variation in time domain. The transient output code can be transformed to frequency domain to calculate the integrated noise level. Whilst dynamic measurement calculates integrated noise, thereby successfully captures the dynamics of the converter, it is hard to extract a relevant parameter for direct ToF applications. Instead, the linearity and the statistical output code distribution provide better insights in precision of the TDC output.

Type	Code density test (static)	Single Shot precision	Dynamic
Input	Random delay δT	Constant δT	Time-domain sinusoidal
Measurement	Scanning & repetition	Repetition	FFT
Output [y-axis, x-axis]	DC transfer characteristics [Code, T_{in}]	Histogram [Counts, output code]	Power spectral density $[\frac{dBps^2}{Hz}, Hz]$
Parameter [Unit]	Linearity (INL & DNL) [LSB]	σ [-]	Integrated noise $[Sec_{rms}]$

Table 4.2: TDC characterization techniques

The characterization method to obtain differental (DNL) and integral (INL) non-linearity for TDCs resembles of ADCs. TDC receives an input of which the timing delay from the reference is randomly distributed within the dynamic range and converts to digital output, which are stored in histogram. The process is repeated to yield histogram, thus referred as code density test. Ideally, distribution is uniform, but the actual histogram shows fluctuation of codes. Accordingly, the code width for each bin is calculated and compared with ideal linear transfer function to yield INL and DNL. Linearity provides insight in an averaged code error for a given input time delay, but omits the statistical distribution of the output code. The same code density approach can be used achieve distribution, but with a constant delay as an input to calculate the precision. The process is known as single shot precision (SSP) and is used widely to characterize TDCs for LiDAR applications. However, SSP is not considered in FoM_{TDC} as the parameter is application specific. The introduced three TDC characterization methods used in TDCs are summarized in table 4.2

As the technology node advances, a minimally achievable delay decreases which in turn may have an impact on power consumption. For instance, instead of differential delay lines to achieve a short delay, a single

delay line at advanced node can be implemented. Furthermore, the minimally achievable delay may also decrease LSB, thereby allowing lower proposed FoM. Thus, $\overline{FoM_{TDC}}$ according to eq.(4.1) over technology node is shown in Figure 4.1 for different architectures.

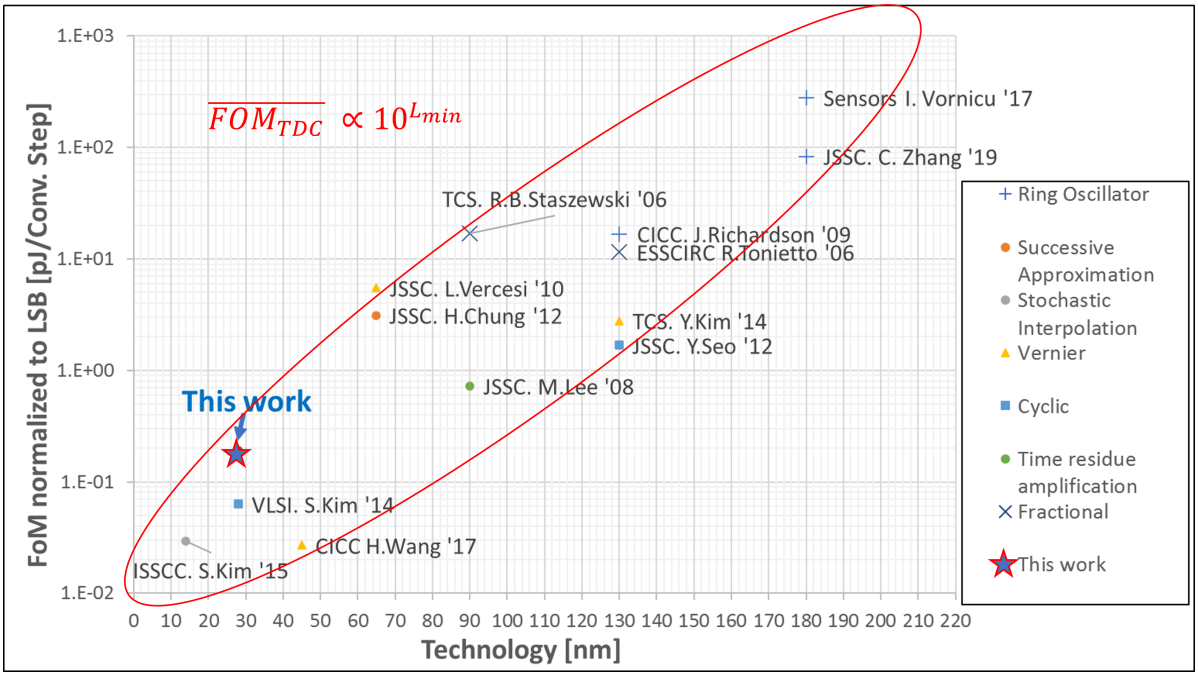


Figure 4.1: Normalized FoM_{TDC} over Technology node

First order approximation suggests a logarithmic relation between the normalized FoM and technology node (equivalently, minimally achievable gate length). Such trend clearly indicates the benefit of achieving low normalized FoM with lower technology node due to reduced gate delay.

Note that $\overline{FoM_{TDC}}$ in Figure 4.1, in particular ring oscillator based TDCs implemented in an array, denotes the performance results of the individual timing unit. $\overline{FoM_{TDC}}$ of this work presents the performance results of the 7-bit ring oscillator based TDC based on the simulation results, as presented in section 4.4. When considering the 14-bit coarse-fine structure and dividing the power consumption of the external PLL by the number of pixels, to account for the consumed power to cover the dynamic range for each pixel, $\overline{FoM_{TDC}}$ is expected to achieve 0.0027 [pJ/Conv.step] per pixel. The detailed PLL power estimation is given in section 4.3.1.

Another important factor to consider is area. Assuming 3D integration with SPAD on top layer and the time stamping logics on the bottom layer, the pixel area is $2,500\mu m^2$, of which the actual area allocated for TDC is reduced by memory blocks to enable on-chip histogramming. Although area and architectures are not considered in eq.(4.1), $\overline{FoM_{TDC}}$ is considered over area for different architectures, in the interest of miniaturization. Figure 4.2 depicts normalized FoM over area for different architectures.

The red box in Figure 4.2 denotes a target region and is based on the specification in Table 4.1. Assuming a coarse-fine structure, the target $\overline{FoM_{TDC}}$ is:

$$\overline{FoM_{TDC}} < \frac{100\mu}{2^{14-log_2(1+1)} \times 1M} \times \frac{50p}{1p} \approx 0.61 \left[\frac{pJ}{conv.-step}\right] \tag{4.2}$$

which includes the PLL power contribution for the coarse timing resolution.

The limited area eliminates area-consuming architecture candidates such as delay-line and Vernier. Vernier TDCs achieve sub-gate delay by differentiating the delays from different delay lines. However, the relative mismatch between delay elements reduces linearity. The area inefficiency can be alleviated by increasing the dimensions into 2D and 3D, but the number of auxiliary blocks to compare multiple delay lines grow simultaneously [50, 52, 53]. More importantly, the required LSB is greater than the minimum achievable gate delay in 28nm technology (order of a picoseconds), thus is not considered in this thesis.

Architectures such as cyclic or successive approximation shown in Fig.4.1 form a loop around delay elements and generally comprise a decision logic, a time adder, a subtracter, and a timing generator [54–56]. At

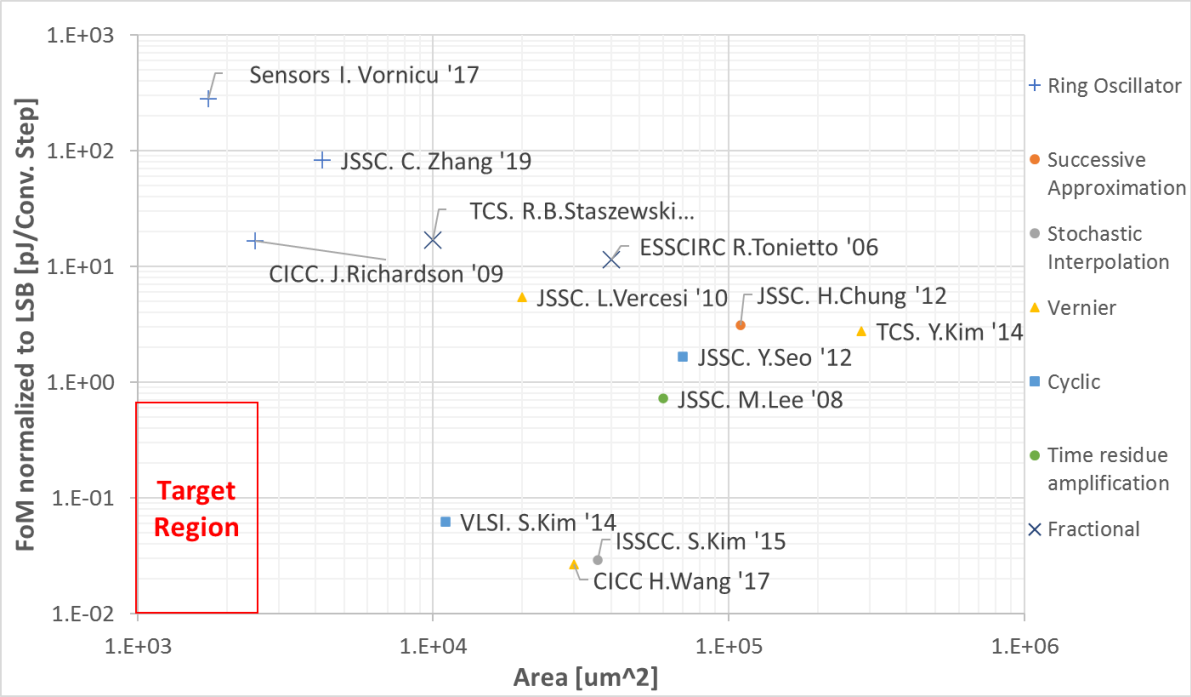


Figure 4.2: Normalized FoM_{TDC} over area

every iteration, the decision logic compares signal to the reference and determines its location. Accordingly, time is either added or subtracted to the reference signal, the reference is produced through timing generator, and is fed to the decision logic in the next iteration. The dynamic range is extended through the number of iterations, making the technique ideal for long range applications. However, the control logic and timing generators scale with linearity, thereby limits an efficiency. Moreover, a wrong decision at an earlier iteration leads to a maximum error of half a detection range which does not satisfy the linearity requirement for an automotive application.

Along with the advancement with technology, the delay cells become more prone to PVT, noise, and mismatch. In such cases, conventionally, a calibration method holds a key to achieve high linearity with fine resolutions. [57] suggests an approach to exploit the inherent statistical uncertainties of delay units and interpolate to convert timing, known as stochastic interpolation. The technique purposely adds redundant delay elements to eventually achieve a uniform distribution of timing edges, where the input δT is calculated simply by counting the number of edges and truncating LSBs to yield required LSB. Counter-intuitively, linearity improves with the number of delay elements, mismatch, and jitter, and renders the system calibration-free. Nevertheless, the redundancy and the counter size pose serious challenge in miniaturization, especially for high dynamic range.

