CMOS BASED SINGLE PHOTON AVALANCHE DIODE AND TIME-TO-DIGITAL CONVERTER TOWARDS PET IMAGING APPLICATIONS

CMOS BASED SINGLE PHOTON AVALANCHE DIODE AND TIME-TO-DIGITAL CONVERTER TOWARDS PET IMAGING APPLICATIONS

By
Zeng Cheng
B.A.Sc. Xi'an Jiaotong University, 2009
M.A.Sc. Xi'an Jiaotong University, 2012

A Thesis
Submitted to the School of Graduate
Studies in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

McMaster University
Hamilton, Ontario, Canada

Doctor of Philosophy (2016)	McMaster University
(Electrical and Computer Engineering)	Hamilton, Ontario

TITLE:	CMOS Based Single Photon Avalanche Diode and Time-to-Digital Converter towards PET Imaging Applications
AUTHOR:	Zeng Cheng, B.A.Sc. Xi’an Jiaotong University, Xi’an, China M.A.Sc. Xi’an Jiaotong University, Xi’an, China
SUPERVISOR:	Prof. M. Jamal Deen Prof. Hao Peng (Co-supervisor)
NUMBER OF PAGES:	xvii, 139

Abstract

Time-of-flight (ToF) positron emission tomography (PET) is a non-invasive, functional biomedical imaging modality. It can be used to determine the metabolic activity difference between lesion and normal cells, thus, making it possible to detect tumors at their early stages. In such a system, photomultiplier tubes (PMTs) are typically employed as the photon sensors. However, PMTs are of large size, have fragile package, consume large power, require high bias voltage and are costly. To pursue the benefits of ToF PET system, there is a growing need for research on new types of photodetectors and photo-detection systems.

This work focuses on studying and building a compact, low-cost time-of-flight photo-sensing system. To achieve this goal, we choose a standard digital CMOS technology to design and fabricate the photodetectors and associated electronics. A CMOS single photon avalanche diode (SPAD) is selected because of its low-cost, ultra-high light sensitivity and fast speed. Being implemented in IBM's 130nm CMOS process, the impacts of silicide layer on the overall performances of the SPAD are investigated. By eliminating silicide on the active area of SPAD, a fivefold improvement in both dark noise and photon detection efficiency are demonstrated. Then, a SPAD comprehensive analytical model is proposed and implemented in Verilog-A hardware description language. This model includes all the noise sources and will provide useful guidance in optimizing SPAD and the associated circuits.

A time-to-digital converter (TDC) is proposed and designed in the same 130nm CMOS process. The TDC is capable to digitize time intervals with 7.3ps resolution and covers up to 9ns dynamic range. The proposed TDC achieves state-of-the-art low power consumption. It will be used to extract the time-of-flight information, and to improve the image's single-to-noise ratio (SNR) in a ToF PET imaging system.

Finally, a time-of-flight sensing system prototype is built by integrating the CMOS SPADs and TDC on a printed circuit board. Based on the preliminary measurement results, this system achieves 440ps coincidence timing resolution. Factors of 2.5 and 6.1 improvements in image SNR and effective sensitivity, respectively, are expected with this prototype in ToF PET imaging applications.

Acknowledgements

Firstly, I would like to express my sincere gratitude and appreciation to my supervisor, Distinguished University Professor Dr. M. Jamal Deen, for his support and guidance over the past four years. It is a great honor and fortune for me to work under the supervision of such an outstanding mentor and scholar. Being always motivational and supportive, he has helped, encouraged and guided me throughout this research work. He is a role model to me, from professional research, to teaching and to personal aspect, and so on. I will never stop learning from him.

I would also like to express my deep appreciation and thanks to my co-supervisor Professor Hao Peng for his help, guidance and encouragement. He guided me into this interdisciplinary research. His door is always open for me for technical supports and insightful discussions. All these helps are greatly appreciated.

My sincerest thanks also go to Professor Nicola Nicolici and Professor Chih-Hung Chen for being members of my supervisory committee. They offer insightful advices and support during my research and take their valuable time to review my thesis.

Next, I would like to the administrative and technical staff of my department at McMaster University: Cheryl Gies, Tyler Ackland and Ron Harwood. Especially, Tyler spent lots of time helping me with the printed circuit boards used in my measurements.

Thanks Canadian Microelectronics Corporation (CMC) for their timely technical support and arranging the fabrication of my test chips.

My thanks also go to my dear colleagues at the Nanoelectronics-Optoelectronics Research Laboratory. Dr. Zhiyun Li, Dr. Darek Palubiak and Dr. Tianyi Guo offered numerous help and suggestion for my work. It is also a great pleasure and memorable experience working together with the other members, Hythm Afifi, Hani Alhemsi, Mrwan Alayed, Yiheng Qin, Si Pan and Sumit Majumder.

Last but not the least, I sincerely thank my parents for their unconditional support and encouragement over all these years, and my dear wife, Xiaoqing Zheng, for her endless love and for always being there for me and putting me first. This thesis is dedicated to them.

Table of Contents

List of Abbreviations

ADC	Analog-to-Digital Converter
ADPLL	All-Digital Phase-Locked-Loop
APD	Avalanche Photodiode
ASIC	Application Specific Integrated Circuit
CIS	CMOS Imaging Sensor
CMOS	Complementary Metal-Oxide-Semiconductor
CR	Counting Rate
CT	Computed Tomography
DCR	Dark Count Rate
DFF	D-Flip-Flop
DLL	Delay-Locked-Loop
DNL	Differential Nonlinearity
dSiPM	Digital Silicon Photomultiplier
DSM	Deep Sub-Micron
DT	Dead Time
FDG	Fluorodeoxyglucose
FF	Fill Factor
FLIM	Fluorescence Lifetime Imaging Microscopy
FoV	Field-of-View
FPGA	Field Programmable Gate Array
FWHM	Full-Width at Half-Maximum
GRO	Gated Ring Oscillator
HDL	Hardware Description Language
IAT	Inter-Avalanche Time
INL	Integral Nonlinearity
I/O	Input/Output

I-V	Current-Voltage
LoR	Line-of-Response
LSB	Least Significant Bit
LUT	Look-Up-Table
MRI	Magnetic Resonance Imaging
PCB	Printed Circuit Board
PDE	Photon Detection Efficiency
PET	Positron Emission Tomography
PMT	Photomultiplier Tube
PQPR	Passive Quench, Passive Reset
PVT	Process, Voltage and Temperature
RO	Ring Oscillator
SEM	Scanning Electron Microscope
SiPM	Silicon Photomultiplier
SNR	Signal-to-Noise Ratio
SPAD	Single Photon Avalanche Diode
SRH	Shockley-Read-Hall
TDC	Time-to-Digital Converter
TDL	Tapped-Delay-Line
TEM	Transmission Electron Microscope
ToF	Time-of-Flight
VDL	Vernier Delay Line

List of Figures

List of Tables

Chapter 1

INTRODUCTION*

Optical, photonic, and optoelectronics components and systems are intensely studied and deployed in the field of biomedical imaging applications. Examples include Positron Emission Tomography (PET), Raman Spectroscopy, and Fluorescence Lifetime Imaging Microscopy (FLIM). This is mostly because of the non-invasive property of optical sensing and imaging techniques. In these applications, traditional photomultiplier tubes (PMTs) are used as the optical sensors due to their excellent performance in terms of excellent detection efficiency in the visible light range, fast response time and high single-to-noise ratio (SNR). However, they suffer from some important disadvantages such as bulky size, high operating voltage, low integration capability and high cost. Moreover, their performances will be greatly degraded under a high magnetic field. This limits their usage in some medical imaging applications, for example in magnetic resonance imaging (MRI). Thus, a study of optical sensing systems based on detectors not affected by magnetic fields, for example using Complementary Metal-Oxide-Semiconductor (CMOS) technology, is of great importance.

In this chapter, first, we introduce the application background of a PET imaging system. Then, a brief review of commonly used optical sensors and timing circuits will be provided. Next, the research motivation is highlighted, followed by the thesis organization and research contributions from this study.

* Part of this work was published as: Z. Cheng, X. Zheng, M. J. Deen and H. Peng, "Recent developments and design challenges of high-performance ring oscillator CMOS time-to-digital converters," IEEE Transactions on Electron Devices, vol. 63, no. 1, pp 235-251, 2016. Copyright granted by IEEE.

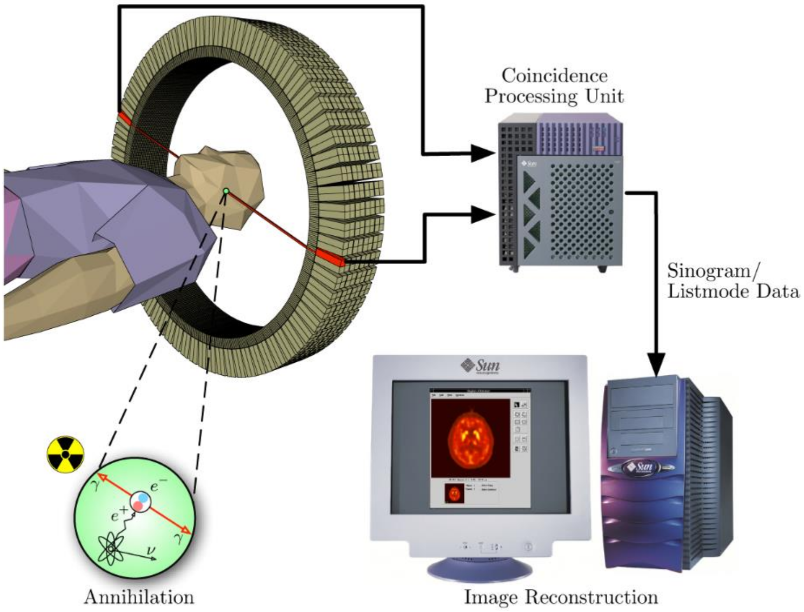


Figure 1-1, Schematic illustration of a PET imaging system. (Reprinted from [1], [2] with permission granted).

1.1. Application Background

1.1.1. Positron Emission Tomography (PET) Imaging

PET is a non-invasive imaging modality in nuclear medicine. As its name suggests, it is based on the coincidence detection of two 511keV gamma-rays emitted from positron annihilation events from radioactive tracers injected into the patient's body. PET imaging is categorized as functional imaging because of the fact that the reconstructed images can present the patient's physiology.

The schematic illustration of a PET imaging system [1], [2] is presented in Figure 1-1. As the first step in PET scanning, the radioactive tracer is injected into the patient's body. The tracers used in PET are of particular interest since they can be used to form analogs of common biological molecules. For example, ^{18}F is used to produce ^{18}F-fluorodeoxyglucose (FDG) which is analogous to glucose and it can be used to indicate levels of cellular metabolism, and ^{11}C is used in ^{11}C-L-methionine as it is analogous to the amino acid, which can indicate cancer malignancy based on the amino acid utilization [3]. Since these isotopes feature short half-life times (e.g. ^{18}F is 110 minutes and ^{11}C is 20.4 minutes), positrons are generated in form of radioactive decay (i.e. the β+ decay). After the positron is emitted it will travel a short distance through the surrounding tissue before it combines with an electron from surrounding atoms. Then, the mass of these two particles is converted to electromagnetic energy in an annihilation event. Two gamma-rays are then generated in opposite directions with energy of 511keV [4], [5].

To detect these annihilation resulting gamma-rays, PET imaging systems make use of scintillation crystals to convert the high energy photons to low energy photons in visible wavelength range. Then, solid-state photon detectors, such as PMTs, are employed to translate the optical signals to electrical signals. Usually, the scintillation crystals and detectors are coupled together and placed in a ring formation around the patient, as shown in Figure 1-1. The electrical signals will be filtered in a coincidence process unit in order to select true events that are from the same annihilation. Then, the data is used to reconstruct the images for the region under scanning.

Since PET imaging has high sensitivity to the biological and metabolic activity differences at the molecular level, it has become one of the most powerful modalities in nuclear medicine for cancer detection and treatment. PET is currently being used in various fields, including: oncology for cancer diagnosis, staging, and therapy [6]; cardiology in myocardial perfusion and viability studies, and coronary artery disease [7]; and neurology in the study of epilepsy, movement disorders, and Alzheimer's disease [8]. While radiopharmaceutical research will provide new applications for PET, improving the optoelectronic system's capabilities can provide new imaging techniques to increase the effectiveness of PET in current applications and open the door to new ways to apply PET in healthcare.

1.1.2. Time-of-Flight PET

The concept of time-of-flight (ToF) technique has been proposed and demonstrated in early 1980s [9], [10]. The ToF information is the time difference Δt in detection of the two 511keV gamma-rays events. ToF PET takes advantage of this timing information to correlate the detection events to the position Δx of the annihilation point with respect to the center of the field-of-view (FoV) according to the following equation.

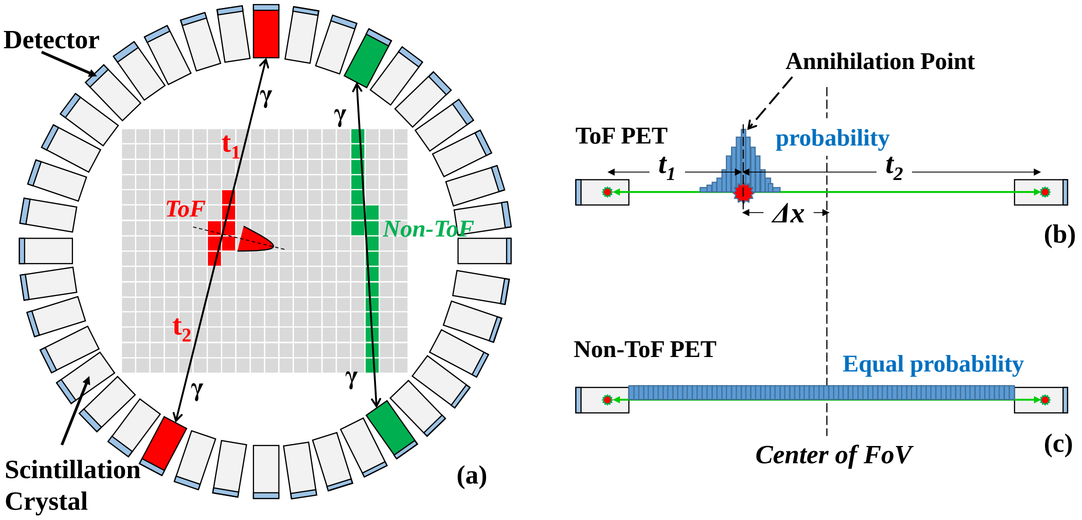


Figure 1-2, Concept diagram of ToF PET comparing with conventional PET.

$$t_1 = \frac{{}^{1}\!/_{2} D - \Delta x}{c}, t_2 = \frac{{}^{1}\!/_{2} D + \Delta x}{c} \Rightarrow \Delta x = \frac{\Delta t \times c}{2}, \tag{1-1}$$

where *c* is the speed of light. It is noted that the factor of 2 in equation (1-1) is introduced since the Δ*x* is the distance between the annihilation point and the center of the FoV.

The concept of ToF PET and its differences with the conventional PET are shown in Figure 1-2.

In conventional (non-ToF) PET, each detected photon is tagged with a detector position and a detection time. If the detection time difference between two photons is smaller than a set coincidence window (e.g. traditionally 5-10 ns), these two events are considered physically correlated to the same annihilation event. A line-of-response (LoR) between these two correlated detectors, as presented by the green part in Figure 1-2(a). Thus, in the non-ToF PET, the timing information is only used to filter the time coincidence events. When a LoR is determined, the point that the annihilation occurs can happen anywhere along this LoR. All the voxels along this LoR are assigned the same probability of emission to locate the annihilation position, as shown in Figure 1-2(a) and Figure 1-2(c). This will accumulate the noise contributions from all the voxels and degrade the SNR, which subsequently results in image blurring and degraded image contrast [11]–[14].

However, in the ToF PET, the ToF difference Δt is directly related to the distance Δx by Equation (1-1). In Figure 1-2(a), the time t_1 is proportional to the distance between the annihilation point and the top detector, and the time t_2 is proportional to the distance between the annihilation point and the bottom detector. Because of the advantageous ToF information, only a few voxels on the LoR will be considered for the emission source. Since the measured timing information (t_1 and t_2) have jitters, the coincidence timing resolution is limited. This causes a distribution spread in the location probability as shown in Figure 1-2(b), with its most possible position Δx corresponding the measured ToF value.

Even though the ToF concept could not improve the spatial resolution in a PET imaging system [13], it helps reducing the statistical noise in the reconstructed image and enhance the image contrast. The following equation gives the benefit in terms of the gain in SNR,

$$\frac{SNR_{ToF}}{SNR_{Non-ToF}} = \sqrt{\frac{D}{\Delta x}} = \sqrt{\frac{2D}{c \times \Delta t}}, \tag{1-2}$$

where SNR_{ToF} and $SNR_{Non\text{-}ToF}$ are the SNRs of the reconstructed images from ToF PET imaging and conventional PET imaging, respectively, and D is the diameter of the object under scanning [11], [14]–[16].

Another term to represent the advance available from ToF PET is the effective sensitivity increase (G), as presented in the following equation [11], [13],

$$G = \frac{2D}{\mathrm{c} \times \Delta t}, \tag{1-3}$$

With coincidence timing resolution as fine as hundreds of picoseconds, considerable SNR gain and effective sensitivity increase in the reconstructed images are achievable from ToF PET system. An illustration of the difference between these two PET systems is given in Figure 1-3. This implies better lesion detectability, reduced radioactive dose needed by the patient and/or reduced scan time [12].

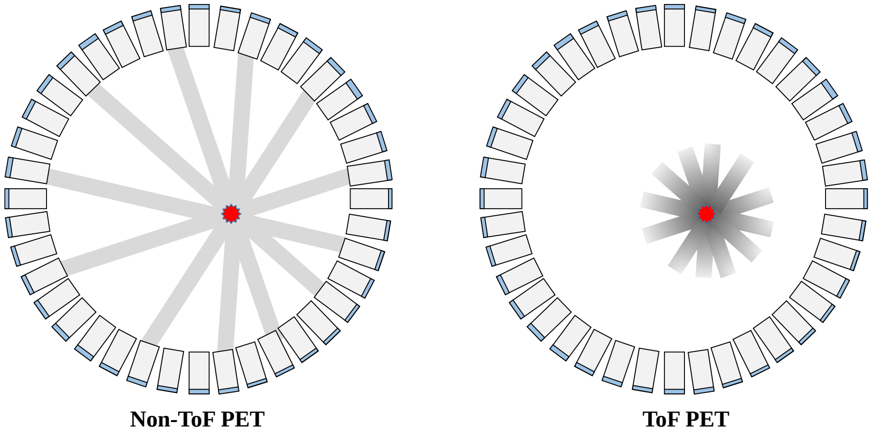


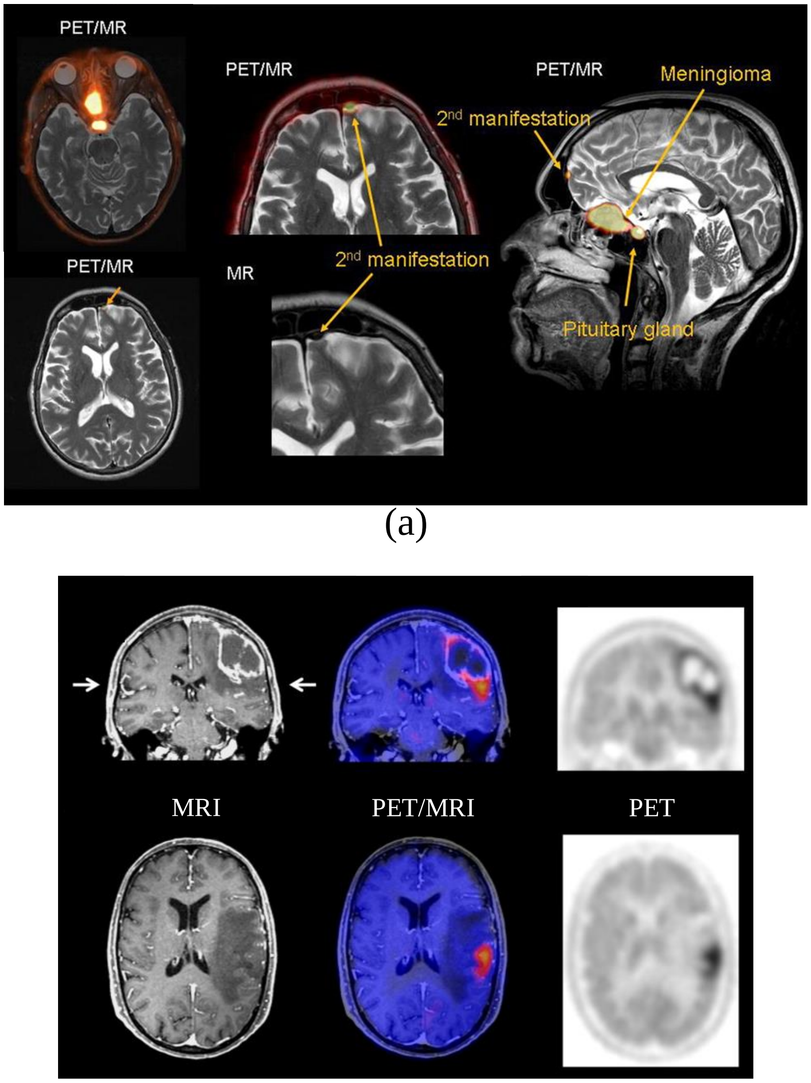
Figure 1-3, Illustration of the SNR increase of a ToF PET system over a conventional PET system.

1.1.3. Combined Imaging Modality: PET/MRI

Despite the advantages from the standalone PET scanning, it suffers from image blurring due to patient movement. And the low contrast in PET images increase the difficulty to locate the lesion region precisely. As an important technical evolution in the field of nuclear imaging, PET/MRI multimodality imaging can add anatomic features with high soft-tissue contrast available from MRI to the biological information delivered by PET imaging [17], [18]. Two examples of the images from PET/MRI are given in Figure 1-4 below [18], [19].