Ring oscillator (*RO*) based TDCs generally have a very small form factor as the minimal delay elements with a counter suffice. The counter registers transition with oscillations and extend dynamic range through the number of bits, making it an excellent candidate for high dynamic range applications. Nevertheless, ring oscillator accumulate jitter periodically with oscillations, thus require calibration to synchronize frequency. In section 4.2.2 the reference RO TDCs are discussed in terms of design choices from system level to delay cell.

4.2.2. Ring oscillator based TDCs in LiDAR

Conventionally, the frequencies of TDC array in a LiDAR system is calibrated through reference signals generated by external PLLs [58, 59]. Alternatively, neighbouring ring oscillators in TDCs may be mutually coupled to synchronize, the approach is referred as mutual coupling [60]. The mutually coupled oscillators correct deviation in phases through local feedback system and lock the frequency to achieve synchronization. The frequency locking allows PVT calibration with external PLL through any single TDC of the array, reducing the distribution power. Furthermore, in comparison to an event-driven ring oscillators, always-running ring oscillators records reduced IR-drop that is caused during initialization phase.

A RO-based TDC consists of a ripple counter to convert oscillation periods to timing. Elongating the counter bits extend the dynamic range, making the architecture area efficient for long range. Finer resolution is achieved by each delay unit of oscillator driving individual counter upon reference signal to store the states in thermal code. The code can be converted into binary code and is combined with counter bits to provide converted timing. Inherently, counters do not register in the absence of transition leading to phase error of maximally 1 LSB. A conventional TDC for LiDAR, after each conversion, resets the oscillator by discharging the nodes, for the succeeding measurement cycle. Alternatively, quantization error can be 1st order noise shaped by preserving the phase during off-time and start the oscillation from the held phase [61]. The input phase is given by:

$$\Phi_{IN}[n] = 2\pi D_{out}[n] + \Phi_Q[n] - \Phi_Q[n-1] \tag{4.3}$$

where $\Phi_{IN}[n]$, $\pi D_{out}[n]$, and $\Phi_Q[n]$ denote input phase, digital output code, and quantization phase at n^{th} sample, respectively. As a result, the noise shifts to higher frequency and in-band phase noise decreases, thereby yielding a higher effective resolution when fed through low pass filter. However, the leakage or charge injection may introduce phase error and the physical delay in start and stop of the oscillator introduces dead-zone. The dead-zone is eliminated by switching to lower supply to run the oscillator at a lower frequency during the hold period [62]. The effectiveness of noise shaping technique is evident in a dynamic system such as a frequency synthesizer based on a fractional PLL, where the TDC is placed in a loop to form a correlation between consecutive conversion samples. However, the correlation between samples is, in principle, not preconditioned in a static system such as LiDAR, hence noise shaping does not improve the static measurements results.

The timing uncertainty of individual delay unit accumulates with propagation, resulting in non-linearity. When jitter dominates, the uncertainty grows in power domain assuming uncorrelated variables, that is the noise arising from the preceding delay unit is statistically independent of the succeeding unit [63]. The growth is independent of the number of the delay units within the ring oscillator loop, thus minimal number of delay units to enable oscillation are used. As the channel length decreases, the transistors are more susceptible to mismatch which leads to timing error [64]. Thus, section 4.3 investigates the accumulation of mismatch-dominated noise in a ring oscillator empirically. Four prevailing architectures of delay units are compared with respect to area, power, supply sensitivity, jitter, and mismatch. Accordingly, the optimal delay unit architecture and the number of delay stages are suggested.

4.3. Ring oscillator-based TDC design in 28nm

A ring oscillator consists of N inverting delay cells. The oscillation frequency is inversely proportional to the time taken for signal to propagate around the loop twice

$$f_{osc} = \frac{1}{2Nt_p} \tag{4.4}$$

where t_p denotes propagation delay of each cell, which is defined as the average of rise time t_{LH} and fall time t_{HL}

$$t_p = \frac{1}{2}(t_{LH} + t_{HL}) \tag{4.5}$$

t_{LH} denotes the time for falling input to rise the output by driving the load. More specifically, it is the time difference at which the signal crosses half-point of excursion, known as transition instance, between output and input. Similarly, t_{HL} is the time difference between transition instances of output and input of a delay cell. A LiDAR with RO-based TDC utilizes propagation delay of each cell to decode the phase. As a result, oscillation frequency can be reduced for steps with fine resolution. The minimum propagation delay in 28nm is in the order of pico-seconds, thus 50 ps LSB requirement reduces power constraint at the cost of noise.

4.3.1. Clock distribution and Auxiliary blocks

Ring oscillator operates for maximally one reference clock period provided by an external phase-locked loop (PLL). A PVT-tracking PLL generates a reference clock with four phases at $f = 320MHz$, which is distributed over the array of TDCs to achieve synchronization, thereby correcting phase noises. Moreover, the distributed reference clock also gates TDCs within the gating window. The two step approach shortens the TDC range to one clock period, $T = 3.125ns$, thus require 6 bits with LSB = 50 ps.

Dynamic power for clock distribution is given by

$$P_{dyn} = C_L V^2 f_{CLK} \tag{4.6}$$

where C_L, f_{CLK}, V, and C_L denote a load capacitance, a clock frequency, and a supply voltage, respectively. The load capacitance is a function of interconnect length. Assuming H-tree clock distribution, the total length of the tree is calculated as

$$L_{H-tree} = \sum_{i=1}^{3} \frac{w}{2^i} + \frac{h}{2^i} \tag{4.7}$$

where w, h, and i denote pixel array width, pixel array height, and distribution points of each level. The increase in length with each level decreases, thus $i = 3$ is taken with insignificant truncation error. The total length for the array of 320×100 pixels with pixel area $50 \times 50 \mu m^2$ yields the total length $L_{H-tree} = 9187 \mu m$.

Table.4.3 summarizes RC parasitics of the wire with calculated length which yields lower capacitance for top layer (metal layer 6) in comparison to lower layer (layer 1)

	Metal layer 1 (M1-FOX)				Metal layer 6 (M6-FOX)			
linewidth [μm]	0.05 (min)	0.5	0.1	2.25 (max)	0.05 (min)	0.2	0.5	0.75 (max)
C_c [fF]	12.2	22.9	34.1	60.2	160	186	231	265
C_T [fF]	637	1025	1437	2423	544	624	770	885
R [Ω]	34700	5100	3010	1320	29400	13100	5680	3860

Table 4.3: RC parasitics of the interconnect wires at metal layer 1 & 6

where C_c denotes coupling capacitances between the wire of interest and a neighbouring wire and C_T denotes a total capacitance of the wire of interest. Assuming supply voltage as the drive voltage $V = 0.9V$ and based on the largest obtained C_T, dynamic power is estimated to dissipate power in the order of mW. The estimation is only valid for the H-tree of the calculated wire lengths without auxiliary circuits such as repeaters. Assuming a lumped RC structure, the 1^{st} order dominant time constants of the wire in both layers are in the order of nanoseconds, $\tau = \alpha RC \approx ns$, thus the distribution network may require repeaters which will contribute to dissipated power for distribution significantly.

A comparable LiDAR system employing a coarse-fine structure is designed with 252×144 SPAD-based pixels and a PLL implemented in 180nm provides a coarse clock reference at the frequency of 960 MHz and consumes 176 mW [13]. The power consumption in the comparable system, together with the power estimation of the power distribution, serves to estimate the PLL power contribution to the individual pixel. That is, PLL provides a coarse timing reference and together with per-pixel TDC, produces 14-bit range for each pixel. Thus, the PLL power consumption is divided by the number of pixels to yield the estimation presented in section 4.2.1. More specifically, 10 μW is assumed as the PLL power contribution to the 14-bit time reference generation of an individual pixel.

On-chip ToF retrieval constructs histogram on-chip thus requires memory. The occupied area by memory for each pixel are estimated using the standard TSMC 28nm digital library. As discussed earlier, the considered two step approach with PLL period of period, $T = 3.125ns$ which is equivalent to 6 bits of LSB. That is, 64 histogram bins, where each bin window size is LSB, covers one PLL period. For each bin, 8-bit D-type flip flop store a single time-stamp for each measurement cycle. The standard cell HD28-8SDFPRHQX5 occupies 21.924 μm^2 for each bin, yielding total memory area of 1403 μm^2, equivalent to 56 % of the pixel area, thus necessitate the miniaturization of per-pixel TDC.

The output of the last delay cell in a ring oscillator is connected to a multi-bit ripple counter which counts oscillations with positive transitions and resets for each frame. The output of each remaining delay units drives a single-bit counter which can be triggered for both positive and negative edge to store transition. The readout signal freezes the ring oscillator, then the oscillation states are readout which can be converted into binary by using a lookup table or counting zeros of frozen states. Verilog-A implementaion of counters and thermometer-to-binary conversion can be found in the appendix A.2.

4.3.2. Delay cell topologies and propagation tuning

The topology of individual delay cell is chosen based on the performance parameters of a ring oscillator such as low power and area consumption, low jitter, 50% duty cycle of output waveform, a wide output swing range, common-mode signal rejection capability, and power supply noise rejection capability. Often in a voltage controlled oscillators (VCO), wide tuning range is also considered to compensate for the variations

in process, supply voltage, and temperature (PVT). The delay cell architectures under consideration are shown in Figure 4.3.