Figure 1-4(a) shows the clinical PET/MR images of a 71-year-old woman with frontobasal meningioma in the olfactory region [18]. In addition to the detection of the main tumor, the secondary smaller and previously unknown frontal meningioma was seen on PET scanning. It is not detected by computed tomography (CT) and hardly by MRI. In Figure 1-4(b) [19], the MRI images showed the contrast-enhancing lesion and surrounding edema. PET images showed the tumor’s location with better accuracy according to the increase of the isotope intensity. MRI images can help to improve the soft tissue contrast in PET images.

The combination of PET and MRI requires a solution to several technological challenges in terms of the photodetector. First, the photomultipliers used in PET scanner must be replaced by magnetic-field insensitive photon sensors. The strength of magnetic field in the magnet bore in a MRI system is typically between 0.5T and 10T, with 3T for most human MRI scanners. The performance of conventional PMTs used on PET system will be greatly deteriorated under such a high magnetic field strength



(a)

(b)

Figure 1-4, Two examples of the fusion images obtained from the combined PET/MRI imaging. (Reproduced from [18], [19] with permission granted. Copyright by the Society of Nuclear Medicine and Molecular Imaging Inc.)

due to their intrinsic analog properties. Second, the size of PET detectors needs to be compact. The detector ring for PET scanner has to be invisible to the MRI and must not interfere with the field gradients or the MR radio frequency [17], [18]. Thus, these requirements motivate the research and development in new types of photodetectors.

1.2. Brief Review on Existing Technologies

We divide the review of existing technologies into two categories. One is for optical detectors including PMTs and other types of silicon-based solid-state detectors. The

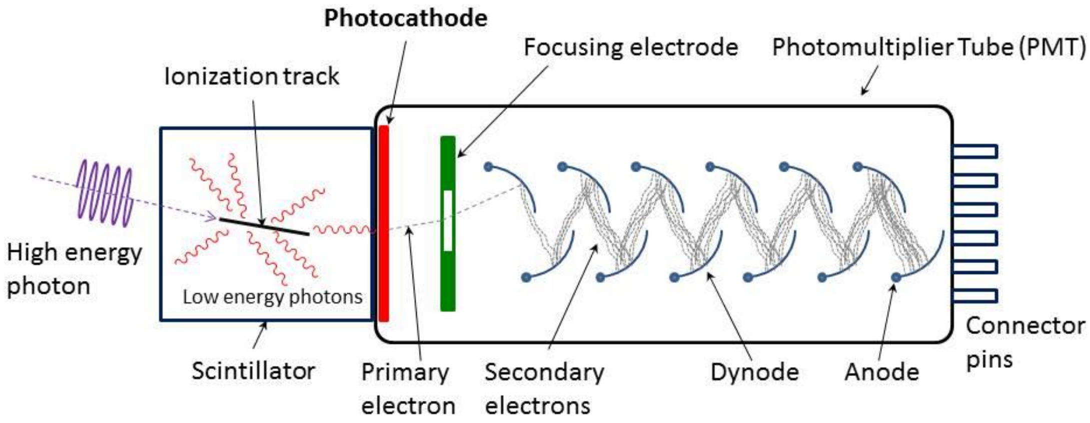


Figure 1-5, Schematic diagram of a typical PMT with a scintillation crystal. (Reprinted from [20] with permission granted.)

other is the timing circuitry, i.e. the time digitizer used to extract the ToF timing information.

1.2.1. Photodetector Technologies

The photodetector that is most commonly used in PET imaging systems is the PMT. PMTs are commonly coupled with scintillators to form the scintillation detectors due to their large area, fast response time, high sensitivity, high gain and low noise. As other types of solid-state photodetectors, the silicon-based photodiodes have attracted intensive research interests over the past decades [5].

(a) PMT

Figure 1-5 presents the typical schematic diagram of a PMT [20], which is coupled with a scintillator to convert the high energy gamma-ray in PET imaging system to low energy visible photons. Basically, a PMT is composed of a photocathode, a series of multiplication stage (i.e. a series of electrodes called *dynodes*) and an anode. All these components are enclosed in a vacuum environment by borosilicate glass.

When scintillation photons strike the photocathode, photoelectrons (i.e. primary electrons in Figure 1-5) will be released by the photoelectric effect given that the incident photon energy is much higher than the work function of the photocathode’s material. An important term called quantum efficiency is used to describe the probability of liberating photoelectrons in the PMT’s photocathode. It is usually measured as a ratio of the number of generated photoelectrons relative to the number of incident photons. The PMT's quantum efficiency is wavelength dependent and has a typical value of 20~30% [5].

The photoelectrons are then focused by the focusing electrode and accelerated towards the first dynode. The dynode plates are specifically shaped and arranged in a PMT to maximize the collection of electrons. Since each dynode is held at a higher voltage potential than the previous one, electrons will be accelerated under the electric field. Every time the electrons trike dynodes at high velocities, the plate will emit secondary electrons and they will accelerate towards the next plate. This process is repeated for 10 to 12 times depending on the number of dynodes in a PMT. Usually, a PMT delivers an amplification gain of 10^6 [5].

Besides the high gain, PMTs have other attractive properties such as stability, low dark current, linear amplification and low noise. These features make PMT the most used photodetector in a PET imaging system. Recently, several advanced modes of PMTs were developed and are now commercially available. For example, the micro PMT from Hamamatsu features very compact size of 40mm×30mm×12mm, around 5V input voltage and light weight of ~50g [21].

(b) Avalanche Photodiode (APD)

The avalanche photodiode (APD) is essentially a PN junction which works under the reverse-biased condition, as shown in Figure 1-6. A depletion region exists due to the drift of electrons and holes. And it becomes wider in the reverse bias than that in zero bias. When incident photos are absorbed, electron-hole pairs will be generated in the depletion region. These photon generated electron-hole pairs can migrate towards electrodes under the external applied electric field, thus contributing to the so-called photocurrent.

Because the depletion region in APD is very thin, this makes APD response very fast to incident photons, thus, APD can provide better timing performance than PMT. The quantum efficiency of APD is much higher than PMT (60% ~ 80%) as photoelectrons are not required to escape from the photocathode (which is the case in PMT to overcome the work function of the photocathode material). Besides, silicon-based APDs are easy to miniaturize and can be fabricated in standard semiconductor

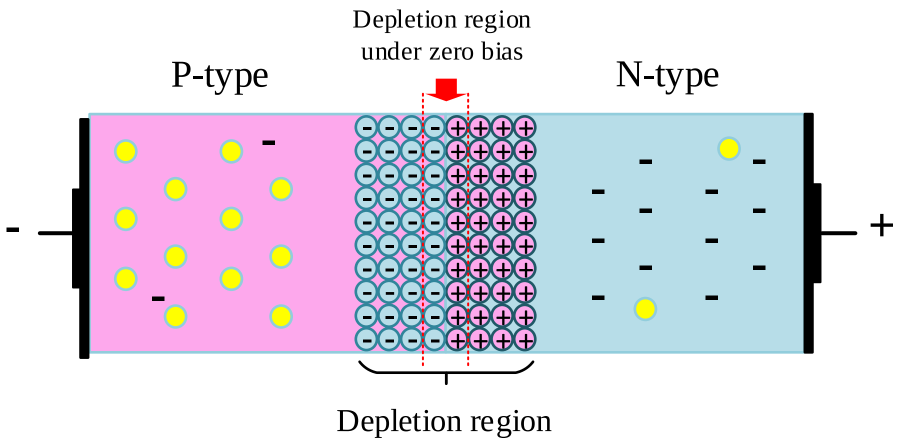


Figure 1-6, Depletion region of an APD device.

technologies. It is more robust, less expensive, and requires lower operating voltage when compared with PMT.

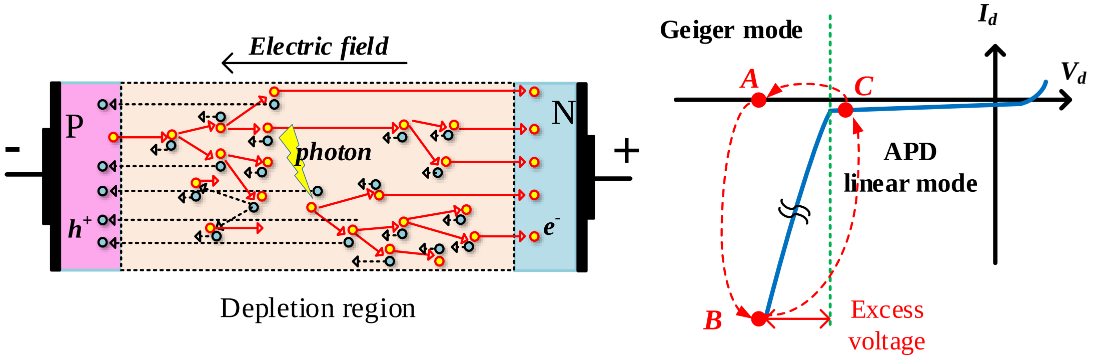
However, the internal gain available from APD is usually on the order of 10^3. This is much lower than that of PMT (~10^6) and is not sufficient to make it working as a single photon detector in a PET system. Therefore, low noise amplifiers are usually required for APDs [5], [22].

(c) Single Photon Avalanche Diode (SPAD)

Single photon avalanche diode (SPAD) is an APD being biased in the Geiger mode [5], [22], [23]. This mode refers to the condition that the photodiode is biased above its breakdown voltage.

In Geiger mode, the electric field in the depletion region of the SPAD is sufficiently high that both electrons and holes generated in this region can be multiplied by impact ionization process. As depicted in Figure 1-7(a), first, photon-generated electron-hole pair will be accelerated by the electric field, and then impact ionize to generate more electron-hole pairs. These secondary electron-hole pairs will further ionize more atoms. This multiplication process repeats in the depletion region before the free carriers are collected by electrodes. Thus, a self-sustained avalanche process is resulted in an SPAD.

Figure 1-7(b) shows the typical current/voltage characteristics of a SPAD device. The dashed line presents the breakdown point which also divide the reverse biasing into Geiger mode and APD linear mode. At the beginning, the SPAD is biased beyond



(a) (b)

Figure 1-7, Principle operation of SPAD. (a) The avalanche breakdown process and (b) The typical current/voltage curve.

the breakdown voltage and stays in OFF state (*A*). When a photon strikes its active (depletion) region and introduces initial carriers, avalanche multiplication process will be started. The SPAD is in ON state (*B*). Other than triggered photons, avalanche could happen due to other sources (as will be discussed in Chapter 2 and Chapter 3), and these are considered as dark noise. Consequently, a macroscopically avalanche current is built up flowing through the device. Usually, the avalanche current follows though a resistor to provide negative feedback. Thus, a large voltage drop across the resistor will bring the SPAD to the point *C*, where the avalanche cannot be self-sustained. Now, the SPAD is quenched, and it will be recharged back to the point *A* through the same quenching resistor that is connected to the supply voltage.

Due to the internal avalanche multiplication process in the device, SPAD provides a large gain (~10^6) that is comparable with PMTs. It features many attractive properties as a solid-state silicon-based photodetector, such as single photon sensitivity, fast response time, low dead time with different quenching/recharging circuits, low operating voltage, compact size and fully compatible with standard CMOS process. Especially, it can easily scale-up to large SPAD arrays and be integrated with peripheral electronics [24]. As a state-of-the-art example, the work in [25] implemented an 96×240 SPAD array, and integrated timing measurement and real-time energy counting circuits on chip. Being fabricated in a 130nm imaging CMOS process, the design only required 3.3V supply voltage, consumes 300mW power, and worked at 100Msps.

SPAD has several important performance metrics [24], [26]. Below we list several factors that reveal the optical property of a SPAD:

- As mentioned above, the SPAD can be triggered by sources besides light illumination and these false triggering brings dark noises to the SPAD. These sources will be discussed in detail in the following two chapters. Dark count rate (DCR) refers to the avalanche counts per second when a SPAD is falsely triggered by means other than photons. It is a parameter that presents the noise level of a SPAD. Typically, it is greatly affected by the SPAD fabrication technology. It can vary from <100Hz to several MHz. DCR is also highly dependent on both excess bias voltage and temperature.
- Since some associated circuits (e.g. quenching and recharging resistors and transistors) are integrated with the SPAD and due to design rule requirement, only a portion of the surface area of a SPAD can be used to collect photons. Fill-factor (FF) is the ratio of the active area of a SPAD to the total pixel area. A SPAD design with large FF means that more silicon area are dedicated to photon sensing.
- Photon detection efficiency (PDE) is a measure of the ratio of the number of detected photons to the number of incident photons on the active area of the SPAD. This ratio depends on absorption probability and the triggering efficiency [24]. PDE changes with both excess voltage and wavelength.
- The SPAD timing jitter (or timing precision) is the statistical spread of output pulse of SPAD relative to the actual photon arrival time. This parameter is usually represented as the full-width at half-maximum (FWHM) of the distribution histogram of SPAD response time. CMOS SPAD features good timing performance and timing jitter can be less than 100ps [27].

The past decade has witnessed the development of SPADs with higher detection efficiency, lower dark count level, higher detection rate, and higher fill factor. They can be designed and fabricated in various CMOS processes, from the high performance but high cost custom process [28], [29], to inexpensive mainstream digital CMOS process [30]–[33]. A comprehensive review of SPADs developed in CMOS technology can be found in these references [24], [26].

(d) Silicon Photomultiplier (SiPM)

Even though PMT is still the mainstream photodetector used PET imaging system, its drawbacks make researchers start to investigate new types of semiconductor photodetectors. Among all the candidates mentioned earlier, silicon photomultiplier (SiPM) is a novel version of single-photon level photodetector in the photodiode family.

SiPM comprises a high-density matrix of SPADs that are connected in parallel. For example, the C-series SiPM that is commercially available from SensL has an array size varying from 14.2×14.2 mm^2 to 57.4×57.4 mm^2, containing a large number of SPAD microcells from 75920 (for 2×2 pixels) to 687456 (for 12×12 pixels) [34]. The output signal of the SiPM thus is proportional to the number of operating SPAD microcells (or summation of outputs from each pixel). SiPM has the high gain characteristic inherited from SPAD microcells ($\sim 10^6$). Thus, it eliminates the needs for external amplification circuits as required with APDs.

The PDE of a SiPM is typically dependent on the quantum efficiency, pixel fill-factor, SPAD microcell's triggering/breakdown probability and its recovery time [12]. The quantum efficiency of silicon is quite high for visible light photons (80-90%) and it heavily depends on wavelength [35]. UV range photons do not penetrate very deep into the material and electron-hole pairs are often lost due to short recombination time of surface defects. Longer wavelength IR photons penetrate deeper into silicon and require deeper depletion layers. The effective photon sensitive area on SiPM can be anywhere from 25% to 80% depending on the implementation of the detector. The breakdown probability is simply the probability that the initial electron-hole pair causes breakdown. This is related to the electric field strength in the SPAD, but can often approach 100% due to its very thin depletion region. The recovery time is the recharging period for a microcell. After firing, it typically takes the microcell about 1μs to recharge its capacitance. The product of these factors gives a PDE in the range of 14% to 51% [36], [37] for SiPM in recent commercial devices.

Due to the arrangement of the SiPM array, optical crosstalk may occur. For every 10^5 carriers produced above the band gap of silicon, approximately 3 photons are emitted [35]. These photons may go on to cause avalanche breakdown in neighboring microcells, adding to the overall noise via a stochastic process. This effect can be reduced by using lower gain, or introducing a physical barrier between microcells, but the cost is reduced gain and/or smaller sensitive area. This crosstalk, combined with the dark counts of individual microcells are the major sources of noise in SiPMs. Due to their much higher gain, however, SiPMs still have an SNR advantage over large area APD detectors.

It should be noted that a SiPM has a relatively higher DCR compared with a PMT. This is because of the thermally generated free carriers in the conduction band and the large number of microcells integrated in one device. Another important feature of SiPM is that it is immune to magnetic fields, which makes it an attractive choice for PET/MRI multimodality imaging as discussed before. All these features of SiPM have led to growing attention to use SiPM as an alternative to the traditional PMTs in low-light detection applications such as PET and FLIM imaging [12], [17], [38]–[40].

An important innovation in the development of SiPM, the digital SiPM (dSiPM), has become a hot research topic in the past few years [25], [30], [31], [41]–[54]. In dSiPM, one-bit analog-to-digital converters (ADCs) are integrated in each SPAD microcell to convert the avalanche pulses to digital signals immediately after the SPADs. The dSiPM allows direct access to each microcell in its pixel, thus the noisy cells (i.e. those cells that show considerably high DCRs) can be disabled. This feature can greatly reduce the median noise performance of the dSiPM. On-chip time-to-digital converters (TDCs) are also implemented to explore the timing information during the photon detection. Currently, dSiPM is commercially available from Philips [53], [55].

1.2.2. TDC Technologies

TDCs have been intensely used in the past decade due to their excellent timing resolutions and time stamping capabilities. These capabilities are important for various applications such as All-Digital Phase-Locked Loop (ADPLL) [56]–[58], and biomedical imaging including ToF PET [25], [43], [59], FLIM [49], [60] and various ToF ranging applications [61]–[63]. With the digitization function which is analogous to an ADC, a TDC converts time intervals to digital codes with sub-nanoseconds or even sub-picoseconds resolution when implemented in a 90nm CMOS technology [64]. In order to facilitate system integration and reduce cost, on-chip TDC in standard CMOS technology is the preferred choice [65].

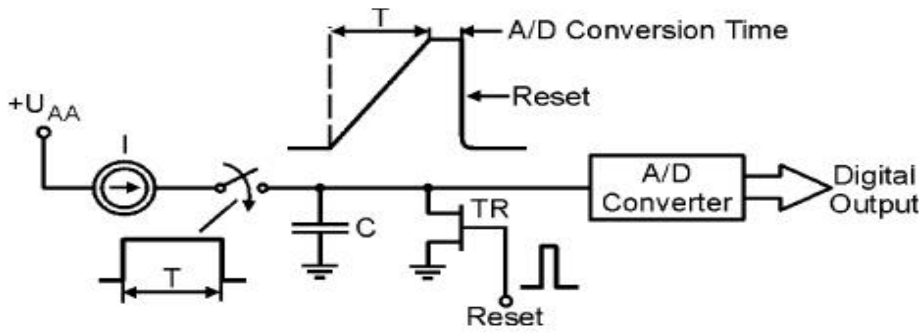


Figure 1-8, Conceptual illustration of an analog-type TDC. (Reprinted from [75] with permission granted.)

TDCs can be divided into two categories according to their operating principles. The analog-type TDCs use time-to-amplitude conversion followed by an ADC. These TDCs achieve good resolution and linearity at the expense of high power dissipation, large size, low scalability with CMOS technology nodes and high noise susceptibility. TDCs based on digital techniques can be further classified as delay-locked-line (DLL) [66], [67], Vernier-delay-line (VDL) [68], and multiple interpolations [69]–[71].

Another variation of the digital types is the ring-oscillator (RO) based TDC [16]. This type of RO-based TDCs is very suitable for array implementation because the ring oscillator used to generate the required multiple-phases for the whole array can be shared on-chip [45], [60], [72], [73]. Also, RO-based TDCs can simultaneously achieve high resolution and wide dynamic range, with small area and low power dissipation. A comprehensive discussion on RO-based TDCs can be found in [15].

Next, a brief discussion and comparison of conventional analog-type TDCs and digital TDCs implemented in both CMOS and field programmable gate array (FPGA) technologies is given. A more detailed review of these TDCs can be found in [74]–[76].

(a) Analog-type TDCs

Generally, in an analog-type TDC, a capacitor is discharged by a constant current source during the time interval to be measured. This will produce a voltage change across the capacitor, which is linearly dependent on the discharging time. Then, the voltage change will be converted to digital codes by either a voltage comparator or a conventional ADC. This concept is presented in Figure 1-8 [75].

In order to get better resolution, the pulse stretching (or time stretching) method is often employed in analog-type TDCs. Figure 1-9 shows the conceptual diagram of such

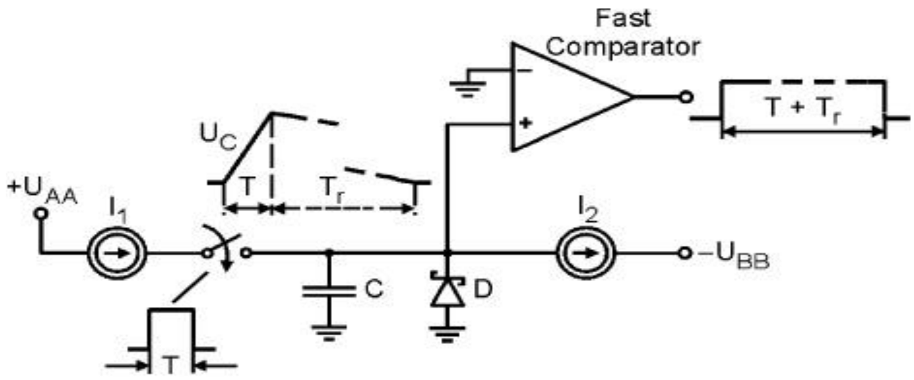


Figure 1-9, Conceptual illustration of a time stretching TDC. (Reprinted from [75] with permission granted.)

an example [75]. In such a method, the capacitor *C* is discharged by a current source I_1. Then, the capacitor is recharged back to the same voltage by another current source I_2, which is *K* times smaller than I_1. Thus, a stretching/amplification factor *K* is obtained, and the time resolution can be improved by the same factor [77]–[80]. For example, in [78], using discrete logic circuits, a pulse-stretching factor of about 5000 and a time resolution of 7.8ps were achieved.

To further improve the stretching factor and time resolution of the TDC, the dual-slope pulse stretching approach can be used [81], [82]. In this method, a second capacitor is used to take advantage of both the current ratio and the capacitance ratio. First, the node (with an initial voltage of V_0) is discharged by the capacitor C_1 with the current source I_1, during the time interval to be measured. Then, a larger capacitor C_2 (=MC_1) starts to recharge the node by a smaller current source I_2 (=I_1/N), until the voltage of the node returns to its initial value. A counter driven by a reference clock records the number of clock cycles during this charging period. In this way, the interpolation factor (= *MN*) can be very large, yielding an enhanced time resolution of the TDC.

Examples of the dual-slope TDC are now discussed. The TDC implemented in a 0.8µm BiCMOS process in [81] achieved 32ps resolution, and 2.5µs dynamic range with a 100MHz reference clock. However, it had the drawback of a large power consumption (350mW) and large temperature drift (±125%, i.e. ±40ps in a 100ºC range). Another implementation in [82] achieved 50ps resolution, 250ns dynamic range, and 0.75mW power consumption with a 80MHz reference clock in 0.35µm CMOS process. Here, the temperature drift was from -15.2% to +13% over a temperature range of -40ºC to 80ºC. A state-of-the-art analog-type TDC in [83] proposed a triple-slope time stretching scheme and implemented it in a 0.35 µm CMOS technology. This TDC had 357ps resolution and 1.46µs dynamic range with a 175MHz reference clock. However, the temperature drift performance was not reported.

Through careful design and layout, the analog-type TDC can have excellent precision and linearity for short measurement ranges. But it suffers from a large temperature drift, hence, a low stability, and a degraded INL for wide dynamic range implementations. Moreover, since the analog-type TDC uses large area consuming capacitors, it is very costly in terms of area, especially for integrated implementations. Also, the dead time of the analog-type TDC will limit its high-speed applications. In addition, the analog nature makes it hard to migrate to advanced CMOS technologies [80]–[83].

(b) Digital CMOS TDCs

A widely used type of digital TDC is based on the delay lines consisting of several stages of buffers or inverters. As shown in Figure 1-10, the *Start* pulse propagates through the delay line, and its statuses along the delay line will be recorded by flip-flops being triggered upon the arrival of the *Stop* pulse. Then, the time interval between *Start* and *Stop* signals is obtained by decoding the statuses in the flip-flops [84], [85]. This basic tapped delay line architecture offers a relatively straightforward approach to perform the time-to-digital conversion. However, its resolution is typically limited by the propagation delay of an inverter, which is the shortest delay specified in a certain CMOS technology. Also, the delay of a delay unit is subject to process, voltage and temperature (PVT) variations, unless the DLL technique is used [86]–[88]. Using a DLL as in Figure 1-11, the delay of the delay unit is voltage-controllable and is locked to a fraction of a reference clock’s period, regardless of ambient or process variations. A phase detector, charge pump and loop filter are used to detect the frequency/phase difference and provide a negative feedback.

The VDL, in which two chains of different delay units are used, is a solution to achieve sub-gate delay in a standard CMOS technology. For example, 5ps resolution was demonstrated with a VDL TDC in a 0.7µm CMOS process [68]. By introducing

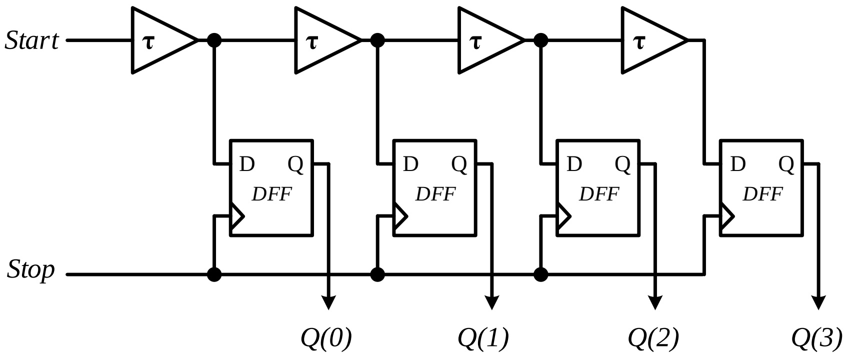


Figure 1-10, Diagram of a conventional tapped delay line.

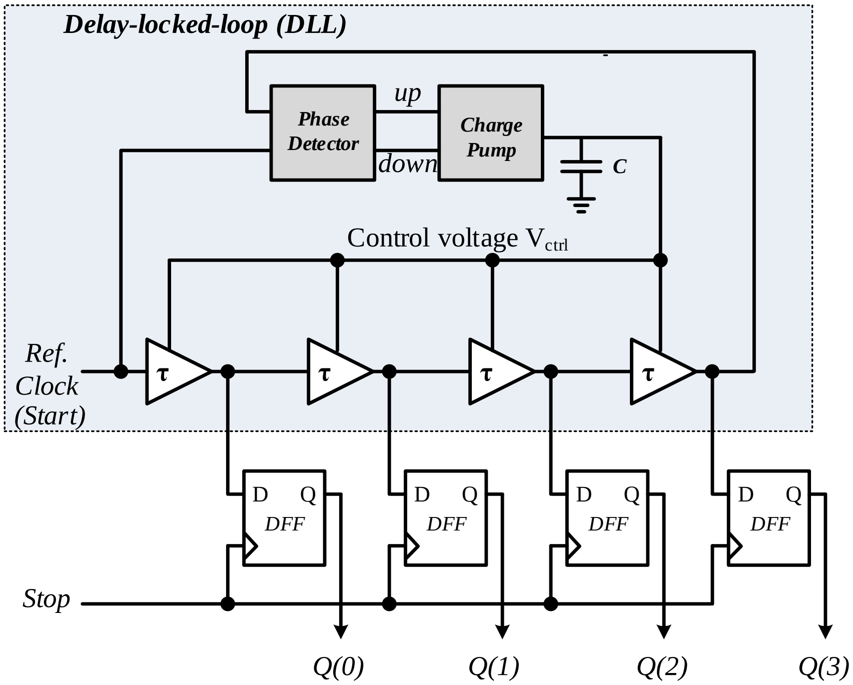


Figure 1-11, Diagram of a DLL TDC.

additional levels of interpolations using both VDL and tapped DLL, the resolution was improved to 1ps with a hierarchical VDL topology implemented in a 90nm CMOS process [89]. VDL-based TDCs can get superior (sub-gate) time resolution, but their dynamic range are usually limited by the total number of delay units, which would increase exponentially in order to cover a large dynamic range. To solve this problem, 2-D VDL-based TDC was proposed [90]. By arranging the delay lines in two orthogonal dimensions, the total length in this 2-D VDL was reduced. This yielded a better efficiency since only a third of the delay lines was required to cover the same dynamic range as compared to the traditional linear VDL.

Another way to achieve fine time resolution and cover a wide dynamic range is to use the multiple interpolation method. For example, a two-stage coarse-fine TDC could amplify the residue after the coarse conversion and subsequently perform the fine conversion. Implemented in a 90nm CMOS process, the coarse-fine TDC in [58] had 1.25ps resolution, 9 bits range, occupied 0.6mm^2 area and consumed 3mW power. Because the time residue amplification was used, this TDC suffered from gain uncertainty due to PVT variations. Similar to the cyclic ADC, a 8-bit cyclic time-to-digital conversion was achieved by repetitively using a 1.5-bit time-domain DAC [91]. When implemented in a 0.13μm CMOS process, this TDC had a resolution of 1.25ps, and a measurement range of 320ps. It occupied 0.07mm^2 and the power consumption was 4.3mW.

(c) FPGA-based TDCs

Instead of a CMOS ASIC (application-specific integrated circuit), FPGA is an alternative implementation approach for fully-digital TDC architectures. Several groups [92]–[97] have published excellent demonstrations of FPGA-based TDCs. FPGA's attractive benefits includes significantly shorter development cycle and much cheaper development cost than the ASIC approach. It also offers programmable flexibility. Moreover, since FPGA devices are designed for parallelism, they are good candidates for multi-channel TDCs [92]–[94], [98]–[100].

At present, most FPGA-based TDCs use the carry chain to form the delay-line. The tapped delay line (TDL) architecture [92], [95], [100], [101], the VDL technique [102]–[104], the pulse shrinking method [97], [99], the wave union method [105]–[107], and the multiple interpolation approach [98], [108], [109] were successfully implemented and demonstrated in FPGA devices. Usually, the time resolution of the FPGA-based TDCs is determined by the propagation delay of the delay unit (in the TDL case), the difference of two delay chains (in the Vernier method), or the fine interpolator (in the multiple interpolation method), just the same as those architectures designed in the CMOS ASIC approach.

Technical improvements in FPGAs will result in finer time resolution and less standard uncertainty in FPGA-based TDCs, because the propagation delay of the delay units in FPGA will decrease. For example, the time resolutions of FPGA-based TDCs implemented in a Virtex-2 device (Xilinx, 0.15μm process) [98], a Virtex-4 device (Xilinx, 90nm process) [92], a Virtex-5 device (Xilinx, 65nm process) [101], and Virtex-6 devices (Xilinx, 40nm process) [95], [100] are 65.3ps, 50ps, 17ps, 10ps and 10ps, respectively. The time interval measurement precisions of TDCs based on the Virtex-4 [92], Virtex-5 [101], and Virtex-6 [95], [100] devices are 25ps, 24.2ps, 19.6ps and 12.83ps, respectively.

1.3. Research Motivations

PET is a powerful nuclear medical imaging tool which offers high sensitivity at the molecular level. ToF PET, as an advanced implementation to the conventional PET, can further boost the image quality by increase of the SNR and contrast of the reconstructed image. For this application, high sensitivity, compact size and low cost single-photon detectors are needed. And to implement the ToF PET imaging, advanced circuits that can measure time intervals with high resolution, sufficient measurement range, low power, and small size are required.

In traditional PET imaging systems, PMTs are widely used due to their excellent timing response, high sensitivity and good SNR. However, they suffer from several limitations, such as their bulky size, high cost, requiring over 1000V operating voltage, and fragile glass package. Besides, PMTs are susceptible to magnetic field. This feature holds it back from being used in multimodality medical imaging, e.g. the PET/MRI.

As a solid-state photodetector implemented in CMOS process, SPAD is a promising alternative to PMT. Working above its breakdown voltage, SPAD can offer single-photon level sensitivity. It requires <20V operating voltage. Since it is fully compatible with standard CMOS processes, it is easy to obtain features like compact size, low power consumption, integration with functional electronics, and remarkably reduced cost. Moreover, due to its intrinsic digital property, SPAD is immune to magnetic field, which makes it a perfect choice of detectors in PET/MR imaging.

Targeting the miniaturized size and cost reduction of the photodetectors, this research focuses on the design and characterization of SPAD in a standard mainstream digital CMOS process. Taking into consideration the fact that this technology is not optimized for optical detection, research efforts are devoted to analysis the dark noise sources in SPAD fabricated in a standard digital CMOS process. A versatile and comprehensive analytical model was studied in this work. Both simulations and measurements were performed and compared.

To fully explore the potential of SPADs in a standard digital CMOS process, the silicide layer on the active area of the SPAD was studied. Two test structures focusing on the silicide layer effects were designed, fabricated and characterized. The impacts of silicide layer on the DCR, PDE, after-pulsing, breakdown voltage of the SPAD were fully studied and analyzed. These efforts led to great improvement in terms of both noise reduction and detection efficiency increase of the SPAD.

Another main goal of this study is to design and implement a high performance TDC in the same technology as the SPAD. This TDC is targeting ToF PET imaging applications. For this purpose, a TDC prototype chip was developed to meet specifications such as large measurement range to cover the whole field-of-view in a PET system, high resolution and low jitter to increase the image's SNR and contrast, and compact size to maximize the fill-factor of photodetectors. With the TDC architecture proposed in this work, a small nonlinearity of its transfer characteristic was achieved, resulting in good measurement accuracy over the entire measurement range. For PET imaging applications, it should have the potential that it can be built in an array and be integrated with SPADs on the same silicon die. While keeping this requirement in mind, the TDC featured state-of-the-art low power consumption and compact footprint. It could be easily extended to a large matrix design and be shared by array of SPADs in future work. It also works in asynchronous mode, thus eliminating the need for a global reference clock.

A customized printed circuit board (PCB) was developed to integrate the SPAD and TDC to build a ToF measurement prototype. The results will be presented in this thesis, and compared with reference setups using commercial components. The prototype is suitable to build a miniaturized, multiple channel and low cost sensing system for ToF PET imaging application. High measurement accuracy and precision are expected.

1.4. Research Contributions

The research work conducted in this thesis aims at designing and building a compact, high performance and low cost solid-state sensing system for PET imaging based on CMOS technology. The major contributions of this work are summarized as follows.

- A set of SPAD was designed in a mainstream standard digital CMOS process (IBM’s 130nm). The devices have been characterized and analyzed.
- The impact of silicide layer on the performance of SPADs was studied. For this purpose, two test structures were designed and fabricated in the 130nm CMOS process. Their breakdown voltage, DCR, after-pulsing probability, PDE and shallow level traps were characterized and compared. The role of silicide layer in SPAD design was explored and highlighted.
- A novel analytical SPAD model was proposed and implemented in Verilog-A hardware description language (HDL). This model included the static behavior, dynamic behavior and statistical behavior. All noise sources, such as carrier diffusion, thermal generation, band-to-band tunneling and after-pulsing mechanisms, were considered in this model. Extensive measurements were performed to accurately extract the physical parameters that were required in the model. Demonstration and validation of the model were accomplished with SPADs at different temperatures, biasing voltages and active area sizes.
- A compact, high resolution and low power TDC was developed based on Vernier ring oscillator technique. Critical circuit blocks, such as the gateable ring oscillator and end-of-conversion detection array, were designed and extensively simulated to meet the performance expectation under all process corners. Careful layout of the design and post-layout simulation were accomplished. Full characterization of fabricated TDC chips were done. The TDC achieved the start-of-the-art low power consumption due to its gating operating feature.
- A prototype sensing system was designed to integrate the previously developed SPADs and TDC on a single PCB. Application demonstration of

ToF measurement was performed. The results showed promising coincidence timing resolution and good accuracy of this prototype.

Below, a list of the journal papers and conference presentations that resulted during the period of work conducted in this thesis is given.

Published Journal Papers

1. Z. Cheng, M. J. Deen, and H. Peng, "A low-power gateable Vernier ring oscillator time-to-digital converter for biomedical imaging applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 2, pp 445-454, 2016.
2. Z. Cheng, X. Zheng, M. J. Deen and H. Peng, "Recent developments and design challenges of high-performance ring oscillator CMOS time-to-digital converters," IEEE Transactions on Electron Devices, vol. 63, no. 1, pp 235-251, 2016.
3. Z. Cheng, X. Zheng, D. Palubiak, M. J. Deen and H. Peng, "A comprehensive and accurate analytical SPAD model for circuit simulation," IEEE Transactions on Electron Devices, vol. 63, no. 5, pp 1940-1948, 2016.
4. Z. Cheng, D. Palubiak, X. Zheng, M. J. Deen and H. Peng, "Impact of silicide layer on single photon avalanche diodes in a 130nm CMOS process,", *Journal of Physics D: Applied Physics*, vol. 49, no. 34, pp. #345105-1-11, 2016.

Conference Presentations

1. Z. Cheng, X. Zheng, M. J. Deen, H. Peng, and L. Xing, "Noise modeling of single photon avalanche diode (SPAD) for photon counting CT applications," oral presentation in *American Association of Physicists in Medicine (AAPM) Annual Meeting*, Washington, DC, July 31- Aug.4 , 2016.
2. X. Zheng, Z. Cheng, M. J. Deen, H. Peng and L. Xing, "Impact of charge sharing effect on sub-pitch resolution for CZT-based photon counting CT systems," oral presentation in *American Association of Physicists in Medicine (AAPM) Annual Meeting*, Washington, DC, July 31- Aug.4 , 2016.

3. Z. Cheng, X. Zheng, M. J. Deen and H. Peng, "Development of a time-to-digital converter and 8×8 single photon avalanche photodiode array towards the digital SiPM sensor," poster presentation in *IEEE Nuclear Science Symposium & Medical Imaging Conference (NSS/MIC),* San Diego, CA, Dec. 31-Nov.7, 2015.
4. X. Zheng, Z. Cheng, M. J. Deen and H. Peng, "Sub-pitch spatial resolution in CZT detectors: simulation study," oral presentation in *IEEE Nuclear Science Symposium & Medical Imaging Conference(NSS/MIC),* San Diego, CA, Dec. 31-Nov.7, 2015.
5. Z. Cheng, X. Zheng, M. J. Deen and H. Peng, "Development of a high performance digital silicon photomultiplier (dSiPM) for ToF PET Imaging," oral presentation in *Society of Nuclear Medicine and Molecular Imaging Annual Meeting (SNMMI),* Baltimore, MD, Jun. 6-10, 2015.
6. X. Zheng, Z. Cheng, M. J. Deen and H. Peng, "Investigation of the sub-pixel spatial resolution and charge-sharing effect in CZT detectors for PET imaging," poster presentation in *Society of Nuclear Medicine and Molecular Imaging Annual Meeting (SNMMI),* Baltimore, MD, Jun. 6-10, 2015.
7. Z. Cheng, H. Peng and M. J. Deen, "High Performance Integrated Circuits for Biomedical Imaging Applications," oral presentation in *225th Meeting of the Electrochemical Society (ECS),* Orlando, FL, May 11-15, 2014.

News Article

1. M. J. Deen, Z. Cheng, and D. Palubiak, "High-performance SPADs and advanced digital circuits for improved medical imaging," *SPIE Newsroom,* 2016.

1.5. Thesis Organizations

This thesis is organized as follows. In Chapter 1, an introduction of different nuclear medical imaging modalities including PET, ToF PET and PET/MR imaging, are presented. The advantages and disadvantages of current photodetector technologies and TDC implementations are discussed. Then, the motivation of focusing on high sensitivity silicon-based photodetectors and advanced timing circuits are provided.

Finally, a brief summary of the main contributions of this research and the structure of this thesis was given.

In Chapter 2, an analytical model of the SPAD using Verilog-A HDL is presented. Detailed modeling on the static, dynamic and statistical behaviors is given. In particular, efforts are devoted to the noise modeling in SPAD. Both primary dark noises and secondary noises are considered in this chapter. Since many physical parameters are involved in this model, then a dedicated section regarding the method to extract the parameters from measurements are presented. Then, the simulation results at different temperatures and excess voltages are presented and compared with measurement results.

In Chapter 3, since silicide layer is widely used to achieve a reliable and reduced resistive source/drain contact in standard CMOS processes, its impact on SPAD is studied. For this purpose, two SPAD test structures are developed and fabricated in a 130nm CMOS. Then, characterization of SPAD performance, including the breakdown voltage, dark noise, after-pulsing and detection efficiency, is performed. Detailed analysis and discussions are conducted in order to reveal the reasons that cause these changes when the silicide layer above the active area of a SPAD is removed. Light transmittance calculation based on transfer matrix is also presented.

In Chapter 4, a TDC design featuring high resolution, low jitter, large dynamic range, good linearity is presented. The prototype chip is fabricated in IBM’s 130nm CMOS process. Full details of the circuit design, simulation and performance characterization are presented.

In Chapter 5, in order to demonstrate the proposed SPAD and TDC design towards PET imaging application, a custom PCB is designed to host both the SPAD and TDC. A setup that can mimic the PET imaging is built. Time-of-flight measurements are performed using this system. Results, data processing and their benefits when applied to PET imaging system are also analyzed.

In Chapter 6, this thesis is concluded with a summary of the research and several recommendations for future improvements for both SPAD and TDC designs, and the on-chip system integration.

Chapter 2

ANALYTICAL MODELING OF SINGLE PHOTON AVALANCHE DIODE[†]

SPAD is an attractive photodetector due to its high sensitivity and low dead time. Since a SPAD is usually coupled with its associated front-end circuits, an accurate circuit simulation model will help the designer to optimize the circuit performance. Especially, a HDL based model will facilitate the digital circuits design using HDL code as well, which is synthesizable with electronic design automation (EDA) tools.

In this chapter, a comprehensive, accurate analytical SPAD circuit simulation model will be introduced and implemented in Verilog-A HDL. The modeling process of all the SPAD behaviors will be discussed in detail. Then, the parameter extraction processes from inter-avalanche time measurement of a free-running SPAD device will be discussed. At last, the simulation results will be validated with measurement results to verify the accuracy of the proposed model.

2.1. Background of SPAD Modeling

Due to its low dead time and high sensitivity in the 400-900nm spectral range, silicon-based SPADs have become an attractive solid-state detector technology for biomedical applications. Examples of these applications include FLIM [26], Raman spectroscopy [110] and ToF-PET imaging [25]. For single photon detection, the SPAD is biased above its breakdown voltage, operating in what is termed the Geiger mode. In such a mode, an avalanche current results when a photon strikes the active region of

[†] Part of this work was published as: Z. Cheng, X. Zheng, D. Palubiak, M. J. Deen and H. Peng, “A comprehensive and accurate analytical SPAD model for circuit simulation,” IEEE Transactions on Electron Devices, vol. 63, no. 5, pp 1940-1948, 2016. Copyright granted by IEEE.

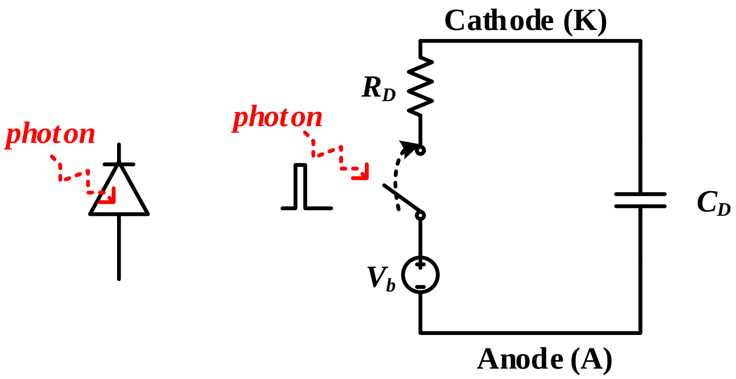


Figure 2-1, Basic model of a SPAD.

the junction. Subsequently, an associated front-end circuitry is needed to protect the diode from thermal burn out and to restore the SPAD bias to Geiger mode to detect the next incoming photon. This front-end circuitry is usually referred as the quenching/reset circuit.