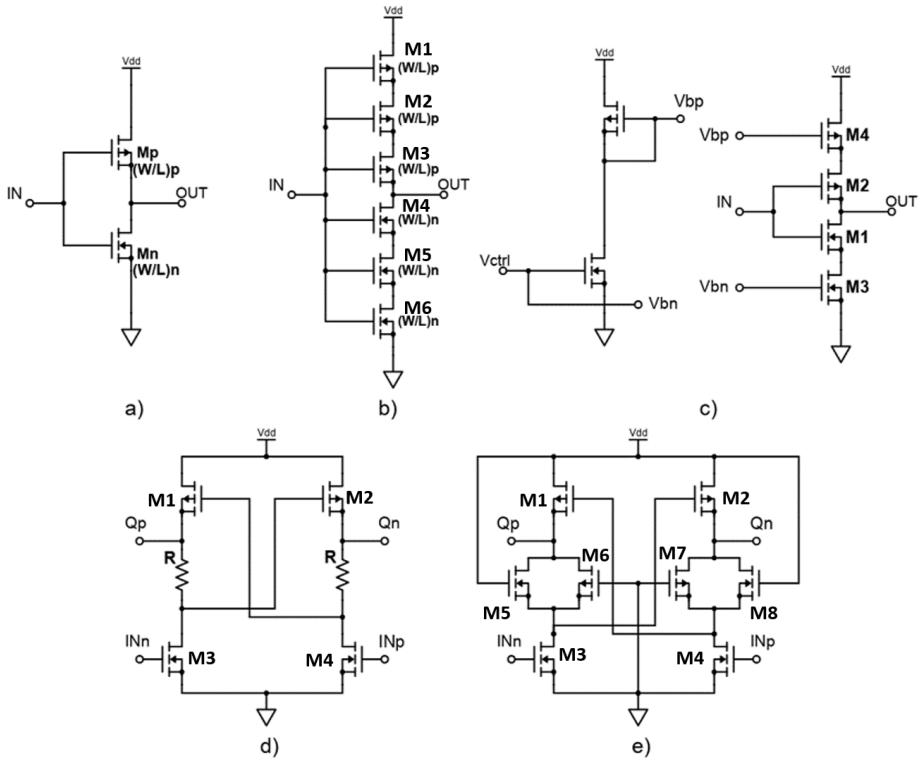


Figure 4.3: Delay cell topologies under consideration. a) CMOS inverter b) Stacked-inverter c) Current mirror and current starved inverter d) Differential Cascode Voltage-Switch-Logic with resistive enhancement (DCVSLR) e) DCVSLR with transmission gate implementation

Figure 4.3a-c are single ended and Figure 4.3d-e are differential inverters. In general, a single ended topology exhibits low power and offers full range output swing, but suffers from poor rejection capabilities of common-mode signal and power supply noise. In contrast, fully differential topology increases common-mode rejection ratio ($CMRR$) with current source, but requires headroom and current source. Moreover, the output swing is limited which results in higher phase noise, thus is not favoured by designs under low supply voltages. In contrast, pseudo differential counterpart offers compromise allowing full swing operation, but with degraded CMMR. Furthermore, differential output allows even number of delay cells in a ring by reversing the polarity at one end.

Pseudo differential topologies commonly implement cross-coupled load, which forms a positive feedback thereby enables oscillation with sharp transitions [65]. Furthermore, cross-coupled load act as an active differential negative transconductance, improving the common-mode rejection ratio [66, 67]. [68] presents the simplest low transistor counts of a differential cascode voltage-switch-logic ($DCVSL$), with differential input pair and a cross-coupled load. However, the coupling of the load requires one branch to pull-up before the other branch can pull-down and vice versa. As the result, there exists an inherent delay asymmetry. [69] achieves symmetry by implementing a resistor between the output node and the gate of PMOS load to increase the gate overdrive. The resistive enhancement technique reduces power consumptions without affecting the phase noise performance for the counter part $DCVSL$, is thus considered in this thesis.

Modeling the exact dynamic behaviour of a delay cell is challenging, as the rail-to-rail operation put transistors through every operation region. Moreover, unlike analogue amplifiers with fixed operating point, the input of an inverter also varies largely, rendering the small-signal analysis inapplicable. Nevertheless, modelling is possible with assumptions and provides useful insights.

Let us consider a CMOS inverter shown in Figure 4.3 with step input at t_H. Assuming PMOS does not conduct any current from $t > t_H$, hence open circuit, an average current NMOS sinks can be calculated for saturation and linear regions, with respect to output voltage. Similarly, input with step down function results

in PMOS driving the load. Using charge balance equation, the rise-fall time and fall-rise time are given by [70]:

$$\begin{aligned} t_{HL} &= \alpha_n \frac{C}{k'_n(W/L)_n V_{DD}} \quad where \quad \alpha_n = \frac{2}{\left[\frac{7}{4} - \frac{3V_{tn}}{V_{DD}} + \left(\frac{V_{tn}}{V_{DD}}\right)^2\right]}, \\ t_{LH} &= \alpha_p \frac{C}{k'_p(W/L)_p V_{DD}} \quad where \quad \alpha_p = \frac{2}{\left[\frac{7}{4} - \frac{3|V_{tp}|}{V_{DD}} + \left(\frac{|V_{tp}|}{V_{DD}}\right)^2\right]}. \end{aligned} \tag{4.8}$$

where $k'_n = \mu_n C_{ox}$ denotes process transconductance parameter for NMOS. The subscript p denotes PMOS. According to Eq. 4.8, the following observations are made:

- $t_{HL} = t_{LH}$ is achieved with W/L ratios of transistors assuming $V_{tn} = -V_{tp}$. Assuming the equal channel lengths, the ratio of width determines the matching of two transistors.
- Delay is proportional to C.
- Delay is inversely proportional to W/L ratio. However, the increase in W/L ratio also increases C, thereby setting the minimally achievable delay. The linear relationship between delay time and W/L ratio holds, only when the load capacitance C is dominated by capacitances that are not related to transistor dimension.
- Delay varies with V_{DD} and V_{th}. The low supply voltage reduces power quadratically, thus can be favoured for power efficient design. However, low V_{DD} increases sensitivity of matching constants (α_n&α_p) to threshold voltage variation.

In essence, the propagation delay is tuned through supply voltage, input and output transconductance, and output capacitance. The output capacitance can be increased by introducing an additional capacitance to the output of the delay cell but require more charges for transitions, increasing the area and power. Thus, the method is not pursued. Instead, transistor dimensions are increased and scaled, which in turn alter transconductance, and supply voltage is adjusted to yield 50 psec delay. Each delay conveys phase information, thus the condition $t_{HL} = t_{LH}$ must be met.

The stack inverter topology elongates channel lengths by cascoding the minimum channel length transistors as shown in Figure. 4.3. The gates of cascoded transistors are connected to the input, thus essentially reduces to the voltage-dependent resistors which controls the output transconductance. The oscillation frequency is tuned through supply, while maintaining the constant ratio W_p/W_n, to yield $t_{LH} = t_{HL}$.

The current starved inverters consist of current source, current sink, and the CMOS inverter in between to drive the load capacitance. The voltage controlled current mirror sets the gate voltages of the current sink and source, and sets the current through the branch. Operating the controlled transistor in saturation mode, the control voltage is a function of the current,

$$V_{ctrl} = V_{th,n} + \sqrt{\frac{2I_{D,n,sat}}{k'_n(W/L)_n}} \tag{4.9}$$

where $I_{D,n,sat}$ denotes the saturation current of NMOS. The topology has two degrees of freedom for tuning; supply variation and current variation. As the target power efficiency is stringent, the supply voltage is lowered which is followed by tuning the control voltage. Assuming step input function, t_{HL} is given by the load capacitance discharge time with current sink operating in saturation and linear region [71]

$$t_{HL} = \frac{2C_L(1-\eta)V_{DD}}{k'_n(W/L)_n(V_{gs,sink} - V_{th})^2} + \frac{C_L}{k'_n(W/L)_n(V_{DD} - V_{th})} \ln\left(\frac{\eta - \kappa}{\kappa}\right) \tag{4.10}$$

where $\eta := (V_{DD} - V_{th,n})/V_{DD}$ and $\kappa := V_{DD}/2$ denote the the lowest output voltage, to which the current sink operate in saturation and linear region, respectively. For $\ln\left(\frac{\eta-\kappa}{\kappa}\right) < 1$, $\Delta t_{HL} \propto \Delta V_{DD}$ holds.

Similarly, the dynamic behaviours of DCVSL can be modelled with assumptions [72] including the operation region, to state the significant assumption. Addition of correction factors, obtained empirically, improves the accuracy [69]. However, as the conducting branch causes voltage drop across resistor, as a function of time and operation region, the overdrive voltage increases

$$V_{G,Mp} = V_{Qn}(t) - I_{D,Mn}(t) \times R \tag{4.11}$$

which in turn shifts the operation region of load on the other branch, hence invalidate the assumptions. The propagation delays are tuned through ratio of transistor width, while using the constant resistance R.

4.3.3. Error sources and jitter

Each delay cell contains noise sources which fluctuates the output voltage. The deviation in voltage domain is translated to time domain, arising an uncertainty in transition instances referred as jitter. In this thesis, jitter after n propagation is defined as the standard deviation of the distribution of an output transition after n intervals or propagations, as presented in [73]. The definition based on the variable transition spacing known as clock jitter, as presented in [74, 75], is not used in this thesis.

The error sources include device noises such as thermal noise and shot noise, external noises such as supply and substrate noise, and local variations known as fabrication mismatch. [76] correlates the device noise to jitter with a figure of merits to optimize the design for device noise dominated ring oscillators. The total device-noise induced jitter through n delay cells is given as:

$$\sigma_{dev,k=n} = n\sqrt{\sigma_{dev}}. \tag{4.12}$$

where σ_{dev} denotes device-noise induced jitter per delay cell. The frequency modulation due to supply and substrate noises for single ended and fully differential ring oscillators are studied in [75, 77, 78]. In contrast to white noise, supply and substrate noises exhibits non-uniform spectral density, thus results in correlation among induced jitter, hence jitter adds linearly with propagations [74]. In other words, the clock jitter is a function of frequency and the transition, hence the dynamic measurements such as cycle-to-cycle jitter is required. In this thesis, supply sensitivity is defined as the difference in oscillation frequencies with small DC variation in supply voltage

$$S_{\Delta V_{DD}} = \left| \frac{f_{osc^+} - f_{osc^-}}{\Delta V_{DD}^+ - \Delta V_{DD}^-} \right| \tag{4.13}$$

Lastly, the local process variation causes the threshold voltage mismatch among transistors. As shown in Eq (4.8), the variation in threshold voltage causes mismatch in pull-up and pull-down network, resulting in asymmetrical transitions. Consequently, the transition instances vary. The variance is inversely proportional to the gate area, known as Pelgrom's law [64]

$$\sigma_{VT}^2 = \frac{A_{VT}^2}{WL} \tag{4.14}$$

where A_{VT} is a process specific constant. The scaling with multiplication factor M reduces the variance linearly, which in turns improves matching and jitter. As the area efficiency is also considered, the trade-off between area and jitter is important. Thus, the following subjects are studied for various delay cells: jitter accumulation pattern, minimum M to meet INL, supply sensitivity, static and dynamic power, and area. The evaluation parameters correlated to the system level decisions and decisions are justified.

4.4. Simulation results of ring oscillators

4.4.1. Simulation set-up and measurements

A ring oscillator consisting of N delay cells is supplied with ideal voltage source V_{DD}. The output node of the first delay cell is initially set to ground. The macro model for transistors were used to estimate layout parasitics. Transient analysis simulates the dynamic behaviour of the ring oscillator, where signal propagates. After k propagations, the transition instance $\tau = k \times LSB$ is recorded.

Supply sensitivity is calculated using transient analysis. As defined in Eq.(4.13), supply sensitivity is difference in frequencies when supply voltage is superimposed by small DC voltage ΔV_{DD}. The altered oscillation period (T'_{osc}) can be expressed as:

$$T'_{osc} = \tau'_{x+2N} - \tau'_x \tag{4.15}$$

where τ'_x is x^{th} transition and T'_{osc} is translated to oscillation frequencies. $\Delta V_{DD} = \pm 25mV$ is used.

Leakage power is obtained with DC simulation. The input and output of each delay cell are set to opposing rail. The static power of ring oscillator is the product of supply voltage, current through all branches, and the number of delay cells

$$P_{leak} = V_{DD} I_{branch} N_{branches} N_{cells} \tag{4.16}$$

and is measured for wide temperature range $[0 - 100°C]$ for outdoor application.

Monte Carlo analysis yield 300 transition instances, from which histogram is built. Assuming normal distribution, the standard deviation, known as mismatch induced jitter, is calculated. Assuming jitter only increases with propagations, the last transition within the TDC range exhibits the highest jitter, thereby achieves

the worst INL. The transistors are scaled with multiplication factor M, to reduce the worst case jitter, thereby meet the INL requirement. Accordingly, minimum power consumption and gate area are calculated. Then, the channel length is increased and delay is tuned to yield 50 ps by adjusting supply voltage or channel width. The simulation is repeated for various N and for all delay cells.

4.4.2. Single ended Inverter based cells

For CMOS inverter delay cell, the channel length is increased while keeping the constant (W/L) ratio for both PMOS and NMOS. Neglecting short-channel effect ($\lambda = 0$), the transconductance remains constant while increasing the load capacitance due to increased gate area of succeeding delay cell (and related parasitics). Thus, higher voltage is supplied. For L_{ch} = 60nm, V_{DD} = 545mV yields 50 ps gate delay. The matching between PMOS and NMOS is achieved through $W_p/W_n = 3$ which results in symmetry in pull-up and pull-down time $t_{LH} = t_{HL}$. Figure 4.4 shows the variance of transition instances over propagations (k)

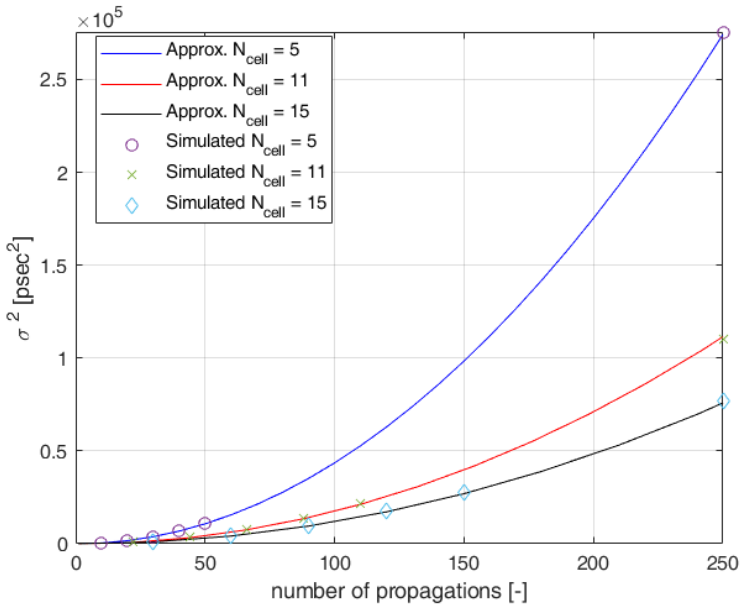


Figure 4.4: Simulated mismatch induced variance in transition instances over propagations. Inverter based ring oscillator with N delay cells, V_{dd} = 545mV & L_{ch}=60nm

The simulated variance can be expressed as:

$$\sigma_k^2 = \alpha_N k^2 \qquad where \quad k = 2N, \quad N \in \mathbb{Z} \tag{4.17}$$

where α_N is a constant specific to N. Equivalently, local variation induced jitter increases linearly for every ring oscillator period. In contrast, within the oscillation period, the variance increases linearly with propagations. Resultantly, the range where variance increase linearly over propagations increase with more delay cells within the ring oscillator, reducing α_N. Eq. (4.14) explains the inverse proportional relation between variance in threshold and gate area. The gate area is scaled with multiplication factor M to reduce jitter. Simulation results show that jitter reduces inversely proportional to square root of multiplication factor

$$\sigma_k \propto \frac{1}{\sqrt{M}} \tag{4.18}$$

Assuming mismatch is the dominant error source, the minimum multiplication factor (M_{min}) to reduce the worst case jitter, within the TDC range, below LSB size is given by

$$M_{min} > \frac{\sigma_{k,max}^2}{LSB^2} \tag{4.19}$$

where $\sigma_{k_{max}}$ denotes the jitter at the last propagation within the TDC range. M_{min} decreases with higher delay cell numbers within the loop due to reduced periodic constant α_N.

In a coarse-fine structure, PLL period represents the dynamic range of the individual TDC. Assuming constant fine resolution and linearity requirement, σ_k^2 must be further reduced in a TDC which covers a longer range. Figure 4.5 shows the relation between the minimum scaling factors and the number of single-ended inverter based delay cells for two different phase correction frequencies f_{PLL} = 80 MHz and f_{PLL} = 320 MHz. That is, for a system with f_{PLL} = 80 MHz, linearity requirement must be satisfied after 250 propagations

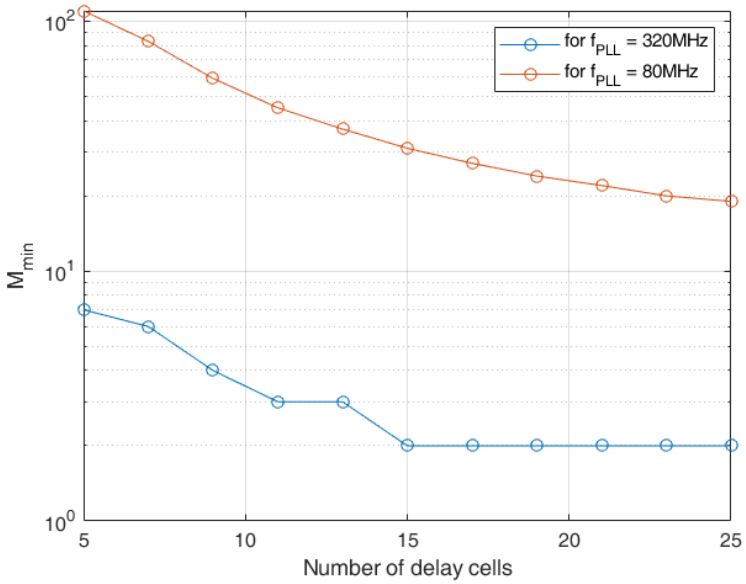


Figure 4.5: Minimum scaling factors over number of delay cells within ring oscillator

whereas for a system with f_{PLL} = 320 MHz, the same requirement must be satisfied after 63 propagations. k_{max} denotes such number of propagations.

In a system with f_{PLL} = 80 MHz, M_{min} decrease linearly with the number of delay cells in the loop, resulting in reduced variance for an RO with 25 delay cells by a factor of 5 in comparison to 5 delay cells based RO. As the PLL frequency increases, the operation range of RO reduces, thus accumulating less jitter. In a system with f_{PLL} = 320 MHz, M_{min} exhibits step-size transitions for close delay cell numbers, and eventually saturates as jitter is reduced below LSB size.

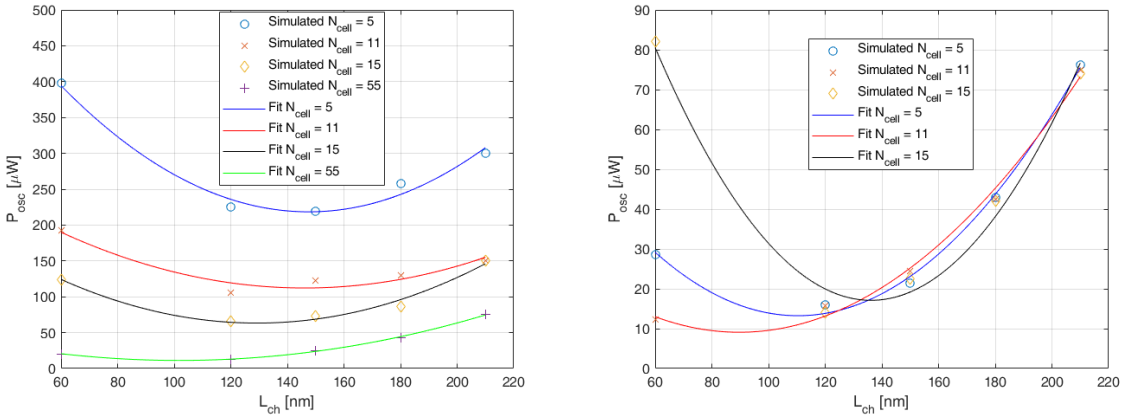
The above result is of interest as M_{min} directly relates to power consumption, thereby aids system decision for f_{PLL} selection. The oscillation power is expressed as:

$$P_{osc} = M_{min} \times \overline{P_{osc,L_{ch}}} = M_{min} \times \frac{1}{T_{osc}} \int_{\tau_x}^{\tau_{x+2N}} \left(V_{dd,L_{ch}} \times I_{L_{ch}}(t) \right) dt \tag{4.20}$$

where $\overline{P_{osc,L_{ch}}}$ denotes normalised oscillation power specific to channel length and τ_a denotes the time instance of a^{th} propagation. Figure 4.6 shows the simulation results of average oscillation power of N single ended inverter based cells over channel lengths. The simulated data points are plotted with circles and continuous lines represent the approximation. Note that the average oscillation power represents the minimum value optimized to satisfy the INL condition, thus account only the ring oscillator. In both systems shown in Figure 4.6a and Figure Figure 4.6b, the ring oscillator comprising cells designed with L_{ch}=120nm consumes the lowest power. The oscillation power in Figure 4.6a changes linearly with N following the relationship in eq.(4.17)-(4.19) and $M_{min} > 1$ for all N. In contrast, in a system with f_{PLL} = 320 MHz, the shortened TDC dynamic range yields lower accumulated jitter, thereby leading to M_{min}=1 for $L_{ch} > 120nm$. As a result, the oscillation power is independent of the number of stages for elongated channels, as shown in Figure 4.6b, and satisfy power requirement for all N. The result indicates the direct trade-off between PLL frequency and power consumption of individual ring oscillator of the respective TDC and justifies the selection of f_{PLL} = 320 MHz.