Since an individual SPAD or an SPAD array is usually monolithically integrated with complex electronics on the same silicon substrate [15], [16], [45], [111], [112], then an accurate simulation is highly desired. This simulation should accurately predict the static, dynamic and statistical behaviors in a circuit design environment. For example, SPADs are often cooled because cooling results in reduced thermal noise. However, with cooling, band-to-band tunneling noise and after-pulsing noise will become dominant, because the former shows less temperature dependence compared to thermal noise, and the latter increases since the trap's lifetime is longer at lower temperatures. Therefore, a comprehensive and accurate circuit simulation model of the SPAD will facilitate the analysis and impact of noise on the performance of entire photon sensing system.

A simple and commonly used model of the SPAD [113], [114] is shown in Figure 2-1. It consists of a DC voltage source (V_b), a series resistor (R_D) and a capacitor (C_D), representing the diode breakdown voltage, diode internal resistance and junction capacitance, respectively. A voltage-controlled switch, which can also be replaced by a NMOS transistor [67], [115], is used to imitate the photon that strikes the active region of the SPAD. By closing the switch, an avalanche event can be triggered.

An improved model was proposed in [116], [117]. It introduced two selfsustaining/self-quenching switches to simulate the positive feedback and self-quench mechanisms when an avalanche was triggered in a SPAD. To better model the current-voltage (I-V) characteristics of SPAD, the work in [116], [117] also presented a piecewise linear curve approach using a nonlinear voltage source, but it has convergence problems. The quenching/recovery time was emulated by taking the parasitic components (i.e. the stray capacitors from the cathode and anode to the substrate) into account. Since it only employs SPICE and the standard library cells available in commercial circuit simulators, this model is very attractive to designers.

It is noted that any piecewise linear function has points where its gradient is discontinuous, and in these points, it is not differentiable. Hence, the simulator solver may not be able to reach the mathematical solution [118]. The convergence problem, which existed in the piecewise linear function, was solved in [118] when the SPAD model was implemented using Verilog-A HDL and Cadence Spectre simulator. An additional resistor was introduced to provide a transitional slope, thus to avoid the convergence problem. Moreover, the voltage-dependence of the diode junction capacitance was considered. Thus, the simulation accuracy was enhanced. However, important statistical behaviors, such as dark count noise and after-pulsing counts were not included in that model [118].

The works in [115] and [119] both included the thermal generation dark noise and after-pulsing dark noise. However, as an important contributor to dark noise, the band-to-band tunneling mechanism was not considered. At the same time, the after-pulsing probability is greatly dependent on the time interval relative to the previous avalanche event [120]–[123]. This temporal dependence of after-pulsing probability was not considered in [119].

Building on previous SPAD modeling works [115]–[119], a comprehensive and accurate SPAD model is implemented with Verilog-A HDL. This model is fully compatible with mainstream commercial circuit simulators, such as Cadence Spectre, Synopsys HSpice and OrCAD PSpice. The proposed circuit simulation model is not only emulating the static and dynamic behaviors of SPAD, but it also includes important noise behaviors, that is the primary dark noise and the secondary dark noise

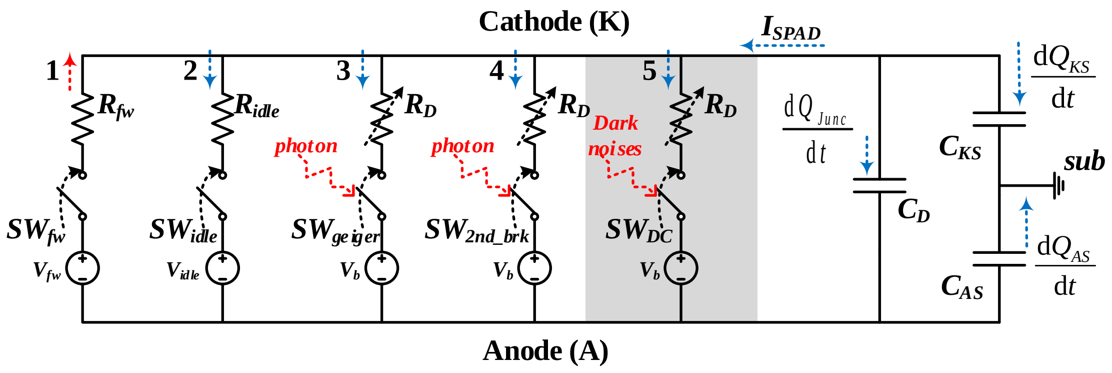


Figure 2-2, Proposed analytical model of SPAD. The red and blue dashed arrows on each path indicate the current flowing through the SPAD. (Note: the path in shadow is specific for the dark noise due to thermal generation, band-to-band tunneling and after-pulsing noises)

(after-pulsing). The simulation results are validated against measurement results to demonstrate the accuracy of the proposed model.

This work in this chapter is significant for several reasons. First, to the best of our knowledge, this is the first SPAD circuit simulation model that considers the band-to-band tunneling noise mechanism and the temporal dependence of after-pulsing probability. Second, the approaches to extract physical modeling parameters are discussed. Third, because this model is implemented in Verilog-A HDL code, it does not rely on any specific technology. This feature brings great application universality for the work. Thus, it can be easily used by other researchers to help in optimizing front-end quenching/reset electronic circuits.

The rest of this chapter is organized as follows. The analytical modeling of SPAD, including the static, dynamic and statistical (thermal generation, band-to-band tunneling and after-pulsing noises) behaviors, is presented in Section 2.2. In Section 2.3, we describe how to extract those physical parameters used in our proposed model. In section 2.4, we give the simulation results and experimental validations. Finally, the conclusions are provided in Section 2.5.

2.2. SPAD Analytical Modeling

Figure 2-2 shows the proposed analytical SPAD model. According to its bias voltage, five different paths are used to present the avalanche process in a SPAD. In each path, a voltage source, a switch and a resistor are used, as parts of the basic SPAD model in

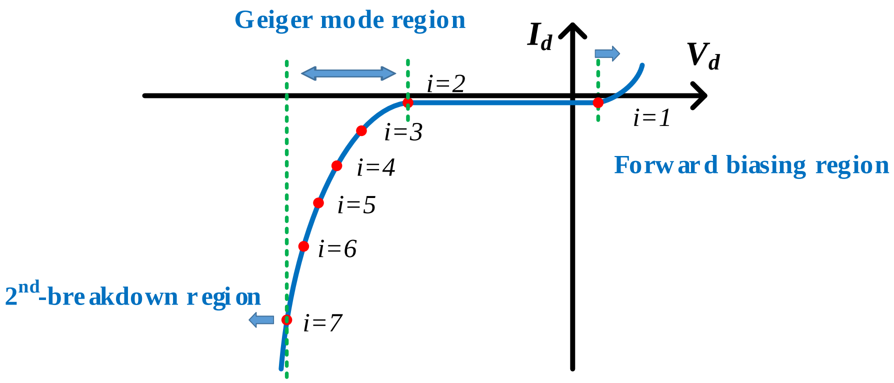


Figure 2-3, Illustration example of the typical SPAD I-V curve.

Figure 2-1. The first path emulates the forward biasing condition, when the anode voltage is higher than the cathode voltage by the junction turn-on voltage. The second path is the reversed biasing condition when the excess voltage is less than the breakdown voltage of the SPAD. The third to fifth paths are used to simulate the cases when SPAD is in Geiger mode. More specifically, the third path emulates the breakdown operation that are triggered by incident photons. The fourth path is for the second breakdown due to edge-junction or punch-through effects [117]. And the fifth path is for avalanches that are initiated by dark noises (e.g. thermal generation, band-to-band tunneling and after-pulsing).

2.2.1. Static Behavior Modeling

The I-V characteristic of a SPAD device in Geiger mode can be well represented by a piecewise linear curve while avoiding any convergence issue, as in [115], [118], [119]. Along the measured I-V curve, several points ($I_{d,i}$,$V_{d,i}$) are chosen to reconstruct the static characteristic of the SPAD. Here, i is the index number of selected points. Figure 2-3 gives an illustration curve and the selected points are highlighted (i=1, 2 … 7 in this example). Several key points should be included.

1): When the SPAD is forward biased,

2): When the SPAD enters Geiger mode,

3): When the SPAD is under second breakdown region.

Also, when more points are chosen, the I-V characteristics are more accurately reproduced [117].

To accomplish the piecewise linear modeling of I-V curve in the breakdown region, a nonlinear SPAD resistance is used, which can be written as,

$$R_{brk,i} = \left(V_{d,i+1} - V_{d,i}\right) \big/ \left(I_{d,i+1} - I_{d,i}\right),\ i = 2,3,\ldots 6. \tag{2-1}$$

Then, the current that flows through the diode during the avalanche process can be calculated by equation (2-2), with a fully differentiable pseudo-max function to avoid the convergence problems in [115], [118], [119],

$$I_{SPAD} = I_{d,i} + \frac{V_n}{R_{brk,i}} \ln(1 + e^{\frac{V_d - V_{d,i}}{V_n}}),\ i = 2,3,\ldots 6, \tag{2-2}$$

In (2-2), V_n is a normalization voltage of about 10mV [118], [119].

Since our model mainly concerns the behavior of the SPAD in the breakdown region, we can simply use straight lines to emulate the I-V characteristics in both the forward biasing and the second breakdown regions.

2.2.2. Dynamic Behavior Modeling

In order to emulate the dynamic behavior of the SPAD, we consider three capacitors (shown in Figure 2-2), and their storage charges are the same as in [115], [117]–[119]. The charges being stored in the depletion region (Q_{Junc}), the cathode-to-substrate (Q_{KS}) and the anode-to-substrate (Q_{AS}) stray capacitors are given as,

$$\begin{aligned} Q_{Junc} &= A_D \frac{\Phi_i C_{j0}}{1 - m_j} (1 + \frac{V_d}{\Phi_i})^{1-m_j}, \\ Q_{KS} &= C_{KS} V_K, \\ Q_{AS} &= C_{AS} V_A, \end{aligned} \tag{2-3}$$

where A_D is the area of the active region of the SPAD, Φ_i is the built-in voltage, C_{j0} is the zero-voltage junction capacitance per unit area, m_j is the junction grading coefficient, C_{KS} and C_{AS} are the cathode-to-substrate and anode-to-substrate stray capacitors, respectively. V_K and V_A are the voltages of cathode and anode with respect to the substrate. Here, the C_{KS} and C_{AS} are considered as constant values to simplify the modeling. However, the values of these two stray capacitors can be precisely extracted from measurement, using the approach proposed in [117].

2.2.3. Statistical Behavior Modeling

Dark counts are the false-triggered avalanche events due to carriers in the depletion region when the SPAD is biased above its breakdown voltage and in the dark. The dark counts can be categorized as primary and secondary dark counts. In our SPAD model, the causes of dark counts are considered. The causes of primary dark counts include carrier diffusion from the neutral region, carrier thermal generation in the depletion region, and band-to-band tunneling. The cause of secondary dark counts is the after-pulsing phenomenon.

For all the primary dark noise sources, we first calculate the carrier generation rate. Then, two successive carrier generation events can be scheduled with a time interval. The average value of this time interval is the reciprocal of the carrier generation rate and it follows an exponential distribution.

(a) Primary Dark Counts due to Carrier Diffusion

When the diode is biased in Geiger mode, minority carriers will diffuse from the neutral regions to the depletion region edges and have certain probability to initiate an avalanche event. The carrier generation rate due to this diffusion process can be approximated by the diffusion equation [124], [125]. According to the results presented in [124], the carrier diffusion process contributes the least to dark noise among all the sources, about 2-3 orders of magnitude smaller than the others. Thus, we can neglect this noise source to simplify the modeling presented here.

(b) Primary Dark Counts due to Carrier Thermal Generation

Usually, the Shockley-Read-Hall theory is used to determine the rate of thermally generated carriers. With the electron/hole concentration much less than n_i (i.e. $n \ll n_i$, $p \ll n_i$) in the depletion region, and assuming the carrier capture cross sections for hole's and electron's are equal, the carrier thermal generation rate ($CGR_{thermal}$) [119], [124], [125] can be simply given by,

$$\begin{aligned} CGR_{thermal} &= \frac{(n_i^{\,2} - pn)A_D W_D}{\tau_e\,(p + n_i e^{\frac{-(E_t - E_i)}{kT}}) + \tau_h\,(n + n_i e^{\frac{-(E_i - E_t)}{kT}})} \\ &\approx \frac{n_i}{\tau_e e^{\frac{-(E_t - E_i)}{kT}} + \tau_h e^{\frac{-(E_i - E_t)}{kT}}} A_D W_D \\ &\approx \frac{n_i v_{th} \sigma_0 N_t}{2} A_D W_D, \end{aligned} \tag{2-4}$$

where N_t is the density of generation/recombination centers, σ_0 is the carrier capture cross section, A_D is the diode's active area, W_D is the effective thickness of the depletion layer, and v_{th} is the electron thermal velocity given by the follow equation,

$$v_{th} = \sqrt{\frac{3kT}{m^*}}, \tag{2-5}$$

in which k is Boltzmann's constant and m^* is the electron effective mass.

(c) Primary Dark Counts due to Band-to-band Tunneling

With the scaled CMOS technology being employed to fabricate the SPAD, the depletion layer becomes thin (around 1μm) and electrical field strength increases accordingly. Carriers generated via the band-to-band tunneling process may also trigger the avalanche events with a certain probability. Hence, the band-to-band tunneling contribution to the total DCR should be considered. For electrical fields approaching 10^6 V/cm in silicon, the effect of band-to-band tunneling on DCR becomes significant [125].

However, in previously reported SPAD simulation models [115], [117]–[119], this mechanism was not considered. In this chapter, the band-to-band tunneling process is included in the circuit simulation modeling for the first time. In particular, its dependences on reverse bias voltage and temperature make it feasible to estimate the DCR from circuit simulation. More importantly, since the tunneling mechanism is less temperature dependent than the thermal generation process, then band-to-band tunneling related dark counts will dominate at lower temperatures (typically 0°C), which is generally the case in real applications where SPADs are cooled to suppress the thermally generated dark noise. Thus, this model will help to determine the proper cooling temperature for applications of SPADs.

The carrier generation rate due to band-to-band tunneling process ($CGR_{tunneling}$) [124]–[131] is given by,

$$CGR_{tunneling} = \frac{\sqrt{2m^*}q^2FV_R}{h^2\sqrt{E_g}}\exp(-\frac{8\pi\sqrt{2m^*}E_g^{\ 3/2}}{3qFh})A_D, \tag{2-6}$$

where h is Plank's constant, E_g is silicon band gap energy, V_R is the reverse bias voltage and F is the average electric field in the depletion region.

The avalanche triggering probability (P_{tr}) depends on both the SPAD structure (i.e. its breakdown voltage V_{brk}) and the excess bias voltage (V_{EX}). It can be numerically calculated according to the electric field profile and the ionization coefficients for electrons and holes, using the method described in [114], [124], [132]. In our model, the triggering probability under reverse biasing condition is approximated by the following equation [115], [119], [133]–[135].

$$P_{tr} = 1 - e^{-\frac{V_{EX}}{\eta_T V_{brk}}}, \tag{2-7}$$

where η_T is an empirical parameter which sets the exponential slope. A η_T value of 0.131 [115], [119] is used in our model.

Based on the discussions above, the primary DCR of the SPAD due to carrier diffusion, thermal generation and band-to-band tunneling mechanisms can be obtained as follows,

$$\begin{aligned} DCR_{primary} &= CGR_{total}P_{tr} \\ &\approx (CGR_{thermal} + CGR_{tunneling})P_{tr}. \end{aligned} \tag{2-8}$$

Each carrier generation process has a Poisson distribution while its average value is CGR_x, with x being either *thermal* or *tunneling*. In our model, the time interval between two successive dark carriers generation is considered as an exponential distribution, whose expected value is τ_x=1/CGR_x. The function *$dist_exponential*(*seed*, *mean*) available in Verilog-A HDL is used to return a pseudo-random number for τ_x. One example of the Verilog-A HDL codes for the tunneling dark counts modeling is given in Figure 2-4.

```
@(timer(t_cg_tunneling)) //dark counts due to tunneling;
    ... //update the carrier-generation counters;
    if ( aval==0 && Vex() >= 0 ) //Vex() gives the excess bias voltage;
    begin
        if( Ptr() > $rdist_uniform(seed,0,1) ) //Ptr() gives the triggering probability;
        begin
            aval=1; //trigger the avalanche;
            ...   //update the tunneling dark count counter;
            //schedule next carrier-release;
            t_cr_1=$abstime + $rdist_exponential(seed_cr_1,tau1);
            ...
            AvalTime=$abstime; //update the time for this avalanche event;
        end
    //schedule the next carrier-generation event;
    t_cg_tunneling=$abstime
                        + rdist_exponential(seed_cg_tunneling,tau_tunneling);
end
```

Figure 2-4, Verilog-A HDL code for the band-to-band tunneling process.

(d) Secondary Dark Counts: After-pulsing Phenomenon

When an avalanche occurs, a macroscopic current flows through the SPAD until it gets quenched via external quenching circuits in either passive or active approaches. During the avalanche process, some carriers may be captured by shallow-level traps, which are defects in the semiconductor lattice causing energy levels inside the forbidden band gap [124], [125]. These trapped carriers will be released after a statistical delay time. This delay is the lifetime of the shallow-level trap. If the trapped carrier is released when the SPAD is biased above its breakdown voltage and is ready to detect the next photon, a false, undesired avalanche will then be triggered. This is the after-pulsing phenomenon of an SPAD, and it contributes to the secondary dark counts.

The shallow-level trap’s lifetime decreases with increasing temperature [119], [124], [136], [137] according to the following equation,

$$\tau = \tau_0 e^{\frac{E_A}{kT}}, \tag{2-9}$$

where E_A is the activation energy which is defined as the energy difference between the trap’s energy level and the bottom of conduction band for an electron trap (or the top of the valence band for a hole trap), and τ_0 is the pre-exponential factor.

The lifetime can vary from tens of nanoseconds to several microseconds. Generally, there is more than one shallow-level trap in an SPAD, and they can be modeled by different traps' lifetimes. It has been reported that four exponentials best fit the experimental measurement data of trap emission by the time-correlated carrier counting technique [133], [136]–[139],[120]. One study [121] demonstrated that power-law temporal dependence could also obtain accurate fitting results, though we did not use the power-law fitting approach in this work.

The probability of after-pulsing is also time dependent. If the time interval between the avalanche and its associated trapped carrier release is small, this release will have a high probability to initialize an after-pulsing event, and vice versa. Thus, this time dependent probability [120]–[123] can be expressed by the following equation,

$$P_{ap}(t) = \sum_{i=1}^{N} A_i \frac{1}{\tau_i} e^{-\frac{t}{\tau_i}}, \tag{2-10}$$

where the index i presents the i-th trap, N is the total number of deep-level traps, A_i and τ_i are the exponential pre-factor and the lifetime associated with the i-th trap, respectively. The time interval between the avalanche and its associated carrier release is t.

2.2.4. Other Considerations of Temperature Dependence

Since SPADs are often cooled for real applications to suppress the dark count noise, it is necessary to include the temperature dependence of some key parameters. In this way, our model will help in evaluating the cooling effect because noise due to band-to-band tunneling and after-pulsing contribute more than that due to thermal generation when the temperature is lowered by 10~20°C below room temperature, for example.

We can model the temperature dependences of the silicon intrinsic carrier concentration n_i, band-gap energy E_g and the SPAD breakdown voltage V_{brk} by the following three equations [125], [140]–[143],

$$n_i = \sqrt{N_C N_V} e^{\frac{-E_g}{2kT}}, \tag{2-11}$$

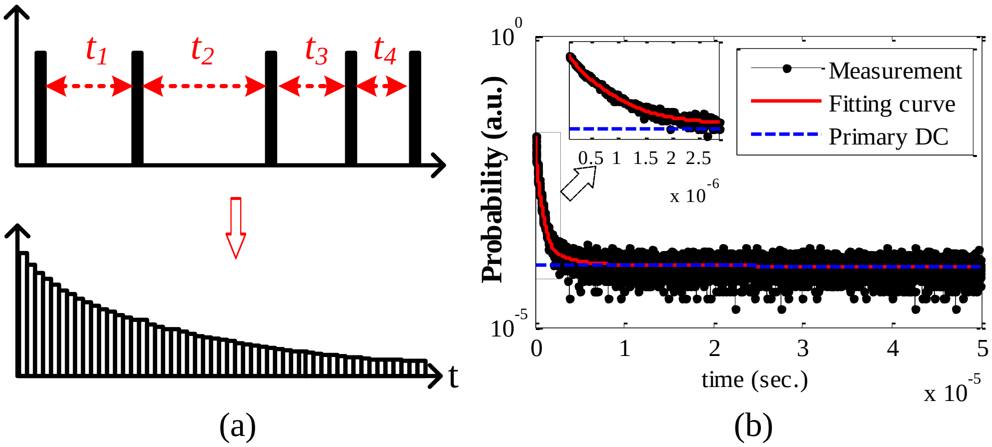


Figure 2-5, (a) The inter-avalanche time interval measurement of a free-running SPAD, and (b) Example of the statistical distribution obtained through the measurement at ~30°C with 0.5V excess voltage. The distribution is fitted with four exponentials.

$$E_g = E_{G0} + E_{G1}T + E_{G2}T^2 + E_{G3}T^3 + E_{G4}T^4, \tag{2-12}$$

$$V_{brk} = V_{brk0}[1 + \beta(T - T_0)], \tag{2-13}$$

where N_C and N_V are the effective densities of states in the conduction band and valence band, respectively. E_{G0} to E_{G4} are temperature coefficients of the band-gap energy. V_{brk0} is the breakdown voltage at room temperature T_0, and β is the linear temperature coefficient.

2.3. Parameter Extraction

In this model, we consider three shallow-level traps in the forbidden band gap. In this section, we describe the method used to obtain the activation energies, lifetimes, and exponential pre-factors that are associated with these three shallow-level traps.

One free-running SPAD [32], [67] is used to extract physical parameters. The inter-avalanche time interval is measured from -30 °C to 40 °C by 10 °C increments, at three different excess bias voltages. Figure 2-5(a) illustrates this measurement, in which the time intervals between two successive avalanche pulses are measured and recorded. After sufficient events are collected, a statistical distribution of the inter-avalanche time interval can be precisely obtained. One example at a temperature of -30 °C and an excess voltage of 0.5V, is shown in Figure 2-5(b).

Next, we fit the distribution with four exponentials. One exponential represents the

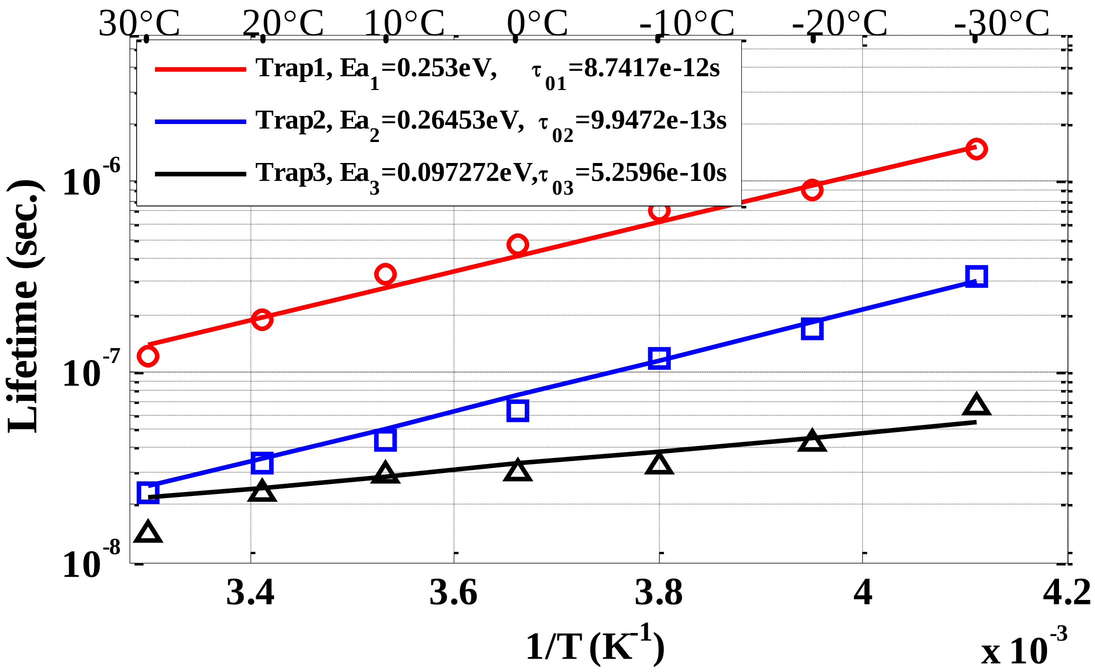


Figure 2-6, Arrhenius plot of the trap lifetimes to extract their activation energies.

primary dark counts that is contributed by thermal generation and band-to-band tunneling processes. The other three exponentials result from the after-pulsing phenomenon that are associated with the three shallow-level traps. Through the inter-avalanche time interval measurement and multiple exponential fittings, we can extract the traps' lifetimes and pre-exponential factors. We also note that when the temperature increases, the after-pulsing counts (i.e. the area between the red line and the dashed blue line in Figure 2-5 (b)) are reduced because the traps' lifetimes become shorter. This causes the fitting process at higher temperatures to be more difficult. The fitting goodness (R-Square) drops from ~99.7% at -30 °C to ~82% at 40 °C.

From equation (2-9), a trap's lifetime is temperature dependent. Hence, after the multiple exponential fittings at different temperatures, the Arrhenius plot [120], [136] is used to extract the τ_0 and E_A associated with the shallow-level traps. The results are shown in Figure 2-6.

In Table 2-1, a summary of parameters that are used in this model is provided. A_D is a design parameter. Since we do not have access to the device's technological information, the value of σ_0, N_t, F, W_D, N_C, N_V, C_{KS} and C_{AS} are estimated and are verified to be physically reasonable [116], [119], [124], [125], [140], [141]. The value

Table 2-1, Parameters that are used in SPAD analytical model

Symbol	Quantity	Value
V_{brk0}	Breakdown voltage at room temperature	11.42 V
β	Temperature coefficient of V_{brk}	6.325×10^{-4} °C^{-1}
A_D	The active area of the SPAD	82 μm^2
σ_0	Carrier capture cross section	4×10^{-19} m$^{2\,*}$
N_t	Generation/recombination center density	$8.7\sim10.7\times10^{17}$m$^{-3\,*}$
F	Average electrical field in the diode	$9.3\sim9.4\times10^{7}$ V/m*
N_C	Conduction band density of states	2.8×10^{19} cm^{-3}
N_V	Valence band density of states	1.04×10^{19} cm^{-3}
$E_{A,1}$	Activation energy of the 1st trap	0.253 eV
$E_{A,2}$	Activation energy of the 2nd trap	0.265 eV
$E_{A,3}$	Activation energy of the 3rd trap	0.097 eV
$\tau_{0,1}$	Pre-exponential factor of the 1st trap	8.7417×10^{-12} s
$\tau_{0,2}$	Pre-exponential factor of the 2nd trap	9.9472×10^{-13} s
$\tau_{0,3}$	Pre-exponential factor of the 3nd trap	5.2596×10^{-10} s
C_{KS}	SPAD cathode-substrate stray capacitor	2 pF
C_{AS}	SPAD anode-substrate stray capacitor	2 pF

* These values are estimated by fitting the SPAD primary dark counts measurement results with the modeling equation (8). N_t and F are different for different excess bias voltages. Here, the N_t and F are estimated and verified to be physically reasonable.

of $E_{A,i}$ and $\tau_{0,i}$ (i=1,2,3) are obtained through the multiple exponential fitting for the SPAD inter-arrival time measurement and the Arrhenius plot. V_{brk0} and β are calculated from I-V measurements.

2.4. Simulation and Experimental Validation

This modeling work is implemented in Verilog-A HDL, which is fully compatible with mainstream commercial circuit simulators. Our simulations have been performed in both Cadence Spectre and Synopsys HSpice as a demonstration of its universality. Figure 2-7 shows the state diagram for this modeling.

2.4.1. Simulation Setup

The SPAD model is operating in free-running mode, with a 50 kΩ passive quenching resistor. The incident photon is emulated as a narrow pulse (for example, 10ps in our simulation) by a pulsed voltage source. The simulation setup is shown in Figure 2-8(a). The anode of the SPAD is biased at -9V, and the cathode is connected to a DC voltage source (V_{CC}) through the quenching resistor. Figure 2-8(b) shows an example of the transient simulation at 40 °C with 1.5V excess voltage. At 20μs, a photon is detected. Then, at ~28μs, a dark count is generated, which is followed by an

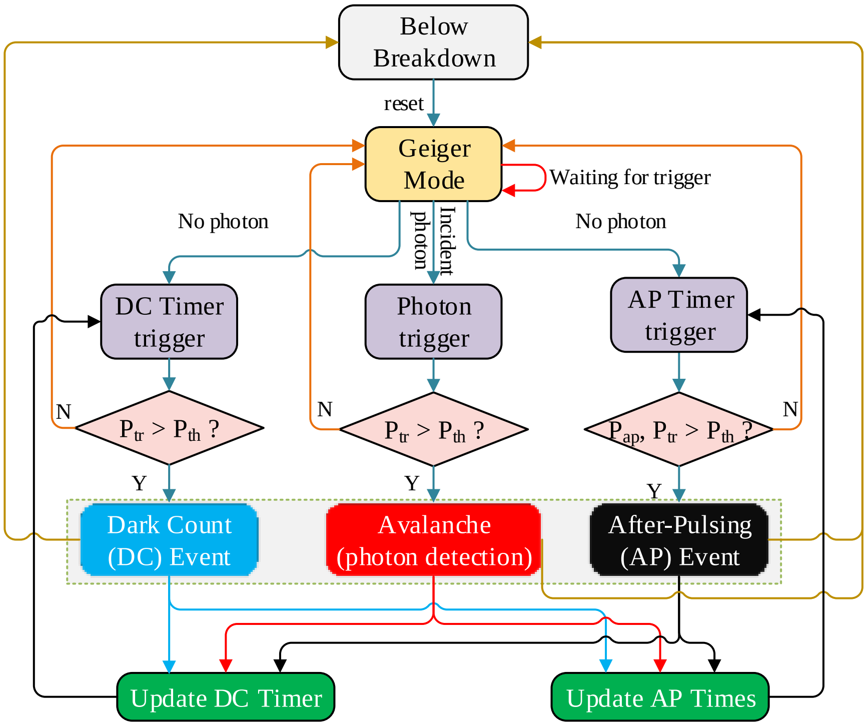


Figure 2-7, State diagram of the proposed analytical model.

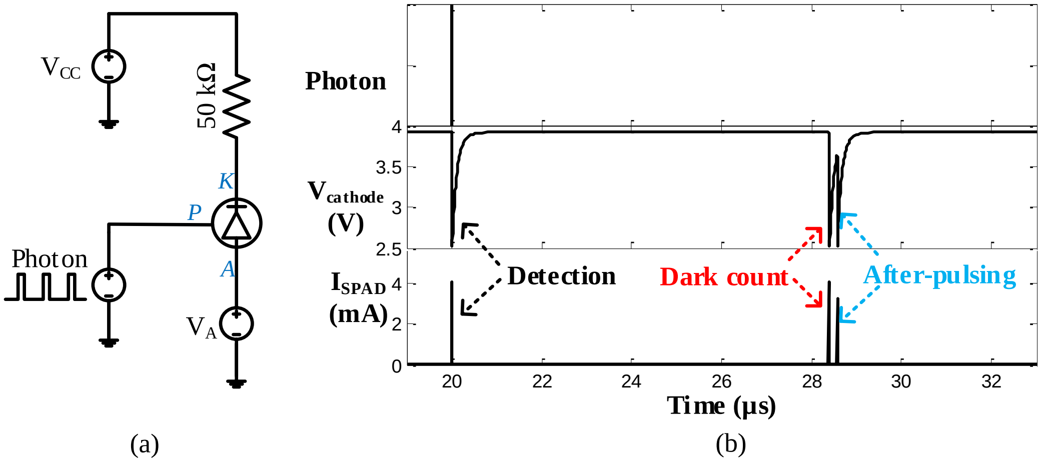


Figure 2-8, (a) Simulation Setup of the SPAD model, and (b) Example of transient simulation that is executed with 1.5V excess voltage at 40°C. This example shows a detection, a dark count and an after-pulsing events.

after-pulsing event.

To test the dark noise performance of the model, the SPAD is kept “in dark” during the simulation. This condition is fulfilled by connecting the “*P*” port of the SPAD model (see Figure 2-8(a)) to ground. Under this condition, the avalanche in the SPAD model can only be initiated by either primary dark counts or an after-pulsing event.

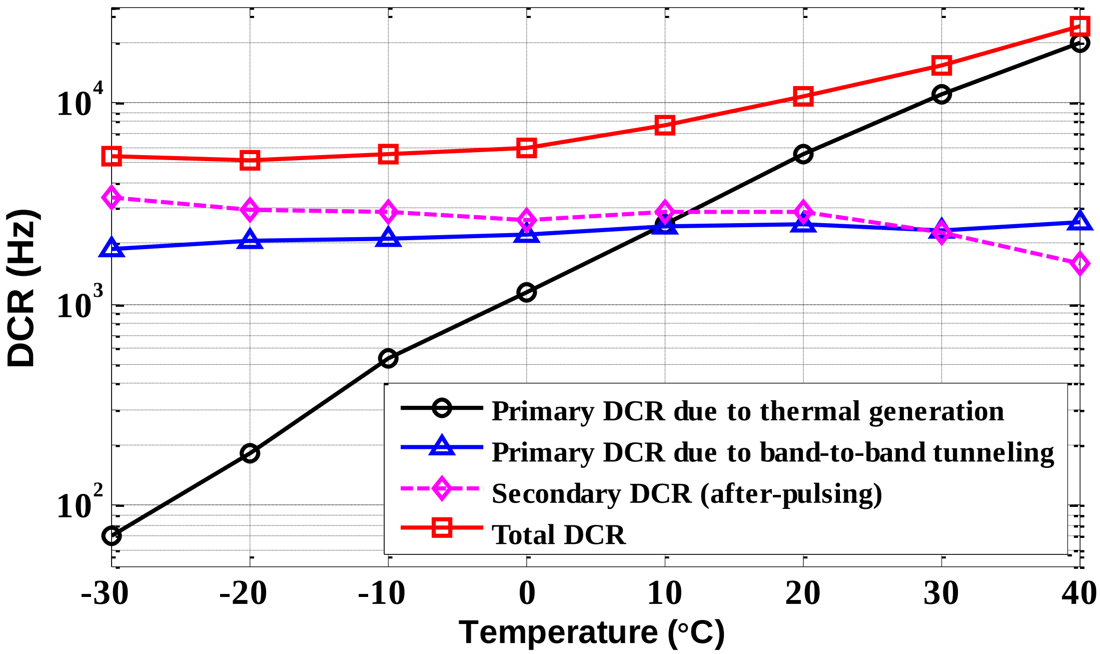


Figure 2-9, Simulation results of the temperature dependences of both the primary and secondary dark counts (V_{ex}=0.5V).

Each simulation is executed for 200ms. Then, it is repeated for three excess voltages (0.5V, 1.0V and 1.5V), and over a temperature range from -30 °C to 40 °C by 10 °C steps.

2.4.2. DCR Results

Figure 2-9 gives the results of the simulated dark counts as a function of temperature, when the excess voltage is set to 0.5V. The dark counts portion of band-to-band tunneling shows less temperature dependence than that of thermal generation. And the after-pulsing effect is reduced with temperature.

Note that at high temperatures (i.e. from 15 °C to 40 °C in Figure 2-9), the thermal generation will be the dominant source of the dark counts. The total DCR reduces when the temperature decreases, as seen from Figure 2-9. However, this benefit becomes less because of two reasons. First, for the primary dark count, the band-to-band tunneling mechanism begins to play a larger role than the thermal generation process at low temperatures. Second, the trap lifetimes become longer at lower temperatures, which slows down the release of trapped carriers. This means there will be more carriers to be released after the SPAD is quenched and reset to Geiger mode, which leads to higher

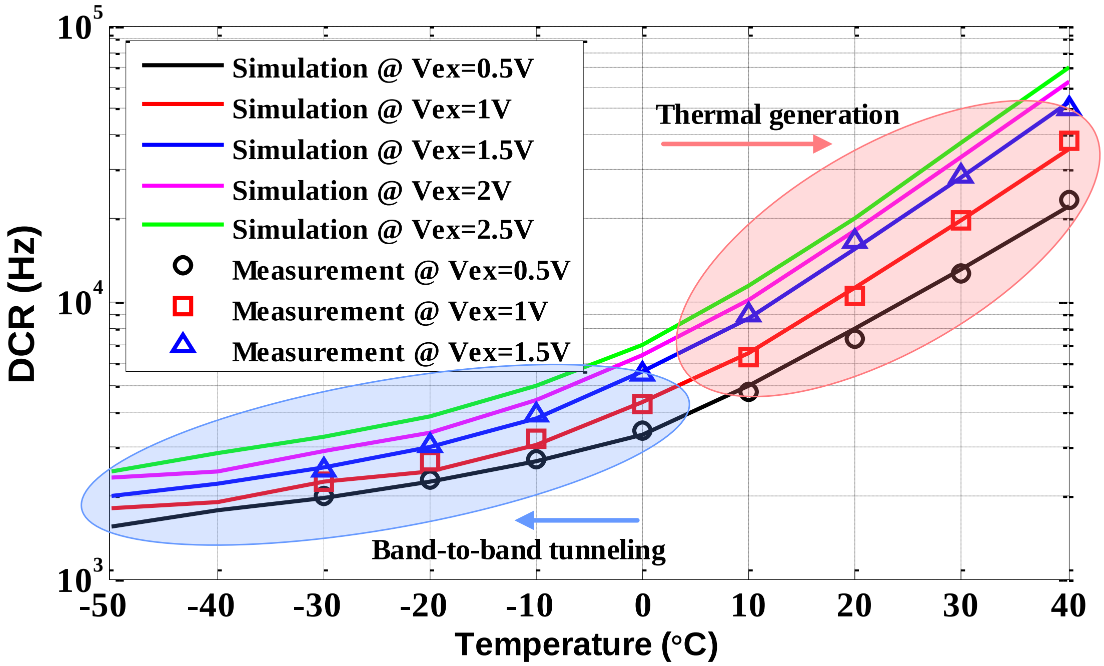


Figure 2-10, Simulation results of the primary DCR with temperatures at five different excess voltages.

after-pulsing dark counts. Therefore, a proper cooling temperature for the SPAD may be around 0 °C ~10 °C for this example, as indicated in Figure 2-9. For the other two excess voltage scenarios, the same trends are observed, while the absolute values of DCR increase with the excess voltages.

The variations of primary dark counts with temperatures, and primary dark counts with excess voltages are given in Figure 2-10 and Figure 2-11, respectively. The temperature ranges for which thermal generation dominates and band-to-band tunneling dominates are indicated in Figure 2-10. The primary dark counts increase with both temperature and excess voltage. By comparing the trends in Figure 2-10 and Figure 2-11, we can see that the DCR-temperature dependence is stronger than its excess voltage dependence. The measurement results of the primary DCRs of a SPAD (fabricated in IBM standard 130nm CMOS process), with temperature range of -30°C to +40°C and excess voltage range of 0.5V to 1.5V, are also included in both Figure 2-10 and Figure 2-11 to validate the accuracy of our proposed analytical SPAD model.

Note that the simulation and measurement results are in good agreement. The maximum relative error is 8.7%, which happens at 20 °C with 0.5V excess voltage. For this worst case, the primary DCRs from measurements and simulations are 7.35 kHz

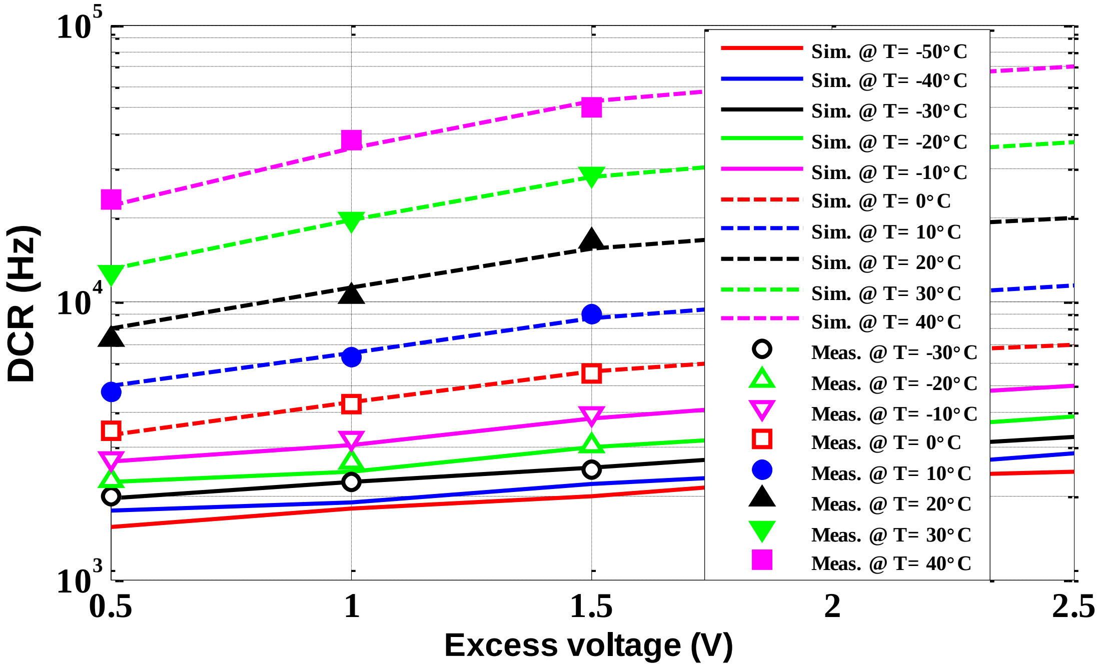


Figure 2-11, Simulation results of the total primary DCR with excess voltages at ten different temperatures.

and 7.99 kHz, respectively. The error between the simulation and measurement results is mainly due to the imperfect fitting process and the estimation error of some physical parameters used in the model, such as N_t, F and σ_0 in the Table 2-1. If these parameters are known accurately (now they are not available from the foundry), then the accuracy of the proposed model is expected to be improved.

In order to future validate the proposed model with various ambient conditions, two more temperatures (-50°C, -40°C) and excess voltages (2.0V, 2.5V) are simulated, and the results are included in Figure 2-10 and Figure 2-11. We can see that the trends are consistent with previous simulations and measurements at other temperatures (i.e., -30°C to +40°C) and excess voltages (i.e., 0.5V to 1.5V). This provides evidence that the proposed model works properly.