The minimum gate area can be expressed as:

$$A_{gate,min} = L_{ch} \ (W_p + W_n) \ \times \ N \ \times \ M_{min} \tag{4.21}$$



(a) A system with $f_{PLL} = 80MHz$ (b) A system with $f_{PLL} = 320MHz$

Figure 4.6: Minimal average oscillation power of an individual ring oscillator over transistor channel lengths

where M_{min}, L_{ch}, W_p, W_n, and N denote minimum multiplication factor, channel length, channel width of pmos, channel width of nmos, and number of delay cells, respectively. Figure 4.7 shows the $A_{gate,min}$ over the stages and for varying channel lengths for a short range system. The minimum transistor area for

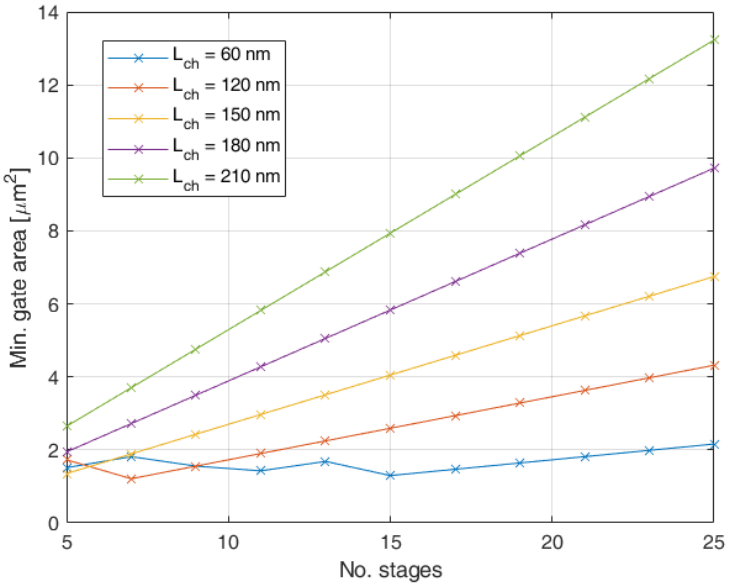


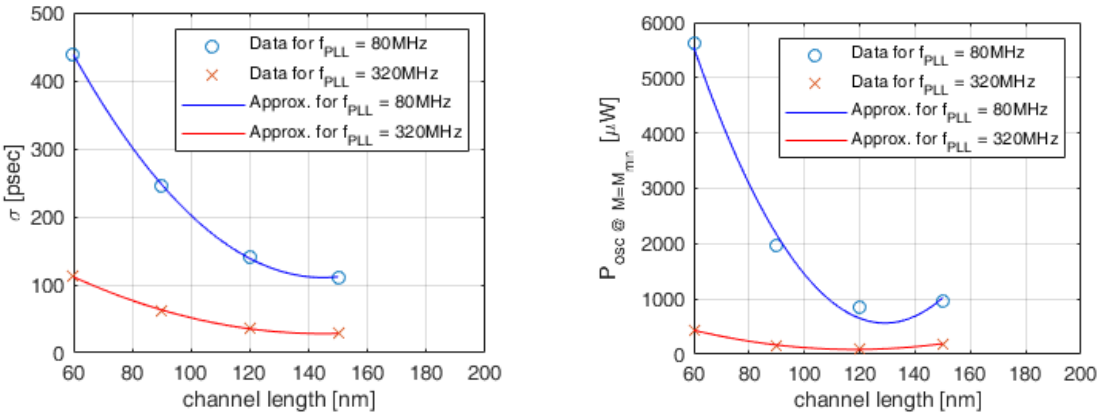
Figure 4.7: Minimal transistor dimensions over the number of delay cells

L_{ch}=60nm fluctuates for N <15 due to step-wise integer reduction in M_{min} with the increase in N, until until reaching to M_{min}=1 (N =15) from which point increases linearly over the number of stages. Similarly, M_{min} increases linearly for longer channel lengths due to saturated M_{min}. Taking into account the layout overhead of 30%, the designed oscillator occupies below $150\mu m^2$, even for $N = 25$. However, the auxiliary block area may dominate, thus low delay cell count is preferred, unless $A_{gate,min}$ reduces significantly with more delay cells. Therefore, N=5 is considered for the stacked and current starved delay cells.

4.4.3. Pseudo differential cells

In contrast to single-ended CMOS inverter based delay cells, the delays of DCVSLR cells shown in Figure 4.3c are tuned with channel length and W/L ratio of PMOS. The increase in channel lengths of input and load transistors reduces short-channel effect and causes deviation in transconductance. Manipulation of

load's channel width ensures symmetrical output transitions with 50 psec delay, while supply voltage and enhancing resistors remain constant. Supply voltage is set to 0.9V and the the resistance of the enhancing resistors are kept constant at 4.9 kΩ. Figure. 4.8 depicts the simulation results using 5 delay cells. Figure 4.8



(a) Jitter at $k = k_{max}$

(b) Average oscillation power at $k = k_{max}$

Figure 4.8: DCVSLR, with ideal resistors, cell based RO simulation results for PLL frequencies: f_{PLL} = 80MHz ($M_{min} = 9$) & f_{PLL} = 320MHz ($M_{min} = 1$).

shows the simulation results of a DCVSLR cell based RO for two different PLL frequencies over the channel lengths. Figure 4.8a shows the jitter performance where for each jitter, M_{min} is simulated to satisfy INL requirement following Eq.(4.19) and the corresponding average oscillation powers are shown in Figure. 4.8b. The data points are approximated with quadratic equations where the minimal power consumption is estimated around channel length of 120nm. A system with $f_{PLL} = 80MHz$ records the minimum $P_{osc} = 848\mu W$, which exceeds the allowed maximum power consumption by more than a factor of 8, whereas in another system with $f_{PLL} = 320MHz$, the average oscillation power is below the specification at L_{ch}=120nm. Therefore again, the system with $f_{PLL} = 320MHz$ is considered for the performance comparison. The remaining simulation parameters are summarized in Table. 4.4.

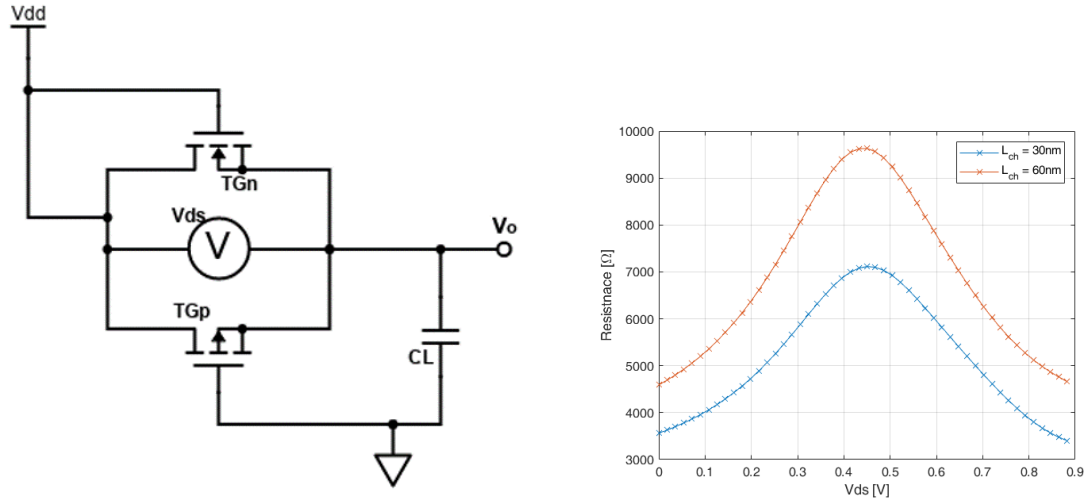
In layout, ideal resistors are implemented with polysilicon, and often occupy significant area with respect to transistor areas. Furthermore, the metal line act as a capacitance when interacting with other lines, leading to undesired parasitic capacitances. Transmission gate, shown in Figure. 4.9a, effectively reduces area while maintaining a relatively constant resistance over the voltage difference applied across it. Either drain or source of NMOS connected to V_{DD}, the equivalent resistance follows Ohm's law,

$$\begin{aligned} R_{N_{eq}} &= \frac{V_{DD} - V_o}{\frac{1}{2}k_n(V_{DD} - V_{tn} - V_o)^2} \qquad for \;\; V_o \le V_{DD} - V_{tn} \\ &= \infty \qquad\qquad\qquad\qquad\qquad for \;\; V_o \ge V_{DD} - V_{tn} \end{aligned} \tag{4.22}$$

where operating regions of the MOSFET determines the current flow, hence two regions are shown [70]. Similarly, the equivalent resistance of PMOS is a function of voltage across drain-source and the current through channel,

$$\begin{aligned} R_{P_{eq}} &= \frac{V_{DD} - V_o}{\frac{1}{2}k_p(V_{DD} - |V_{tp}|)^2} \qquad\qquad for \;\; V_o \le |V_{tp}| \\ &= \frac{1}{k_p\left[V_{DD} - |V_{tp}| - \frac{1}{2}(V_{DD} - V_o)\right]} \qquad for \;\; V_o \ge |V_{tp}| \end{aligned} \tag{4.23}$$

where parallel connection of two MOSFETs results in equivalent resistance $R_{TG,eq} = R_{N_{eq}} || R_{P_{eq}}$ and is shown in Figure. 4.9b. During transitions, the voltage across the transmission gates vary dynamically, minimally $V_{ds} = 0V$ and maximally $V_{ds} = V_{DD}$. Thus, the opposing propagation time at both output nodes equate $t_{LH} = t_{HL}$ when $R_{TG,eq}$ symmetrical around $V_{DD}/2$. The transistors are sized with the same channel length and the widths are varied. The channel lengths were adjusted to yield 50 psec of propagation delay. In comparison to the ideal resistors, transmission gate increases the jitter due to mismatch. Moreover, transmission gate records slightly smaller power consumption at $90.36\mu W$, compared to the ideal resistor implementation



(a) Implemented transmission gate and R_{eq} simulation set-up

(b) R_{eq} over V_{ds} with $W_{n,TG}/W_{p,TG} = 300nm/110nm$.

Figure 4.9: Implemented transmission gates in DCVSLR cell

with equivalent M factor. The voltage-dependent $R_{TG,eq}$ self-regulates the current flow, thereby controls the load's gate overdrive voltage relatively constant upon disturbances when compared to voltage-independent resistors. Therefore, the supply sensitivity is reduced.

Table 4.4 summarizes performance of a ring oscillator with 5 delay cells using different delay cell topologies. Each delay cell is optimized for mismatch-induced jitter and oscillation power.

Topology	CMOS Inverter	Stacked-inverter	Current-starved inverter	DCVSLR with R_{ideal}	DCVSLR with TG
V_{dd} [mV]	615	896	740	900	900
L_{ch} [nm]	120	30 × 4	120	120	120
W_p/W_n [-]	1080/360	(270/100) × 4	1340/360	500/360	500/360
$S_{\Delta V_{dd}}$ [GHz/V]	12.42	4.70	1.8	7.37	4.44
$\sigma_{n=63}@M=1$ [psec]	53.82	43.56	88.79	35.37	41.87
M_{min} [-]	2	1	4	1	1
$P_{osc}@M=M_{min}$ [μW]	26.52	12.44	152.4	94.32	90.36
$A_{min.gate}$ [μm^2]	1.73	0.8640	4.08	0.52 + A_R	0.56

Table 4.4: Performance summary of ring oscillators with $N = 5$ delay cells

An oscillator with CMOS inverter delay cells consumes $26\mu W$, but is susceptible to supply variation, $S_{\Delta V_{dd}} = 12.42GHZ/V$. High supply sensitivity deteriorate synchronicity across TDC array due to IR-drop, when oscillators start upon asynchronous events [60].

Comparatively, stacked-inverter based cells reduces the supply sensitivity by almost a factor of 3, when supplied at higher voltage. Moreover, mismatch-induced jitter is reduced to $\sigma_{n=63,@M=1} = 43.56ps$, thereby allow $M_{min} = 1$. As a result, the topology consumes the lowest power and occupies small area.