2.4.3. After-Pulsing Results

The probabilities of after-pulsing, as functions of the temperature and excess voltage, are shown in Figure 2-12. With 0.5V excess voltage, the after-pulsing probability is 63.5% at -30 °C, and it drops to only 6.6% at 40 °C. Each curve consists of three exponentials, which corresponds to three deep-level traps that are considered

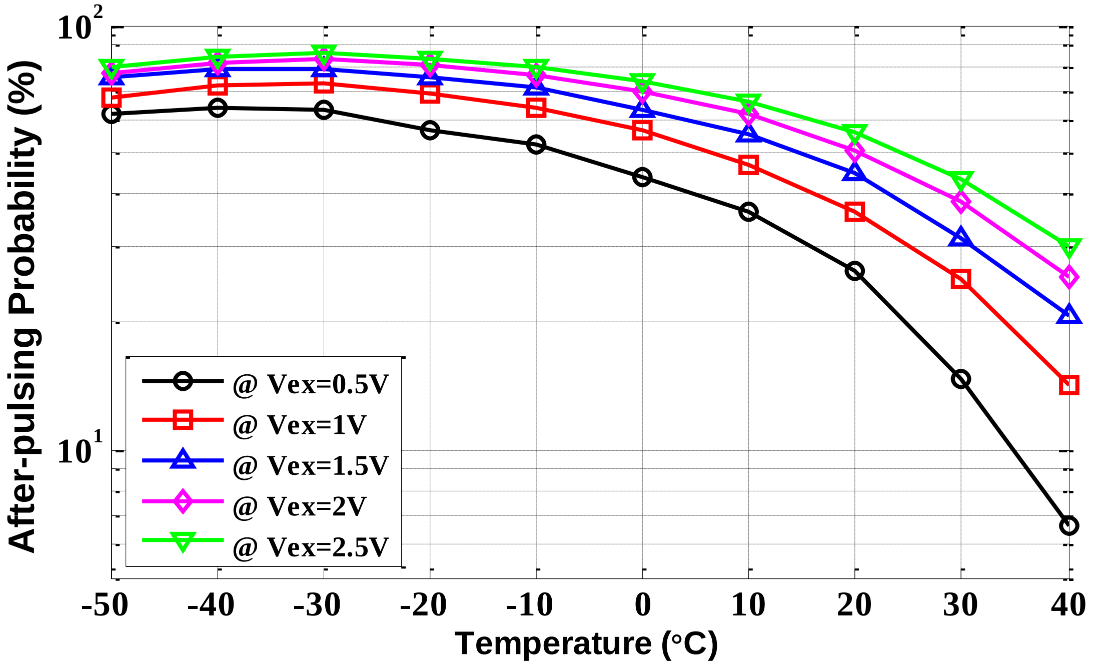


Figure 2-12, Simulation results of the after-pulsing probability with temperatures at five different excess voltages.

in the model. The after-pulsing probability also increases with excess voltage, because the triggering probability is increasing with excess voltage, as revealed in equation (2-7). It is also notable that for the two extended temperature value (-50°C and -40°C), the after-pulsing probabilities seem to be decreasing. The reason is thought to be the imperfection in physical parameter estimation.

Besides the dark counts, we also recorded the total number of carriers generated and carriers released during the simulations. Therefore, we can calculate the triggering probabilities at different temperatures and excess voltages. The results of trigger probability are shown in Figure 2-13. As indicated by equation (2-7), the triggering probability decreases with temperature because of enlarged depletion region depth, but increases with excess voltage because of increased electrical field strength [125]. The probabilities are ~71% and ~60% at -30 °C and 40 °C, respectively, at an excess bias of 1.5V. When the excess voltage is set to 0.5V, it drops to only 46% and 23% at -30 °C and 40 °C, respectively.

In Figure 2-13, it can be observed that these curves have some ripples, for example, in the blue curve for 1.5V excess voltage. A possible reason is that we only considered

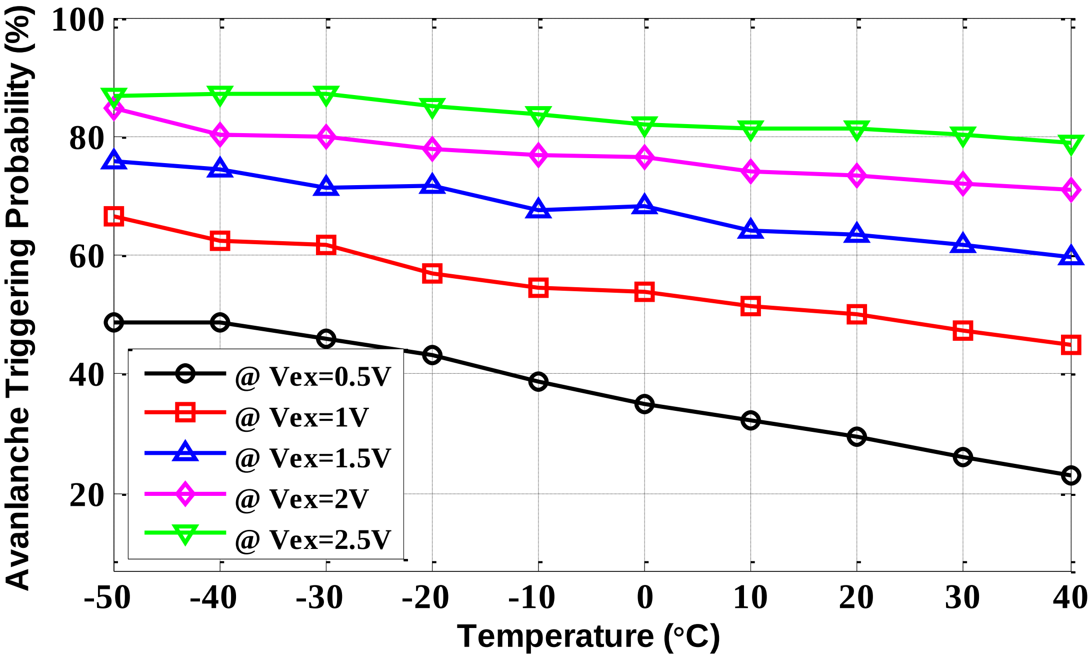


Figure 2-13, Simulation results of triggering probability with temperatures at five different excess voltages.

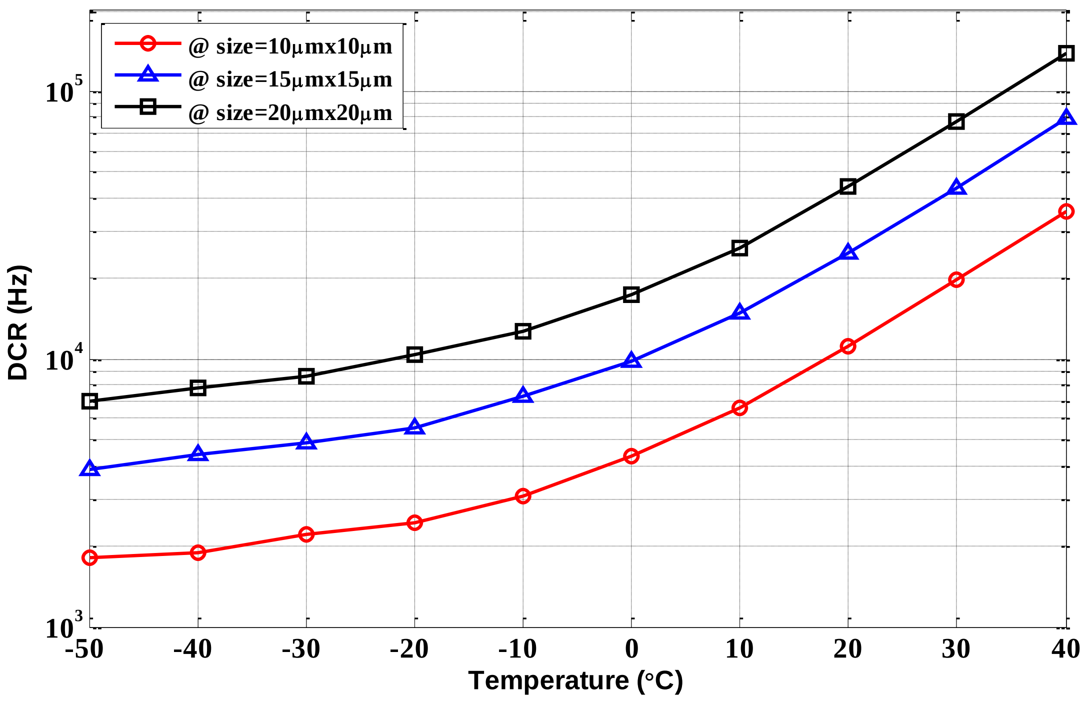


Figure 2-14, Simulation results of DCR with different SPAD size. The excess voltage is fixed at 1V in this simulation.

the temperature dependences of three parameters (i.e. silicon bandgap energy, silicon intrinsic concentration and the breakdown voltage of the SPAD). Under some temperatures and excess voltages conditions, the change in the avalanche triggering probability due to the total number of generated carriers and the number of carriers being released during the emission may slightly increase or decrease.

As an extra step to demonstrate the validity of the proposed model, we use the extracted physical parameters and simulated SPADs with different active areas. The results are presented in Figure 2-14, with a constant excess voltage of 1V. Consistent trends of DCR with temperature is observed for three different SPAD sizes. This means the model can be applied to SPADs with various sizes. It will enable the circuit designers to choose different SPAD design parameters to optimize the associated circuit design, for example to maximize the active area and maximize the fill factor while still maintaining the dark noise within an acceptable level.

2.5. Conclusions

In this chapter, a comprehensive and accurate analytical SPAD model is proposed and validated for circuit simulation purposes. This model includes the static, dynamic and statistical behavior of the SPAD. We believe that this is the first SPAD circuit simulation model that includes the band-to-band tunneling dark noise contribution and temporal dependence of the after-pulsing probability. The process of extracting the model parameters from measurement is discussed. The proposed model is implemented in Verilog-A HDL and the simulations are performed in both Cadence Spectre and Synopsys HSpice tools. The simulation results are validated with measurements, and a maximum relative error of 8.7% is achieved. The work described in this chapter can be used to improve SPAD circuit design and for system evaluation. For example, in a ToF PET imaging system, the timing performance of SPAD detector can be estimated in simulation because it is mainly limited by the noise behavior of the SPAD.

Chapter 3

IMPACT OF SILICIDE LAYER IN CMOS SPAD DESIGN[‡]

Silicide layer has been widely used in standard CMOS processes. Since this layer is close to the active region in a SPAD design, its impact on the characteristics of the SPAD should be studied. In this chapter, two test structures will be compared to investigate the changes in the performance of the SPAD. Detailed measurements and comparison on breakdown voltage, DCR, PDE and after-pulsing probability are to be discussed. The reasons for these changes will also be addressed in detail.

3.1. SPAD in CMOS Technology and the Silicide Layer

As discussed before, SPAD is a widely used solid-state photodetector technology due to its high speed, exceptional photon sensitivity and superior timing performance. It is often used in single-photon counting or single-photon timing applications. When fabricated in CMOS technology, peripheral functional electronics can be easily integrated with SPADs on the same silicon substrate to realize the “optical sensing system-on-chip (SoC)” [15], [16], [25], [45], [110], [143]–[147].

Essentially, the CMOS SPAD is a p-n junction working above its breakdown voltage, also called the Geiger mode. Under such condition, when a photon strikes the diode’s active area, it will trigger the avalanche process, which produces a macroscopic current

[‡] Part of this work was published as: Z. Cheng, D. Palubiak, X. Zheng, M. J. Deen and H. Peng, “Impact of silicide layer on single photon avalanche diodes in a 130nm CMOS process,” Journal of Physics D: Applied Physics, vol. 49, no. 34, pp. #345105-1-11, 2016. Copyright by IOP Publishing.

flowing through the diode. Quenching electronics are then necessary to protect the SPAD from thermal burn out. Resetting electronics are needed to restore the bias voltage of the SPAD and to reset it back to Geiger mode for the next photon detection event. Different quench and reset approaches (i.e. passive, active, mixed) have been studied in the past few years [24], [148]–[153].

In standard CMOS technologies, silicidation is the process of creating a surface layer of metal on silicon in order to reduce the increasing resistance associated with shrinking feature sizes [154]–[158]. Cobalt silicide ($CoSi_2$) is utilized in 130 nm CMOS technology due to its low sheet resistance and high stability [154], [157]. However, silicidation of the source/drain regions becomes a problem because the silicide can penetrate through the shallow junctions, a problem termed the silicide spiking effect. Using transmission electron microscopy (TEM), abnormal $CoSi_x$ spikes were observed to be formed from the Co silicide film towards silicon junctions. These spikes are considered as the origin of random localized leakage current in deep sub-micron (DSM) CMOS technologies [155], [159]–[161]. The study in [160] pointed out that $CoSi_x$ residue between Co film and silicon substrate acts as a solid diffusion membrane which causes the formation of $CoSi_2$ spikes inside silicon. This effect has been shown to greatly affect the leakage current of p-n^+ junctions [154]–[158], [162]–[164]. The silicide layer above the active area will have serious implications for SPADs designed in DSM CMOS process as the silicide leakage current flows not only in the junction periphery, but also in the junction area. Therefore, the motivation of this work is to investigate the impact of silicidation above the active region of the SPAD and how its performance is affected.

In this chapter, we designed, fabricated, measured and compared two different sets of SPAD structures in a standard digital 130nm CMOS process, one with silicidation (SPAD-B) and another without silicidation (SPAD-A) above the active area of the SPAD. The impact of the silicide layer on the performance of the SPAD will be experimentally characterized and analyzed. Detailed characterization of the breakdown voltage, DCR, and PDE of the two SPADs will be discussed, analyzed and compared. Through measurements at several temperatures, the after-pulsing behavior and the activation energies of the shallow-level traps are obtained.

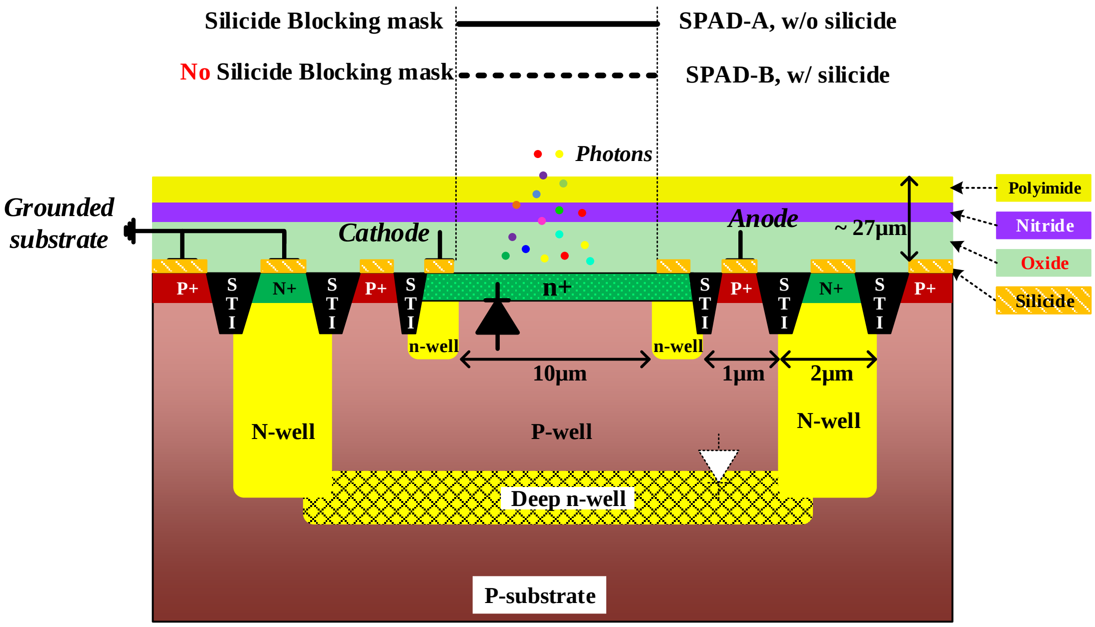


Figure 3-1, Cross-sectional view of SPADs that are used in this chapter.

The rest of this chapter is organized as follows. In Section 3.2, the design and I-V measurements of two SPAD structures are presented. In Section 3.3, we discuss the dark noise measurements. In Sections 3.4 and 3.5, the after-pulsing phenomenon and PDE characteristics of these two SPADs are compared, respectively. Finally, in Section 3.6, the conclusions are provided.

3.2. SPAD Design in 130 nm standard CMOS

3.2.1. SPAD Structure

Figure 3-1 illustrates the cross-sectional diagram of our SPAD devices that were fabricated in IBM's 130nm digital CMOS technology. Note that the dimensions and layer thicknesses in these figures are not drawn to scale, with some numbers labelled in the figure [234]. The SPAD uses an n+/p-well junction for photon sensing. This shallow junction is isolated from the p-substrate by a deep n-well. By implanting an n-well within the p-well and placing the n-well ring around the active area of the SPAD, a guard ring structure is formed to suppress the premature edge breakdown phenomenon. That is, the electric field around the active p-n junction peripheral edges has a higher breakdown voltage than the lateral middle area [30]–[32], [67], [142].

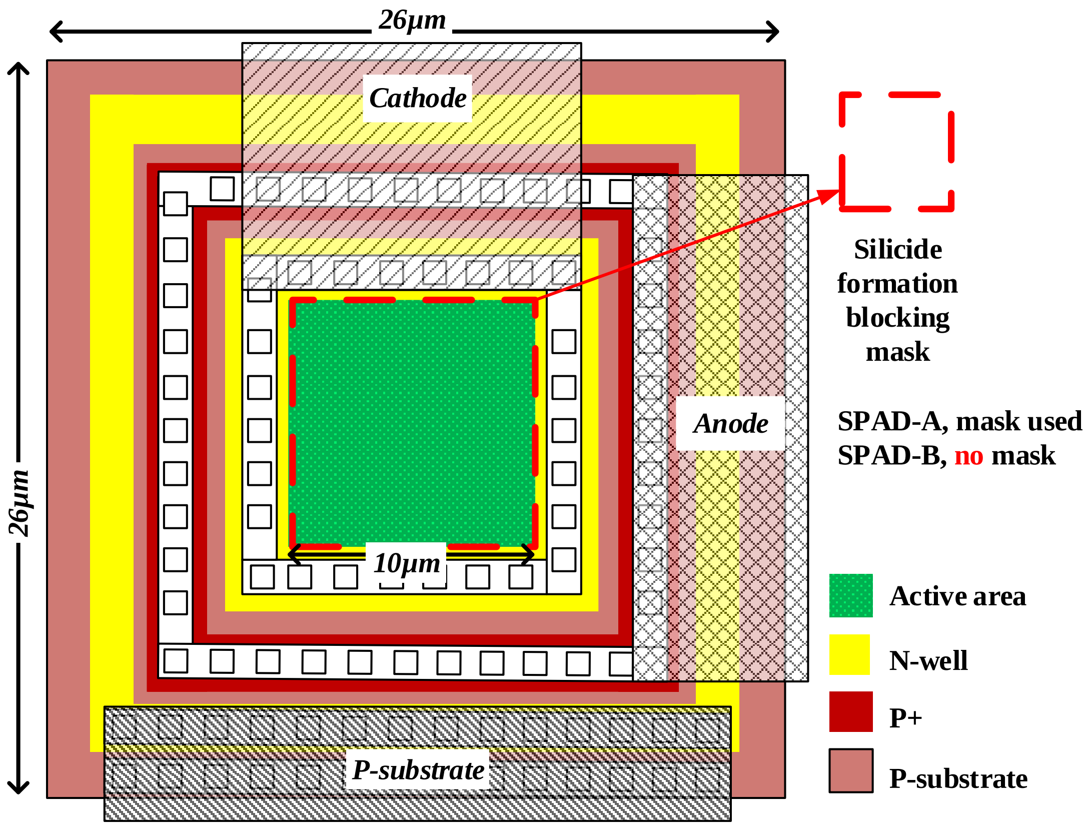


Figure 3-2, Simplified layout view of SPADs that are used in this chapter.

In Figure 3-1, a parasitic diode from the p-well/(deep n-well) junction exists. The breakdown voltage of this deep junction is around 9.4V, that is, it is lower than the shallow junction breakdown voltage. Therefore, when connecting the deep n-well to the substrate, the anode voltage connected to the p-well can be as low as -9V while still avoiding breakdown in the parasitic deep junction.

The SPAD-A and SPAD-B have almost the same design structure except for one difference. A special mask layer available in this technology is used to block formation of the silicide layer above the active region in SPAD-A, while in SPAD-B, this mask layer is not used (see the dashed black line in Figure 3-1). Figure 3-2 shows the simplified layout view of the SPAD. In this figure, the dielectric layers above the silicon are not drawn as they will cover the entire design. Also, we note that the silicide formation blocking layer is only applied to the active region (i.e. the n+ area in green in Figure 3-2) of the SPAD-A to eliminate the silicide layer. The silicided Ohmic contact of the cathode electrodes still exist in both silicided and non-silicided SPADs.

For both SPAD designs, the SPAD signals are sensed on the cathodes, which are connected to a 50 kΩ polysilicon resistor that is used to passively quench the diode. Thus, our SPAD uses a passive-quench, passive-reset (PQPR) front-end circuit in this work. To minimize on-chip local process variations, these structures are placed next to each other. The fabricated SPADs are encapsulated in the PGA69 package and connected to external pins using wire bonding. For both SPADs, they have square shapes, with 10μm×10μm active areas. The entire size of each SPAD pixel is 26μm×26μm. Thus, the SPADs in this work feature a fill factor of 14.8%. The dimensions of the SPAD involves a trade-off between the fill-factor and the dark noise of the SPAD. Since the SPAD is typically combined with front-end circuits, and the area of these circuits are relative small compared to the SPAD, then, by increasing the size of the SPAD, both the fill-factor and optical sensitivity will be improved. However, a larger size of the SPAD also results in an increase in its dark noise, which decreases its SNR. Therefore, there is a compromise in the area of the SPAD when both dark noise and optical sensitivity are considered together [26].

3.2.2. I-V and Breakdown Voltage Measurements

The breakdown voltages of two SPAD structures were measured using a high-precision semiconductor device parameter analyzer, Agilent B1500A, while the diode is in the dark. The breakdown voltage is defined as the point where the I-V curve shows an abrupt increase (i.e. the maximum slope) in current-voltage characteristics.

Figure 3-3 shows the measurement results of breakdown voltages from -30°C to 40°C, for both non-silicided and silicided SPADs. The measurements were repeated for five chips to check for process variations. The thick dashed lines (black and red) present the linear fitting of the mean value of five chips. The standard deviations among five chips are shown as error bars on the dashed lines in Figure 3-3. It is observed that for the entire temperature range, the SPAD breakdown voltage drops about 1V when the silicide layer is eliminated. At 20°C, the breakdown voltages of two SPADs are approximately 10.5V and 11.5V, respectively. The insert in Figure 3-3 gives an arbitrary example of the measured IV curve, in which the current flowing through the SPAD is limited to 1mA in order to protect the diode from thermal burn damage.

We also measured the chip-to-chip variations of the breakdown voltage of the SPAD among 16 test chips at room temperature (~24°C, without temperature control). The results of both sets of SPADs are given in Figure 3-4. The decrease of breakdown

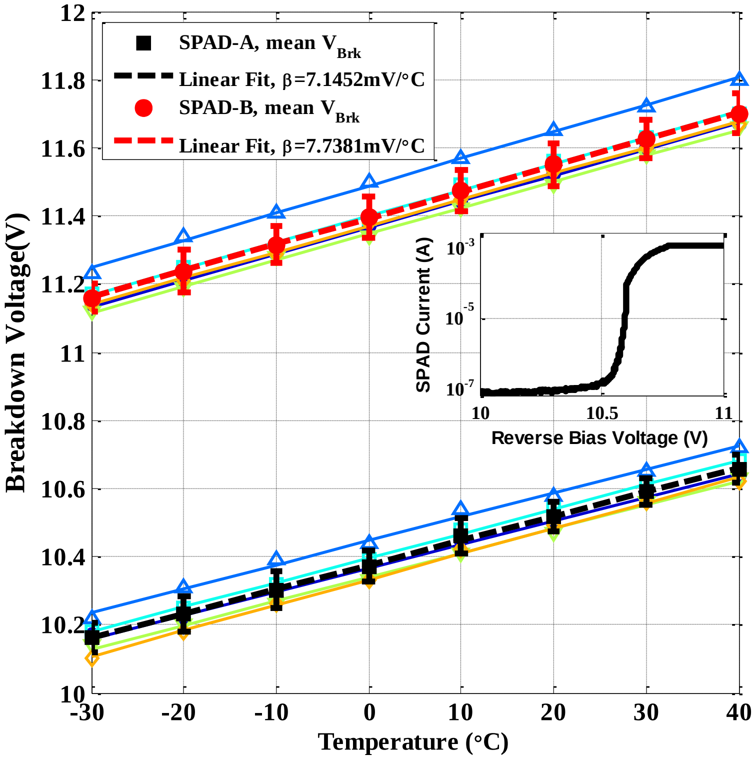


Figure 3-3, Measured SPAD breakdown voltage as a function of temperature. The inset shows an example of the current versus the reverse bias voltage of a SPAD.

voltage in SPAD-A of ~1V is consistent with the previous measurements in Figure 3-3.

In Figure 3-4, the variation of the breakdown voltage of the SPADs is caused by the variations in the doping concentrations. There are both inter-chip and intra-chip variations. The intra-chip variation is due to the slight change in the doping concentration in the active area of the SPAD, which is caused by the formation of the silicide layer above. This results in 1V difference in the breakdown voltage between SPAD-A and SPAD-B. The inter-chip variation among different test chips is due to the doping concentration variation caused by the process induced variations [24], [26]. However, this variation is small, approximately ±0.1V.

In order to investigate the potential causes for the breakdown voltage change, we measured the breakdown voltages of the parasitic junctions (p-well/ deep n-well) in both SPAD-A and SPAD-B devices. They have the same breakdown voltage (~9.4V). This indicates that the doping concentration of the p-well has not being affected by the silicide formation blocking process. Thus, one possible cause may be in the n+ region. The exact doping profile in the n+ region needs more detailed analysis. However, as a preliminary thought, we think that when the silicide layer is formed on top of the SPAD-B's active area, metal impurity migration occurs. Thus, some cobalt atoms

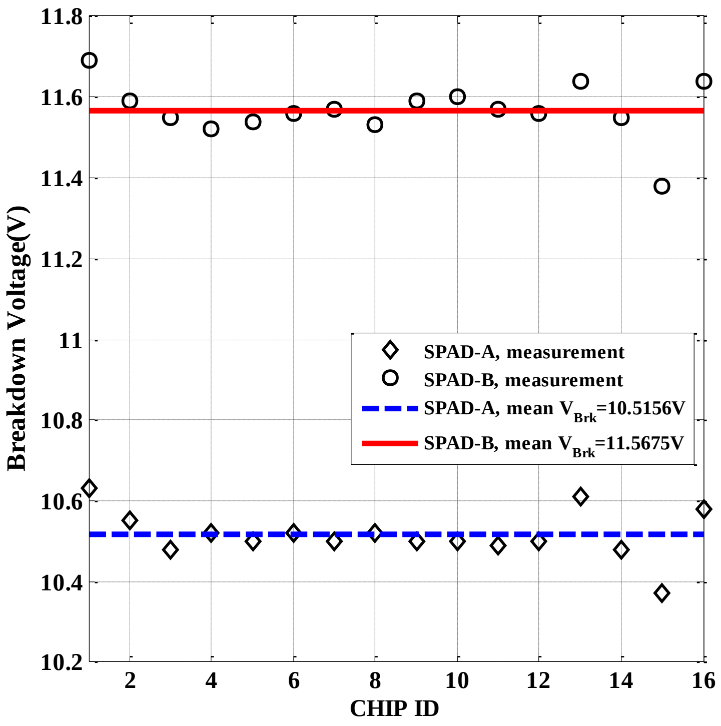


Figure 3-4, Breakdown voltage of the SPAD chip-to-chip variation from 16 test chips measured at room temperature (~24°C, ambient conditions).

diffuse into the n+ region and they act like p-type dopants. Consequently, the doping concentration in the active area of the silicided SPAD is reduced, which results in a lower peak electric field and a higher breakdown voltage compared to the non-silicided SPADs.

The slope of the linear fitting gives the linear temperature coefficient of breakdown voltage [32], [67], [142], which is the factor β in the following equation.

$$V_{brk} = V_{brk0} + \beta \times (\mathrm{T} - \mathrm{T}_0). \tag{3-1}$$

The temperature coefficients are found to be 7.1mV/°C and 7.7mV/°C for SPAD-A and SPAD-B, respectively. These results are consistent with our previous findings [32], [67], [142].

3.3. DCR Characterization and Comparison

Dark counts are avalanche pulses that are triggered when the SPAD is in the dark (i.e. without light illumination). DCR is an important parameter that defines the noise of the SPAD. Because of the high doping concentrations in standard CMOS processes,

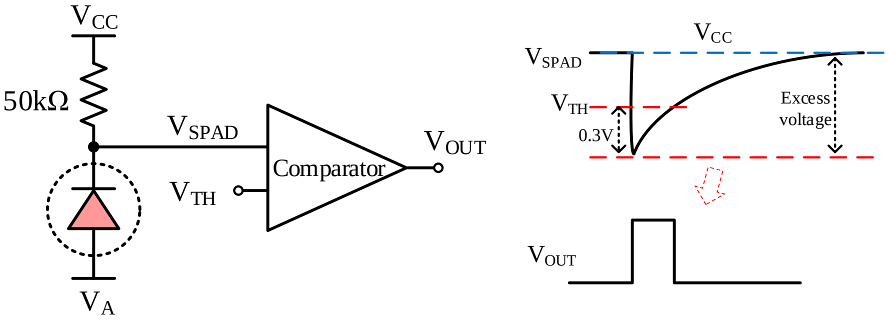


Figure 3-5, Readout circuit for SPAD characterization.

especially in the DSM technologies, SPADs fabricated in advanced nodes of standard CMOS usually have higher DCR per unit area, when compared to PMTs [165] or SPADs fabricated in a custom/imaging CMOS technologies [26].

Sources of dark noise generally fall into two categories [124]–[128], namely, the primary and secondary dark noise. The primary dark noise is mainly contributed by the Shockley- Read-Hall (SRH) generation-recombination and the tunneling mechanisms. The secondary dark noise is due to the after-pulsing process, which is related to the carriers being trapped and subsequently released from shallow traps in the silicon bandgap. The detailed analysis of after-pulsing will be discussed in the next section.

The DCRs of both SPADs are characterized using the read-out circuit shown in Figure 3-5. The threshold voltage (V_{TH}) of the comparator is set to 0.3V above the lowest voltage level of the SPAD cathode voltage (V_{SPAD}). Thus, the analog output from the SPAD can be digitized (V_{OUT}) and then be collected using a 20 GS/s LeCroy WaveRunner 625Zi oscilloscope. For each measurement, a large number of counts are accumulated to obtain good statistical accuracy (i.e. yielding the DCR histogram approaching to Gaussian statistics). From measurements, we observed that SPAD-B (with silicidation) has higher DCR than SPAD-A, so we set the total counts threshold of SPAD-A and SPAD-B to 100,000 and 1,000,000, respectively. These counts are considered sufficient in the characterization of the dark noise of SPADs to ensure negligible measurement error [27], [122], [124].

3.3.1. DCR with Excess Voltage

The DCR increases with excess voltage (V_{ex}), because the triggering probability increases as V_{ex} gets higher [133]–[135]. The measurements were performed with

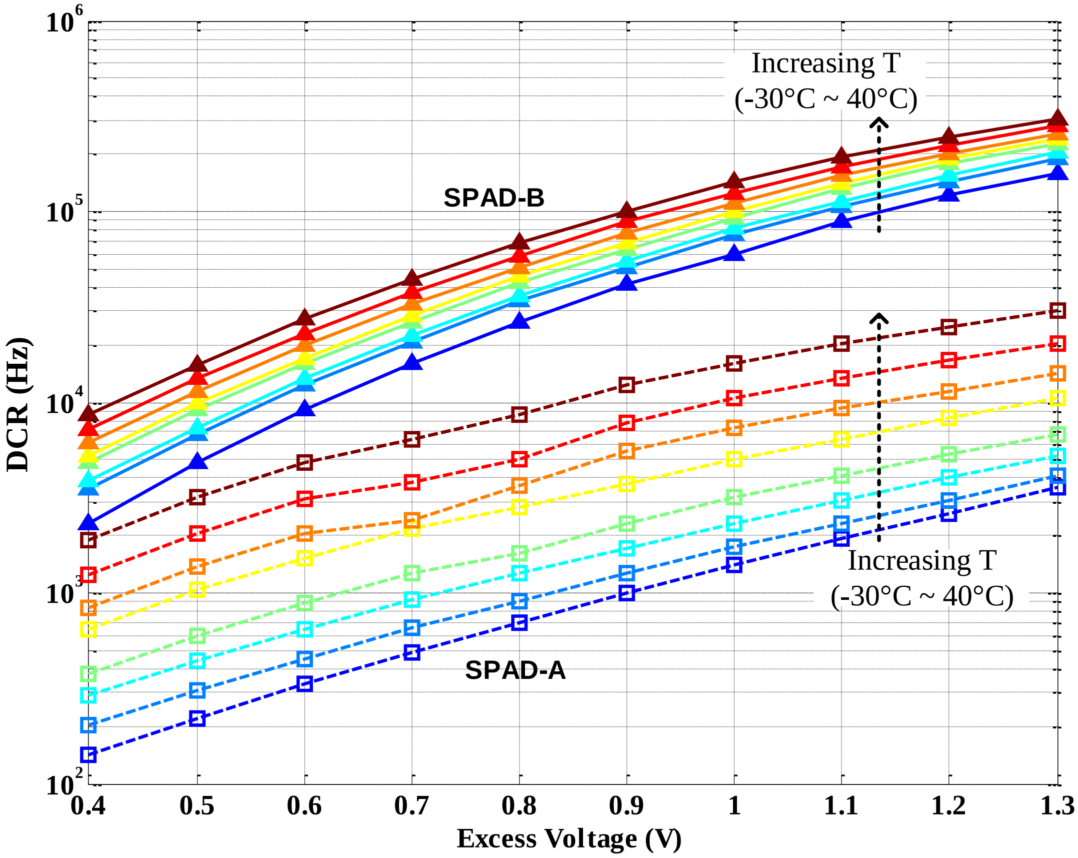


Figure 3-6, Measured DCR with excess voltages. The temperature changes from -30°C to +40°C by 10°C per step.

excess voltages increasing from 0.4V to 1.3V in steps of 0.1V. Then, they were repeated for eight different temperature settings (-30°C ~ 40°C). The temperature dependence of breakdown voltage was also taken into consideration when setting the excess voltage.

As shown in Figure 3-6, the DCRs of both SPADs increase with the excess voltage. Under the entire excess voltage range, the SPAD without silicidation shows much lower DCR than its silicided counterpart. For example, at -30°C and 0.4V excess voltage, SPAD-A has about 16 times lower DCR than SPAD-B. This difference increases with excess voltage and it reaches to a factor of 44 at 1.3V excess voltage. However, the DCR difference between two SPADs becomes weak as the temperature increases. At 40°C, the DCR improvement factor of the non-silicided SPAD is only ~4.5 at 0.4V V_{ex}, and ~10 at 1.3V V_{ex}, respectively.

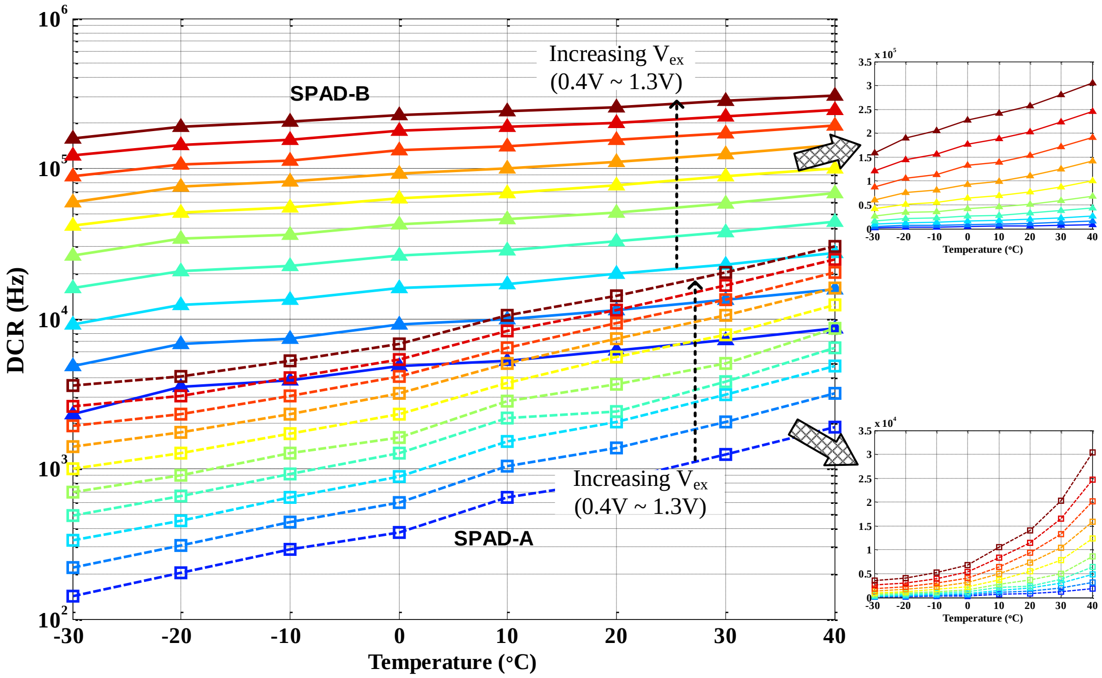


Figure 3-7, Measured DCR with temperature. The excess voltage Vex changes from 0.4V to 1.3V by 0.1V per step.

From the trend of these curves in Figure 3-6, we can see that for SPAD-B, its DCR increase faster with excess voltage by at least one order of magnitude. This suggests that both its electric field profile and the ionization coefficients for electrons and holes differ from the non-silicided SPAD. A detailed analysis for this cause requires TCAD process simulation. But one possible reason may be as follows. Since the non-silicided SPAD has higher donor doping concentration, the width of it depletion region gets smaller. Thus, the electrical field strength is higher with the same excess voltage.

3.3.2. DCR with Temperature

The SRH generation-recombination process has a significant dependence on temperature, while the tunneling process is relatively temperature independent [124], [125]. Therefore, studying the temperature dependence of DCR can help to determine the source of the dark noise of the SPAD and their dominant range.

Figure 3-7 presents the measured DCRs as a function of temperature (from -30°C to 40°C, 10°C per step) at ten different excess voltages (from 0.4V to 1.3V, 0.1V per step) for both SPADs. It is noted that the slopes of DCR versus temperature for the silicided SPAD are almost constant, as shown with solid lines in Figure 3-7. They are smaller compared to the slopes for non-silicided SPAD indicated with the dashed lines. In order to clearly show the differences in the temperature dependence between SPAD-A and SPAD-B, two plots using linear scales are inserted in Figure 3-7. The top insert is for SPAD-B and it shows a linear dependence. On the other hand, for SPAD-A shown in the bottom insert, an exponential relationship with temperature was observed. This means that SPAD-B's DCR is less temperature dependent compared to SPAD-A. Thus, the tunneling mechanism is the main source of DCR for SPAD-B. For the non-silicided SPAD, the slopes reduce exponentially with reduced temperature, as seen in the bottom insert. Thus, it is a combination of both SRH thermal generation and tunneling transport. The SRH contributes over the entire temperature range, but tunneling is the dominant source at lower temperatures (for example within -30°C to 0°C range) [166].

We believe that by eliminating the silicidation layer in SPAD-A, more traps are resulted in silicon because of a direct touch of silicon and silicon oxide [167]–[169]. As will be discussed later in Section 3.4, the SPAD-A has three shallow traps while SPAD-B has two. Thus, trap-assisted tunneling process becomes more evident in the low temperature region in Figure 3-7. This can also explain the observation of an increase in the after-pulsing probability of SPAD-A, which will be discussed in detail in Section 3.4.

3.3.3. DCR Chip-to-Chip Variations

The chip-to-chip variations of DCRs of both non-silicided and silicided SPADs are studied at room temperature using 14 chips. The results are presented in Figure 3-8. The different symbols in this figure represent different test chips. The large variation in the DCR are partially due to variations in process parameters and dimensions of the SPAD. These variations occur because our SPADs were designed in a standard digital CMOS process that was not optimized for optical sensors. In this plot, four chips that appear two orders of magnitude higher than the average DCR are excluded from the data analysis. We consider that these four chips may have severe fabrication defects in the SPAD structures. Their performances are not representative of the majority of SPADs fabricated. Other effects, such as premature edge breakdown phenomenon or junction punch through, might also occur.

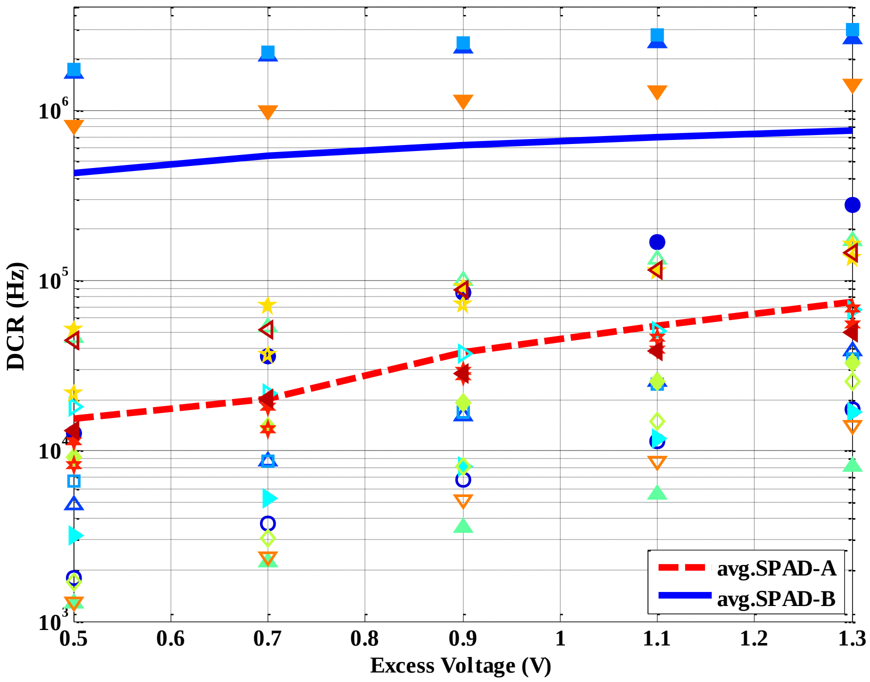


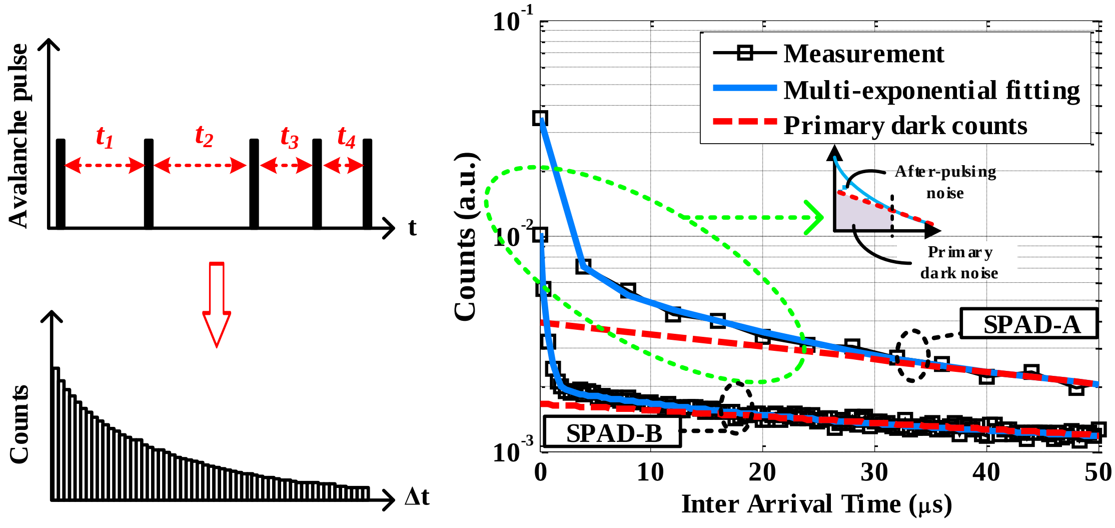
Figure 3-8, DCR chip to chip variations. In total, 14 chips are measured at room temperature. The chips with extremely high dark noise (>10^7 Hz) are excluded from this figure, since their defects are far above the average level. The empty and solid symbols represent SPAD-A and SPAD-B, respectively.