The supply sensitivity is further reduced with conventional current-starved topology, but the voltage-controlled current sources contribute to jitter significantly. Accordingly, the gate area is increased to reduce

the jitter following the Eq.(4.14), at which the power consumption is $152.4\mu W$, which exceeds the design goal.

Alternatively, pseudo-differential cell, DCVSLR is considered. The topology exhibits the smallest jitter when ideal resistor is implemented power consumption within the target $P_{osc} = 94.32\mu W$. The transmission gates replace the additional resistances, which in turn increases jitter. However, the voltage-dependent resistance acting as a current regulator through the branch, improves supply sensitivity with small chip area $A_{min,gate} = 0.56\mu m^2$.

Despite low supply sensitivity achieved by current-started inverter topology, high power consumption renders the topology impractical. Stacked inverter and DCVSLR with TG topologies records comparable $S_{\Delta V_{dd}}$ and $\sigma_{n=63,@M=1}$. However, comparatively, stacked inverter topology consumes less than 14% at the cost of larger minimum gate area. However, a ring oscillator implementing any one of the above topologies occupies less than 0.2 % of the TDC area, thus area is neglected in decision making. Accordingly, $\overline{FoM_{TDC}}$ of the stacked-inverter is calculated and is indicated in Figure. 4.1. Note that the calculated $\overline{FoM_{TDC}}$ only takes into account the average power of a ring oscillator, thus represents the minimum $\overline{FoM_{TDC}}$. Furthermore, with reference to a similar system, PLL power contribution to the individual TDC is assumed and accordingly a 14-bit long range $\overline{FoM_{TDC}}$ is estimated. Lastly, given the large pixel array, 320 × 120, further reduction in power down to below $1\mu W$ should be researched.

4.5. Conclusion

Proposed Figure-of-Merits (FoM) normalizes LSB, thereby takes into account conversion step size. Figure 4.1 shows that normalized FoM increases logarithmically with higher technology node. The mature (=lower) technology benefits from the shortened minimally achievable propagation delay, thus lower normalized FoM is recorded. When plotted with respect to area, ring oscillator based TDCs fall closest to the target region. The ring oscillator architecture has a small form factor owing to the recycling of hardware and minimal auxiliary blocks, thus is considered in this thesis in combination with phase correcting PLL.

The clock distribution is assumed to consume $10\mu W$ per TDC at $f_{PLL} = 320MHz$. On pixel level, memory cells store generated timestamps thereby allowing histograms. Using 8-bit D-type flip flop as a memory bin from the standard TSMC 28nm digital library, the total memory area is 56% of the pixel area. The estimation is valid for PLL running at $320MHz$.

The propagation delay of each cell is tuned to yield LSB with symmetrical output. The mismatch-induced jitter accumulates linearly with every oscillation period, thus dominates jitter. Furthermore, jitter is inversely proportional to multiplication factor (M). The minimum multiplication factor (M_{min}) to satisfy linearity requirement are calculated following Eq.(4.19). The average power (P_{osc}) at the calculated M_{min} suggests that PLL operating at $f_{PLL} = 80MHz$ exceeds the required power consumption, thus 4 phase design with $f_{PLL} = 320MHz$ is further considered. Furthermore, the optimal jitter is recorded for $L_{ch} = 120nm$ for all delay cells, when the transconductance are varied with channel length.

Lastly, a ring oscillator with 5 delay cells are compared for different delay cell topologies. The stacked-inverter, biased by current sources, topology consumes the lowest power of $12.44\mu W$ and susceptibility to supply variation. However, the additional transistors contribute to mismatch-induced jitter, which eventually increases power consumption. DCVSLR with transmission gate exhibits lower supply sensitivity $4.44GHz/V$ and $41.87ps$ of jitter after 63 propagations, owing to the self-regulating transmission gate. Similarly, stacked-inverter topology exhibits comparable supply sensitivity and jitter with respect to DCVSLR with transmission gate, but consumes less than 14% of power, at the cost of larger area. However, a ring oscillator implementing any one of the introduced topologies occupies less than 0.2 % of the TDC area, thus area is neglected in decision making. Therefore, a stacked-inverter topology shall be considered as a delay cell topology when designing a ring oscillator with low $\overline{FoM_{TDC}}$.

5

Conclusion

In this work, a SPAD-based LiDAR system is studied. In particular, the system employs direct time-of-flight (dToF) method to reconstruct a target-reflected pulses using histogram, from which the distance to the target is estimated. More particularly, time correlated single photon counting (TCSPC) method records the flight time of pulses. Such method assumes and successfully reconstructs the returned pulses in a photon starved condition, but fails under strong background noise. The increase in average photon counts per detection range limits the detection range, thereby rendering the system impractical for an automotive application where strong sunlight is present.

In attempt to resolve the shortcomings, the characteristics of SPADs and their building blocks are studied. Then, bottlenecks are identified and evaluated using various receiver operation schemes. Then, in-pixel timing generating blocks are considered. The performance of ring oscillators with various delay cells are compared to yield the candidate for an in-pixel time-to-digital (TDC).

The following summarizes the original contributions and the findings of this work more in detail.

5.1. Original contributions

- Literature review of state-of-the-art SPAD performances
- Literature review of receiver operations of a LiDAR system and identification of bottleneck.
- Performance comparison between synchronous interleaved SPAD gating and asynchronous SPAD operation, in a photon-abundant application.
- Proposal of adaptive TDC gating as a data compression method.
- Literature review of TDCs and proposal of normalized Figure-of-Merits for TDCs.
- Performance comparison of delay cell topologies in a ring oscillator implemented in 28nm.

5.2. Summary of findings

Chapter 2 provides an overview of single-photon avalanche diode (SPAD). In particular, operational principles, minimal working blocks, and the state-of-the-art SPAD performances are presented. Recent works achieve photon detection probability (PDP) up to 55% with a few pico-seconds of full-width half maximum (FWHM), but PDP starts decaying to achieve average of 5% around $\lambda = 900nm$ among different technology nodes. The operation at the near infra-red (NIR) wavelength is critical for outdoor application as the absorption rate of the natural light is the highest within the band. The statistical spread in timing response, known as timing jitter, of 7.8 ps is reported. In conjunction with the SPAD regulatory front-end electronics and further processing units such as time-to-digital converter (TDC), the detection time with high precision is known. In particular, active quenching circuit (AQC) achieves fast transitions thereby reducing the probability of premature avalanche during recovery, which in turn improves maximum counting rate. Moreover, AQC achieves well defined SPAD deadtime which is a considered receiver design parameter in this thesis. 3D stacking technology implements SPAD and the front-end electronics on separate tiers, thereby improving the fill factor

upto 74.4%. Furthermore, the segregated circuit layer increases the circuit area which aids large SPAD-array implementation, such as an automotive application.

Chapter 3 discusses the considered SPAD-based LiDAR system. In particular, basic principle and limitations of time correlated single photon counting (TCSPC) method is introduced, the receiver operation is mathematically formulated, and the proposed SPAD gating scheme as well as the TDC gating schemes are simulated using Matlab based on the state-of-the-art SPAD performances studied in Chapter 2. Our analysis confirms that histogram of asynchronous SPAD gating scheme can be estimated using Markovian chain and comparatively improves SNR by 4dB with respect to the interleaved gating scheme in region 1, where peak counts consist of both signal and noise. Consequently, asynchronous SPAD gating scheme extends reliable detection by 20m, with respect to the intereleaved gating scheme under the same condition, to achieve 74m of detection range. However, the improvement SNR and reliability comes at the cost of 311% increase in photon counts. Simulated photon throughput of the asynchronous scheme requires bandwidth of approximately 112 GB/s for off-chip processing or 235 MB of on-chip memory. The proposed adaptive TDC gating scheme utilizes coarse-fine structure, thus achieves reduction in the required memory size by a factor of at least 14 with respect to the system operating asynchronous SPAD gating without adaptive TDC gating scheme.

Chapter 4 discusses time-to-digital converter. In particular, state-of-the-art TDC are presented and ring-oscillator based TDCs are designed, simulated, and compared. The proposed Figure-of-Merits (FoM) normalized to LSB ($\overline{FoM_{TDC}}$) indicates a logarithmic relation between the proposed FOM and the technology node, and ring oscillator architecture achieves small form factor. The implemented 7-bit ring oscillator based TDC alone achieves $\overline{FoM_{TDC}} \approx 0.2[\frac{pJ}{conv.-step}]$ and a coarse-fine structure is expected to achieve $\overline{FoM_{TDC}} \approx 0.003[\frac{pJ}{conv.-step}]$ per pixel. The simulation results of a ring oscillator shows that the mismatch-induced jitter accumulates linearly with every oscillation period, thus dominates jitter. Furthermore, jitter is inversely proportional to multiplication factor (M). The simulated average power of a ring oscillator in a coarse-fine system operating $f_{PLL} = 80MHz$ exceeds the allowable power for the minimum multiplication factor (M_{min}) to satisfy linearity requirement. Thus, $f_{PLL} = 320MHz$ is further considered. Furthermore, the optimal jitter is recorded for $L_{ch} = 120nm$ for all delay cells, when the transconductance are varied with channel length. A ring oscillator with 5 delay cells are compared for different delay cell topologies. The stacked-inverter, biased by current sources, topology consumes the lowest power of $12.44\mu W$ and susceptibility to supply variation. However, the additional transistors contribute to mismatch-induced jitter, which eventually increases power consumption. DCVSLR with transmission gate exhibits lower supply sensitivity $4.44GHz/V$ and $41.87ps$ of jitter after 63 propagations, owing to the self-regulating transmission gate. Similarly, stacked-inverter topology exhibits comparable supply sensitivity and jitter with respect to DCVSLR with transmission gate, but consumes less than 14% of power, at the cost of larger area. However, a ring oscillator implementing any one of the introduced topologies occupies less than 0.2 % of the TDC area, thus area is neglected in decision making. Therefore, a stacked-inverter topology shall be considered as a delay cell topology when designing a ring oscillator with low $\overline{FoM_{TDC}}$.

5.3. Future work

The following topics may be further researched to enhance understanding in the topic and improve the system performances.

- Impact of pulse shape on the histogram of received pulses.
- Effective and power efficient peak detection algorithms.
- Adaptive TDC sharing among group of pixels.
- Dynamic supply sensitivity of the delay cells.
- Power reduction techniques and/or architectures for ring-oscillator based TDCs.
- Timing synchronization topologies in TDC arrays.

6

List of applied patents

- Asynchronous Gating Technique for Light Detection and Ranging (LIDAR) system based on Single-Photon Avalanche Diode (SPAD) (Application No: 16/939875)
- Adaptive TDC gating Technique for Light Detection and Ranging (LIDAR) system based on Single-Photon Avalanche Diode (SPAD)

A

Appendix

A.1. System level Matlab simulation parameters

Category	Parameter	Value	Unit
Image	Depth Image resolution	320 x 120	pixel
	Field of View	32 x 12 (H x V)	degree
	Frame rate	30	fps
	Detection range	150	m
	Depth resolution	1.5	cm
	Histogram resolution	200	psec
Illumination Source	Repetition rate	1	MHz
	Pulse duration (FWHM)	5	nsec
	Peak power	75	W
	Central wavelength	905	nm
Optical filter and lens	Central wavelength	905	nm
	Optical bandwidth	10	nm
	Optical efficiency	0.7	[-]
	f/N[1]	f/1.4	[-]
	Lens aperture[2]	15.94	Mm
	Focal length	22.32	mm
	Lens area	150	nm^2
Target	Target reflectivity	0.5	[-]
	Target reflection singularity	Single	[-]
Background environment	Background illuminance	100	Klux
	Sunlight irradiance	0.3	$Wnm^{-1}m^{-2}$
	Attenuation coefficient	0.56	$dBkm^{-1}$
Pixel	Pixel pitch	50	um
	Pixel fill factor	30	%
SPAD (BSI 3D stacking)	PDP @λ = 905nm	5	%
	Median DCR	5	Kcps
	Deadtime	12	nsec
	After pulse probability	Negligible	%
	Crosstalk	Negligible	%
	Jitter @ λ = 905 nm	300	psec

[1] The f-number : N = focal length/ effective aperture

[2] Aperture : D = pixel pitch * # of pixels / (2 tan (AFOV/2) * f/#) = (40e-6*320 /(2*tan(16 degrees)*1.4)) = 15.94 mm

Table A.1: Simulated optical, system, environmental and SPAD parameters

A.2. Verilog-A Code

A.2.1. Ideal ripple counter

```
‘include "constants.vams"
‘include "disciplines.vams"
‘define SIZE 5

module msde_lidar_tdc_counter( cnt, in, clk , rst);

        output[‘SIZE-1:0] cnt;
        electrical [‘SIZE-1:0] cnt;
        input in;
        input clk;
        electrical in;
        electrical clk;
        input rst; voltage rst; // Reset input (immediately forces Q output low) (active high)

        parameter integer setval = 0 from [0:(1<<‘SIZE)-1];
        parameter real vtol = 0;                                    // signal tolerance on the in
        parameter real ttol = 0;                                    // time tolerance on the in
        parameter real td = 0 from [0:inf);      // delay from clock to q
        parameter real tt = 1f/1000 from [0:inf);       // transition time of output signals
        parameter real vh = 1;                                      // output voltage in high state
        parameter real vl = 0;                                      // output voltage in low state
        parameter real vth = (vh + vl)/2;                           // threshold voltage at inputs
        parameter integer dir = +1 from [-1:+1];
                        // if dir=+1, rising clock edge triggers flip flop
                        // if dir=-1, falling clock edge triggers flip flop
        parameter integer stepsize = 1;
        integer cnt_int = 0; // intermediate value
        integer cntval ;

        analog begin
        @(initial_step("static","ac")) cntval=setval;
        @(cross(V(in)-vth,dir,vtol,ttol))begin
                if( V(rst) < vth )
                    cnt_int = (cnt_int + stepsize)%(1<<‘SIZE);
        end
        //$display ("cnt_int = %d",cnt_int);

        @(cross(V(clk) - vth, 1, vtol, ttol)) begin
                cntval = cnt_int;
//              $display ("cntval = %d",cntval);
        end

        @(cross(V(rst)-vth,1,vtol,ttol)) begin
                cntval = vl;
                cnt_int = vl;
    end

        generate j (‘SIZE-1,0)
                begin
                V(cnt[j]) <+ transition (!(!(cntval &(1<<j)))*vh+!(cntval&(1<<j))*vl,td,tt);
                end // output assignment

        end // analog beign
```

endmodule

A.2.2. Thermometer-to-binary converter

```
‘include "constants.vams"
‘include "disciplines.vams"
‘define STAGES 5 // number of delay stages in VCO

module msde_lidar_tdc_ThermoboutaryConv( bout, thermo , clk, rst);

// input and output delcariation
// attribute discipline (electrical)
        output [‘STAGES-2:0] bout;
        electrical [‘STAGES-2:0] bout;
        input [‘STAGES-1:0] thermo;
        electrical [‘STAGES-1:0] thermo;
        input clk;
        electrical clk;
        input rst;
        electrical rst;
        electrical clk_delay;

// Instantiate parameters
        parameter real vtol = 0;                                    // signal tolerance on the clk
        parameter real ttol = 0;                                    // time tolerance on the clk
        parameter real td = 0 from [0:inf);      // delay from clock to qQ
        parameter real tt = 1f/1000 from [0:inf);       // transition time of output signals
        parameter real vh = 1;                                      // output voltage in high state
        parameter real vl = 0;                                      // output voltage in low state
        parameter real vth = (vh + vl)/2;                           // threshold voltage at inputs
        parameter integer dir = +1 from [-1:+1] exclude 0;
                        // if dir=+1, rising clock edge triggers flip flop
                        // if dir=-1, falling clock edge triggers flip flop
        integer sum = 0;                    // sum of zeros
        integer val = 0;         // value in decimal

// analog
        analog begin

// delay clk signal - take into account the delay of the readout signal transition
        V(clk_delay) <+ absdelay( V(clk), 15p);

// update value @ read
        @(cross(V(clk_delay)-vth,dir,vtol,ttol)) begin
        generate i (‘STAGES-2,0) begin
                sum = sum + ((V(thermo[i])>vth) ? 0 : 1);                   // no of zeros
        end
        val = ‘STAGES + ((V(thermo[‘STAGES-1])>vth) ? +sum : -(sum+1)); // assign final value
        end

// Reset value @ rst
        @(cross(V(rst)-vth,dir,vtol,ttol))
        generate i (‘STAGES-1,0) begin
                sum = vl;
                val = vl; // assign final value
```

```
        end

// Transition for the final
        generate j (`STAGES-2,0) begin
                V(bout[j]) <+ transition (!(!(val &(1<<j)))*vh+!(val&(1<<j))*vl,td,tt);
                end // output assignment
        end // analog begin

endmodule
```

Bibliography

[1] Edward M.D. Fisher. Single-photon avalanche diodes in cmos technologies for optical communications. (10), 2017.

[2] M.W. Fishburn. *Fundamentals of CMOS Single-Photon Avalanhce Diodes*. PhD thesis, Delft Univerisity of Technology, Delft, 2012.

[3] S. Cova *et al.* Avalanche photodiodes and quenching circuits for single-photon detection. *Appl. Opt.*, 35(12):1956–1976, Apr 1996.

[4] A. Lacaita *et al.* Photon-assisted avalanche spreading in reach-through photodiodes. *Applied Physics Letters*, 62(6):606–608, 1993.

[5] F. Zappa *et al.* Principles and features of single-photon avalanche diode arrays. *Sensors and Actuators A: Physical*, 140(1):103–112, 2007.

[6] Chockalingam Veerappan and Edoardo Charbon. A low dark count p-i-n diode based spad in cmos technology. *IEEE Transactions on Electron Devices*, 63(1):65–71, 2016.

[7] W.J.Kindt. *Geiger Mode Avalanche Photodiode Arrays: For spatially resolved single photon counting*. PhD thesis, Delft University of Technology, 03 1999.

[8] Samuel Burri and E. Charbon. Spad image sensors: from architectures to applications. *Imaging Systems and Applications, ISA 2012*, 06 2012.

[9] D. Bronzi *et al.* Spad figures of merit for photon-counting, photon-timing, and imaging applications: A review. *IEEE Sensors Journal*, 16(1):3–12, 2016.

[10] Yue Xu *et al.* A new modeling and simulation method for important statistical performance prediction of single photon avalanche diode detectors. *Semiconductor Science and Technology*, 31(6):065024, may 2016.

[11] Yux Xu, Ping Xiang, and Xiaopeng Xie. Comprehensive understanding of dark count mechanisms of single-photon avalanche diodes fabricated in deep sub-micron cmos technologies. *Solid-State Electronics*, 129:168 – 174, 2017.

[12] I. M. Antolovic *et al.* Nonuniformity analysis of a 65-kpixel cmos spad imager. *IEEE Transactions on Electron Devices*, 63(1):57–64, 2016.

[13] C Zhang. *CMOS SPAD Sensors for 3D Time-of-Flight Imaging, LiDAR and Ultra-High Speed Cameras*. PhD thesis, Delft University of Technology, 05 2019.

[14] M. Lee *et al.* A back-illuminated 3d-stacked single-photon avalanche diode in 45nm cmos technology. In *2017 IEEE International Electron Devices Meeting (IEDM)*, pages 16.6.1–16.6.4, 2017.

[15] E. Charbon. Spad based image sensors. *Technical Digest - International Electron Devices Meeting, IEDM*, 2015:10.2.1–10.2.4, 02 2015.

[16] Shingo Mandai and Edoardo Charbon. A 4 × 4 × 416 digital sipm array with 192 tdcs for multiple high-resolution timestamp acquisition. *Journal of Instrumentation*, 8:P05024, 05 2013.

[17] M. Mazzillo *et al.* Timing performances of large area silicon photomultipliers fabricated at stmicroelectronics. *IEEE Transactions on Nuclear Science*, 57(4):2273–2279, 2010.

[18] T. Frach *et al.* The digital silicon photomultiplier — principle of operation and intrinsic detector performance. In *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, pages 1959–1965, 2009.

[19] Chockalingam Veerappan *et al.* A 160x128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter. In *2011 IEEE International Solid-State Circuits Conference*, pages 312–314, San Francisco, CA, USA, February 2011. IEEE.

[20] Marek Gersbach *et al.* A Time-Resolved, Low-Noise Single-Photon Image Sensor Fabricated in Deep-Submicron CMOS Technology. *IEEE J. Solid-State Circuits*, 47(6):1394–1407, June 2012.

[21] J. A. Richardson *et al.* Scaleable single-photon avalanche diode structures in nanometer cmos technology. *IEEE Transactions on Electron Devices*, 58(7):2028–2035, 2011.

[22] Roland H. Haitz. Studies on optical coupling between silicon p-n junctions. *Solid-State Electronics*, 8(4):417 – 425, 1965.

[23] Cristiano L. Niclass. *Single-photon image sensors in CMOS: picosecond resolution for three-dimensional imaging*. PhD thesis, École polytechnique fédérale de Lausanne, Lausanne, 2008.

[24] F. Sun *et al.* A simple analytic modeling method for spad timing jitter prediction. *IEEE Journal of the Electron Devices Society*, 7:261–267, 2019.

[25] Josef Blazej and Ivan Prochazka. Avalanche photodiode output pulse rise-time study. *Proc SPIE*, 7355, 05 2009.

[26] A. Spinelli and A. L. Lacaita. Physics and numerical simulation of single photon avalanche diodes. *IEEE Transactions on Electron Devices*, 44(11):1931–1943, 1997.

[27] Z. L. Yuan *et al.* High speed single photon detection in the near infrared. *Applied Physics Letters*, 91(4):041114, 2007.

[28] Heide F., Diamond S., and Lindell D.B. Sub-picosecond photon-efficient 3d imaging using single-photon sensors. *Sci Rep*, 8(17726):261–267, 2018.

[29] Nolet Frederic *et al.* Quenching circuit and spad integrated in cmos 65 nm with 7.8 ps fwhm single photon timing resolution. *Instruments*, 2:19, 09 2018.

[30] Giuseppe Intermite *et al.* Fill-factor improvement of si cmos single-photon avalanche diode detector arrays by integration of diffractive microlens arrays. *Opt. Express*, 23(26):33777–33791, Dec 2015.

[31] Myung-Jae Lee and Edoardo Charbon. Progress in single-photon avalanche diode image sensors in standard CMOS: From two-dimensional monolithic to three-dimensional-stacked technology. *Japanese Journal of Applied Physics*, 57(10):1002A3, sep 2018.

[32] Silvano Donati, Giuseppe Martini, and Enrico Randone. Improving photodetector performance by means of microoptics concentrators. *J. Lightwave Technol.*, 29(5):661–665, Mar 2011.

[33] Juan Pavia, Martin Wolf, and E. Charbon. Measurement and modeling of microlenses fabricated on single-photon avalanche diode arrays for fill factor recovery. *Optics express*, 22:4202–13, 02 2014.

[34] S. Donati *et al.* Uniformity of concentration factor and back focal length in molded polymer microlens arrays. In *CLEO/QELS: 2010 Laser Science to Photonic Applications*, pages 1–2, 2010.

[35] P. Connolly *et al.* High concentration factor diffractive microlenses integrated with cmos single-photon avalanche diode detector arrays for fill-factor improvement. *Applied optics*, 59(14):4488–4498, May 2020.

[36] A Ronchini Ximenes. *Modular time-of-flight image sensor for light detection and ranging: A digital approach to LIDAR*. PhD thesis, Delft University of Technology, 07 2019.

[37] E. Charbon, C. Bruschini, and M. Lee. 3d-stacked cmos spad image sensors: Technology and applications. In *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pages 1–4, 2018.

[38] S. Lindner *et al.* A high-pde, backside-illuminated spad in 65/40-nm 3d ic cmos pixel with cascoded passive quenching and active recharge. *IEEE Electron Device Letters*, 38(11):1547–1550, 2017.

[39] Juan Mata Pavia *et al.* A 1 × 400 Backside-Illuminated SPAD Sensor With 49.7 ps Resolution, 30 pJ/Sample TDCs Fabricated in 3D CMOS Technology for Near-Infrared Optical Tomography. *IEEE J. Solid-State Circuits*, 50(10):2406–2418, October 2015.

[40] R. K. Henderson *et al.* 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)*, pages 106–108, 2019.

[41] Mark Steigemann and Maxim Kulesh. Lidar system and method of operating the lidar system, U.S. Patent 16/207036, November 30, 2018.

[42] Cristiano Niclass *et al.* A 100-m Range 10-Frame/s 340x96-Pixel Time-of-Flight Depth Sensor in 0.18um CMOS. *IEEE J. Solid-State Circuits*, 48(2):559–572, February 2013.

[43] Cristiano Niclass *et al.* A 0,18um CMOS SoC for a 100-m-Range 10-Frame/s 200times ,96-Pixel Time-of-Flight Depth Sensor. *IEEE J. Solid-State Circuits*, 49(1):315–330, January 2014.

[44] A. Ronchini Ximenes. *Modular time-of-flight image sensor for light detection and ranging: A digital approach to LIDAR*. PhD thesis, Delft University of Technology, 7 2019.

[45] Maik Beer *et al.* Background Light Rejection in SPAD-Based LiDAR Sensors by Adaptive Photon Coincidence Detection. *Sensors*, 18(12):4338, December 2018.

[46] Sam W. Hutchings *et al.* A Reconfigurable 3-D-Stacked SPAD Imager With In-Pixel Histogramming for Flash LIDAR or High-Speed Time-of-Flight Imaging. *IEEE J. Solid-State Circuits*, 54(11):2947–2956, November 2019.

[47] Elham Sarbazi, Majid Safari, and Harald Haas. The impact of long dead time on the photocount distribution of spad receivers. In *2018 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2018.

[48] D. Snyder and Michael I. Miller. Random point processes in time and space. 1991.

[49] Joshua Rapp *et al.* Dead time compensation for high-flux ranging. *IEEE Transactions on Signal Processing*, 67(13):3471–3486, 2019.

[50] Yeomyung Kim and Tae Wook Kim. An 11 b 7 ps Resolution Two-Step Time-to-Digital Converter With 3-D Vernier Space. *IEEE Trans. Circuits Syst. I*, 61(8):2326–2336, August 2014.

[51] Wonsik Yu, KwangSeok Kim, and SeongHwan Cho. A 0.22 ps rms Integrated Noise 15 MHz Bandwidth Fourth-Order sigma-delta Time-to-Digital Converter Using Time-Domain Error-Feedback Filter. *IEEE J. Solid-State Circuits*, 50(5):1251–1262, May 2015.

[52] Hechen Wang, Fa Foster Dai, and Hua Wang. A 330μW 1.25ps 400fs-INL vernier time-to-digital converter with 2D reconfigurable spiral arbiter array and 2nd order sigma-delta linearization. In *2017 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–4, Austin, TX, April 2017. IEEE.

[53] P. Dudek, S. Szczepanski, and J.V. Hatfield. A high-resolution CMOS time-to-digital converter utilizing a Vernier delay line. *IEEE J. Solid-State Circuits*, 35(2):240–247, February 2000.

[54] Sungjin Kim, Taeik Kim, and Hojin Park. A 0.63ps, 12b, synchronous cyclic TDC using a time adder for on-chip jitter measurement of a SoC in 28nm CMOS technology. In *2014 Symposium on VLSI Circuits Digest of Technical Papers*, pages 1–2, Honolulu, HI, USA, June 2014. IEEE.

[55] Antti Mantyniemi, Timo Rahkonen, and Juha Kostamovaara. A CMOS Time-to-Digital Converter (TDC) Based On a Cyclic Time Domain Successive Approximation Interpolation Method. *IEEE J. Solid-State Circuits*, 44(11):3067–3078, November 2009.

[56] Hayun Chung, Hiroki Ishikuro, and Tadahiro Kuroda. A 10-Bit 80-MS/s Decision-Select Successive Approximation TDC in 65-nm CMOS. *IEEE J. Solid-State Circuits*, 47(5):1232–1241, May 2012.

[57] Sungjin Kim *et al.* A 0,6V 1,17ps PVT tolerant and synthesizable time to digital converter using stochastic phase interpolation with 16 times spatial redundancy in 14nm FinFET technology. In *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pages 1–3, San Francisco, CA, USA, February 2015. IEEE.

[58] Justin Richardson *et al.* A 32x32 50ps resolution 10 bit time to digital converter array in 130nm CMOS for time correlated imaging. In *2009 IEEE Custom Integrated Circuits Conference*, pages 77–80, San Jose, CA, USA, September 2009. IEEE.

[59] Chao Zhang *et al.* A 30-frames/s, 252x144 SPAD Flash LiDAR With 1728 Dual-Clock 48.8-ps TDCs, and Pixel-Wise Integrated Histogramming. *IEEE J. Solid-State Circuits*, 54(4):1137–1151, April 2019.

[60] Augusto Ximenes, Preethi Padmanabhan, and Edoardo Charbon. Mutually Coupled Time-to-Digital Converters (TDCs) for Direct Time-of-Flight (dTOF) Image Sensors. *Sensors*, 18(10):3413, October 2018.

[61] Matthew Z. Straayer and Michael H. Perrott. A Multi-Path Gated Ring Oscillator TDC With First-Order Noise Shaping. *IEEE J. Solid-State Circuits*, 44(4):1089–1098, April 2009.

[62] A. Elshazly *et al.* A noise-shaping time-to-digital converter using switched-ring oscillators—analysis, design, and measurement techniques. *IEEE Journal of Solid-State Circuits*, 49(5):1184–1197, 2014.

[63] John A. McNeill and David Ricketts. *The designer's guide to jitter in ring oscillators*. Designer's guide book series. Springer, Berlin ; New York, 2009. OCLC: ocn373556056.

[64] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers. Matching properties of mos transistors. *IEEE Journal of Solid-State Circuits*, 24(5):1433–1439, 1989.

[65] J. Jalil, M. B. I. Reaz, and M. A. M. Ali. Cmos differential ring oscillators: Review of the performance of cmos ros in communication systems. *IEEE Microwave Magazine*, 14(5):97–109, 2013.

[66] Demartinos A. C. *et al.* Delay Elements Suitable for CMOS Ring Oscillators. *JESTR*, 9(4):98–101, aug 2016.

[67] Demartinos A.C. *et al.* A 3GHz VCO suitable for MIPI M-PHY serial interface. In *2015 10th International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS)*, pages 1–6, Napoli, Italy, apr 2015. IEEE.

[68] L. Heller *et al.* Cascode voltage switch logic: A differential cmos logic family. In *1984 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, volume XXVII, pages 16–17, 1984.

[69] D. Z. Turker, S. P. Khatri, and E. Sanchez-Sinencio. A dcvsl delay cell for fast low power frequency synthesis applications. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 58(6):1225–1238, 2011.

[70] Adel S. Sedra and Kenneth C. Smith. *Microelectronic Circuits*. Oxford University Press, fifth edition, 2004.

[71] S. Suman, K. G. Sharma, and P. K. Ghosh. Analysis and design of current starved ring vco. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 3222–3227, 2016.

[72] T. Sakurai and A. R. Newton. Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas. *IEEE Journal of Solid-State Circuits*, 25(2):584–594, 1990.

[73] J. McNeill. Jitter in ring oscillators. In *Proceedings of IEEE International Symposium on Circuits and Systems - ISCAS '94*, volume 6, pages 201–204 vol.6, 1994.

[74] A. Hajimiri, S. Limotyrakis, and T.H. Lee. Jitter and phase noise in ring oscillators. *IEEE J. Solid-State Circuits*, 34(6):790–804, June 1999.

[75] F. Herzel and B. Razavi. A study of oscillator jitter due to supply and substrate noise. *IEEE Trans. Circuits Syst. II*, 46(1):56–62, January 1999.

[76] John A. McNeill and David Ricketts. *The designer's guide to jitter in ring oscillators*. Designer's guide book series. Springer, Berlin ; New York, 2009. OCLC: ocn373556056.

[77] A.A. Abidi. Phase Noise and Jitter in CMOS Ring Oscillators. *IEEE J. Solid-State Circuits*, 41(8):1803–1816, August 2006.

[78] T. C. Weigandt, Beomsup Kim, and P. R. Gray. Analysis of timing jitter in cmos ring oscillators. In *Proceedings of IEEE International Symposium on Circuits and Systems - ISCAS '94*, volume 4, pages 27–30 vol.4, 1994.