As seen from Figure 3-8, on average, the non-silicided SPAD shows at least one order of magnitude lower DCR than the silicided SPAD at room temperature over the whole excess voltage range. This is consistent with previous measurements. The higher DCR in SPAD-B is now explained [170].When the silicide layer is formed above the active region, the stress-induced defects and metal impurity contaminations are also introduced in the n+/p-well junction. These defects will act as recombination centers, resulting in higher dark noise. Thus, it is expected that the silicided SPAD-B will have a higher DCR than non-silicided SPAD-A.

3.4. After-Pulsing Phenomenon

3.4.1. Background

During an avalanche process, some carriers will be trapped by shallow traps in the silicon bandgap. After a random time, they will be released, which will then cause a false trigger if the SPAD has been restored to Geiger mode. This is referred to as after-pulsing. After-pulsing counts will add to the total DCR and consequently, the measured signal-to-noise ratio will be degraded in photon counting applications.



(a)

(b)

Figure 3-9, (a) Illustration diagram of the inter-arrival-time (IAT) measurement. (b) SPAD-A and SPAD-B IAT measurement and fitting examples at -20°C and 0.4V excess voltage. The measurement data are fitting with four (SPAD-A)/three (SPAD-B) exponentials. For all the data fitting process, R^2 of 0.99 is achieved. The longest tail presents the primary dark noise. The insert shows that the after-pulsing probability is calculated as the area ratio between the dashed region and total area.

The after-pulsing process will contribute more dark noise when the temperature is reduced. This is because the lifetimes of these shallow traps become longer at lower temperatures, as described by the following equation,

$$\tau = \tau_0 e^{\frac{E_a}{kT}}, \tag{3-2}$$

where E_a is the activation energy which is defined as the energy difference between the shallow trap's energy level and the bottom of conduction band for electron trap (or the top of the valence band for hole trap), and τ_0 is the pre-exponential factor. Also, k is Boltzmann's constant and T is the temperature in Kelvin. Thus, more trapped carriers will be released at lower temperatures when the SPAD is restored back to Geiger mode, and these released charges will have a higher probability to initiate a false avalanche.

3.4.2. Experimental Characterization and Parameter Extraction

We also characterized the dark noise of the SPAD using the inter-arrival-time (IAT) measurement previously described in [32], [123], [139], [171], and illustrated in Figure 3-9(a). The time intervals between two successive avalanche events were recorded using the built-in time stamping function of the LeCroy oscilloscope. After at least 1,000,000 counts were collected, the detection probability histogram can be obtained, as shown in the example in Figure 3-9(b). The IAT measurements were carried out from -30°C to 40°C, with the temperature controlled by a temperature chamber. The temperature chamber has ±3°C variations. It is noted that in order to extract the activation energy and lifetime associated with the shallow traps, the SPADs must be cooled since the trap's lifetime increases with decreasing temperature. If the SPADs are not cooled, then the trap's lifetime is relatively short, making the measurements less accurate.

The pulse detection probability can be modelled with the following multiple exponential equation,

$$P_{pulse}(t)=\sum_{i=1}^{N}A_i\frac{1}{\tau_i}e^{-\frac{t}{\tau_i}}. \tag{3-3}$$

where the index *i* presents the *i*-th trap, *N* is the total number of shallow traps, and A_i and τ_i are the exponential pre-factor and the lifetime associated with the *i*-th trap, respectively. A_i will determine the initial magnitudes of trapped carriers in the bandgap, which can be obtained by setting *t* to zero. Also, *t* is the time interval between the avalanche and its associated carrier release.

Through the multiple-exponential fitting of the measured IAT data, we can extract all the factors involved in equations (3-2) and (3-3). In this analysis, four exponentials for SPAD-A and three exponentials for SPAD-B are needed to achieve better than 0.99 R^2 for the fitting process, respectively. Thus, according to the fitting results, three shallow traps are considered for SPAD-A and two for SPAD-B. A possible reason for this result is that the direct interface between silicon and silicon oxide may introduce shallow-level defects into silicon lattice due to various bonding changes. In SPAD-B, the silicide layer acts as a buffer layer between Si and SiO_2. The publications [167]–[169] provide for more details on how bonding imperfections lead to trap centers. The longest tails in Figure 3-9(b) (also being highlighted with red dashed lines) represent the primary dark counts.

By calculating the area between the blue solid line (i.e. the multi-exponential fitting, total DCR) and the red dashed line (i.e. the primary DCR), and dividing it by the total area under the blue solid line, we can get the after-pulsing probability of the SPAD. The calculated results are shown in Figure 3-10. SPAD-A has about 12% after-pulsing

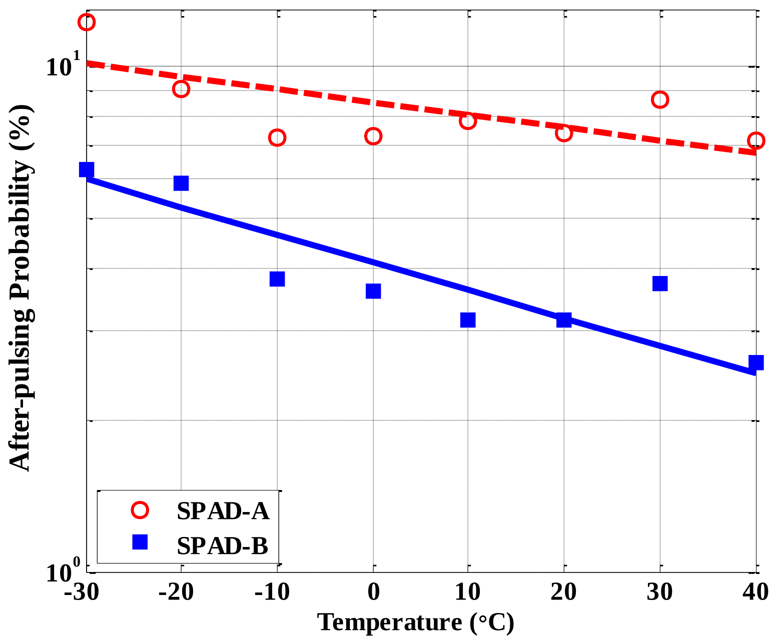


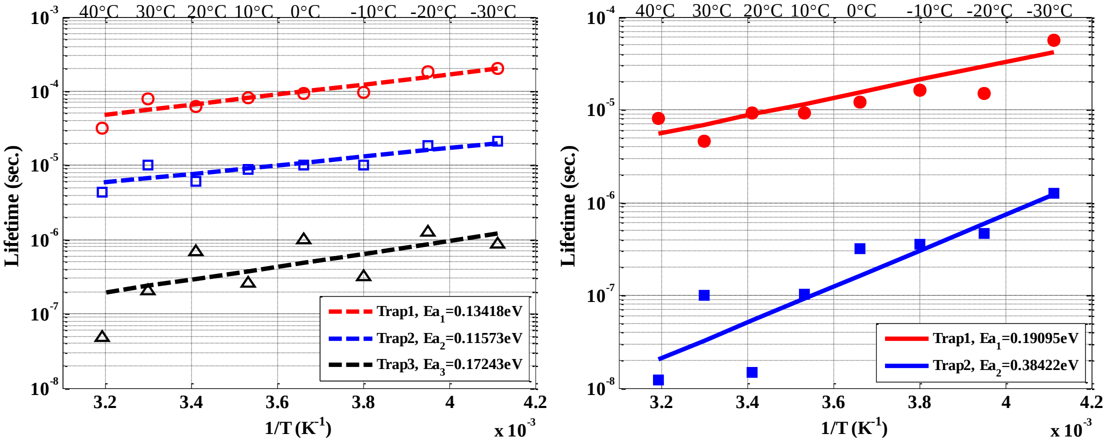
Figure 3-10, After-pulsing probabilities for SPAD-A (no silicidation) and SPAD-B (silicidation). The excess voltage is 0.4V in this example.

probability at -30°C and it drops to 7% at 40°C. The after-pulsing probability of SPAD-B is much lower. It is 6% at -30°C and 2.6% at 40°C, respectively.

In order to investigate the types of shallow traps existing in these SPADs, we further extracted the trap's lifetime at different temperatures. Then, the Arrhenius plots for the trap lifetime were used to determine the associated activation energies. By taking natural logarithms of equation (3-2), we have the following relationship,

$$\ln(\tau) = \ln(\tau_0) + \frac{E_a}{k} \times \frac{1}{T}, \tag{3-4}$$

Therefore, when we plot the log of measured traps' life times versus 1/T (a semi-log plot), the slope (*E*E_a/k) obtained thought a linear fitting will provide the activation energy (E_a) of each trap. The results are given in Figure 3-11(a) and Figure 3-11(b) for SPAD-A and SPAD-B, respectively. The non-silicided SPAD (SPAD-A) has three shallow traps, with activation energies of 0.13eV, 0.12eV and 0.17eV, respectively. For the silicided SPAD, two shallow traps are observed, with active energies of 0.19eV and 0.38eV, respectively. This indicates that these shallow traps are located close to the conduction band in the bandgap.



(a) (b)

Figure 3-11, Arrhenius plots for (a) SPAD-A and (b) SPAD-B. The activation energies of deep-level traps are extracted from these Arrhenius plots

The extracted activation energies and life time constants of the traps can be used to explain why the non-silicided SPAD have a higher after-pulsing probability. Since there are three shallow traps existing in SPAD-A, and they have lower activation energies and longer lifetime constants, they will have higher probabilities to capture carriers during the avalanche and then release these carriers when SPAD-A is reset back to Geiger mode. This yields larger after-pulsing probability compared to the silicided SPAD.

3.5. PDE Measurement

The PDE of an SPAD is the statistical probability that an incident photon will produce an avalanche pulse. It represents the photon sensing efficiency of a SPAD, and it is wavelength dependent. To evaluate the PDE performance, a xenon lamp that generates continuous wavelength light is used in our experiment. Optical bandpass filters with 10nm bandwidth are then used to select the desired wavelength and neutral-density filters are used to attenuate the incident light intensity to keep our SPADs from saturating. Before the actual optical measurement, the set-up was validated using a commercial SPAD module from MicroPhoton Devices, and the incident photon flux was calibrated by a power meter. The incident photon flux can be calculated using the following equation.

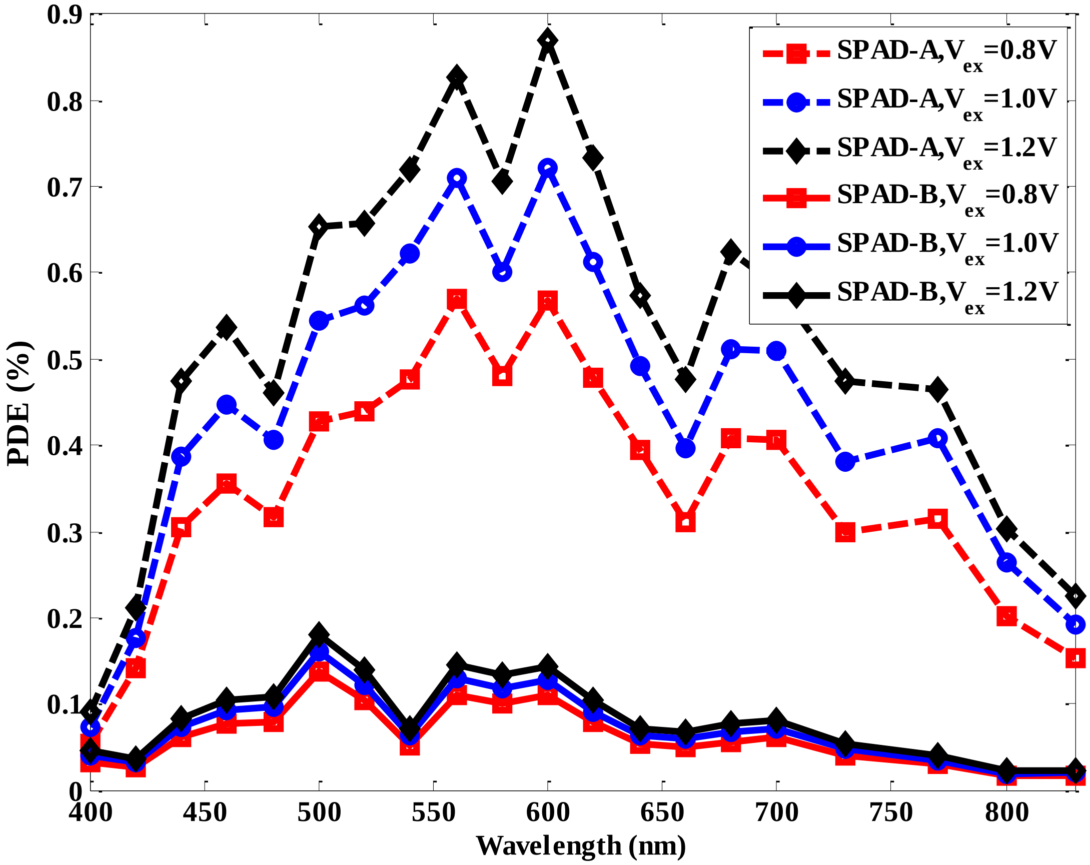


Figure 3-12, PDE measurement results for two SPADs.

$$\Phi_{in} = P_{in}\left(\frac{\lambda}{hc}\right)\left(\frac{A_{SPAD}}{A_{laser}}\right) \tag{3-5}$$

where P_{in} is the light power measured by the power meter, λ is the wavelength, h is the Planck's constant, c is the speed of light, A_{SPAD} is the active area of the SPAD and A_{laser} is the spot area of the laser beam.

The measurement results of PDE are given in Figure 3-12. The ripples in these measured curves are due to light reflection between all the dielectric layers and silicon itself. The curve is expected to be smoother if less layer exists above the silicon.

It is noted that after removing the silicide layer, the PDE of SPAD-A can reach to about 0.9% at 600nm, which is improved by a factor of 6 compared to SPAD-B. The increase in PDE is due to two factors. First is the increase in light transmittance when the silicide layer is removed. The second factor is the increased quantum efficiency in SPAD-A. As pointed out in [170], when a silicide layer is formed above the active area

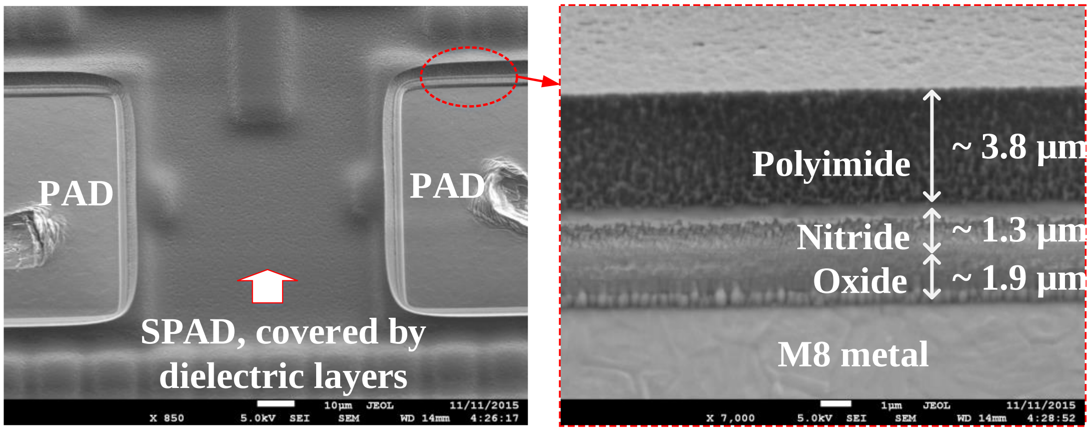

PAD
PAD
SPAD, covered by dielectric layers
x 850 5.0kV SEI 10µm JEOL SEM WD 14mm 11/11/2015 4:26:17
Polyimide
~ 3.8 µm
Nitride
~ 1.3 µm
Oxide
~ 1.9 µm
M8 metal
x 7,000 5.0kV SEI 1µm JEOL SEM WD 14mm 11/11/2015 4:28:52

Figure 3-13, SEM photos of the fabricated SPAD device.

of SPAD-B, deep-level defects (with ~0.55 eV activation energies) are introduced in the silicon. These deep-level defects act as recombination centers of photo-generated carriers. This effect will reduce the photon to carrier conversion efficiency, resulting in a decreased quantum efficiency. In contrast, for SPAD-A, since the silicide layer is eliminated, less defects are in the silicon and we can expect a higher quantum efficiency compared with SPAD-B with the silicide layer [170]. Due to the light reflection and transmittance at the dielectric layers above the active area, SPADs fabricated in standard CMOS process usually have limited PDE, similar to previous publications [67], [142], [172], where their PDEs were 2%, 5%, and 1%, respectively.

Custom SPADs can achieve very high PDE since the doping profile adjustment and post-processing (such as anti-reflection coating, dielectric layers optimization for optical detection, metal thinning and etching steps) are available to maximize their optical detection [24], [173], [174]. Their peak PDEs can be as high as 50%~73%, and for SPADs built in standard imaging CMOS processes, their PDEs vary from 11% to 55% [24]. For the ones in standard digital CMOS, the PDEs are relatively low (from 0.3% to 2.5%) [61], [67], [142], [172], but these processes have the greatest potential to provide the lowest cost for SPADs fabrication and circuit integration.

Figure 3-13 shows the images of the SPAD taken with a scanning electron microscope (SEM). As seen from the left side of Figure 3-13 below, the SPAD is covered by dielectric layers and are not visible from the top view. About 7 µm thick dielectric layers are placed on top of the SPAD. Even thicker oxide/nitride layers (over 20 µm, according to the technology manual) are below the M8 metal layer and above

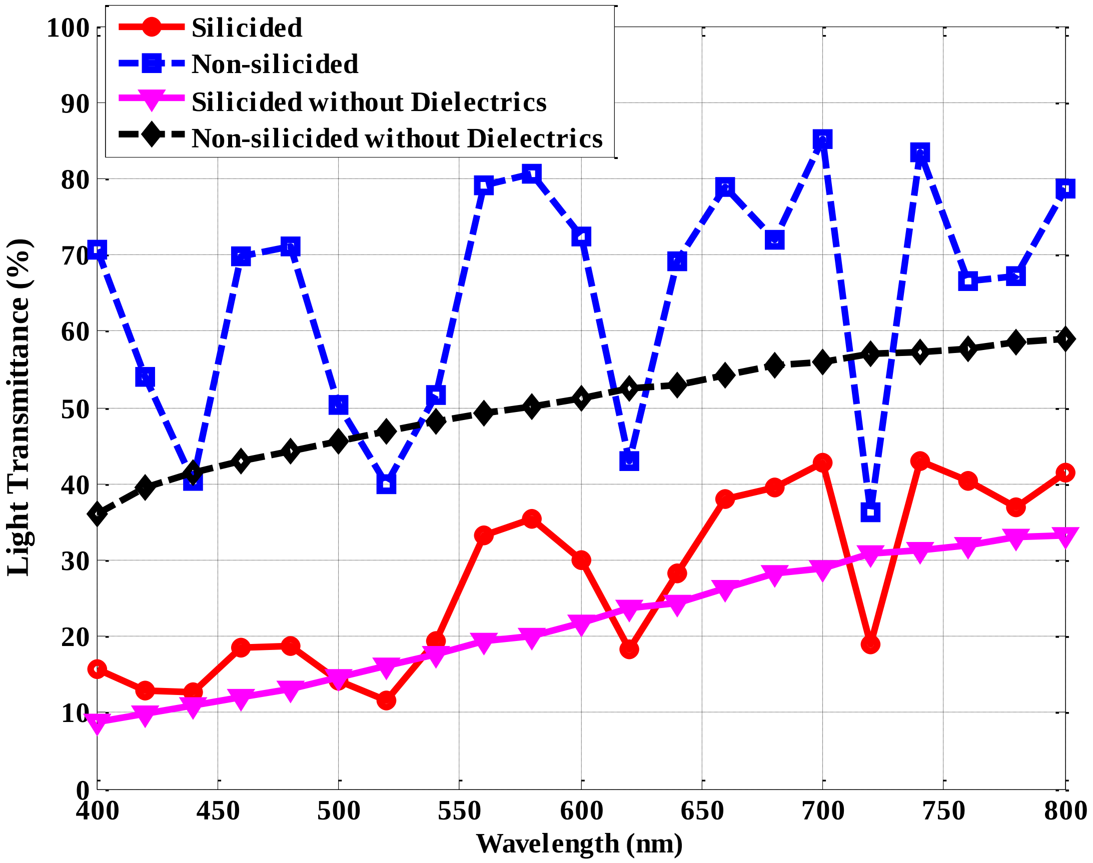


Figure 3-14, Calculated probability that photons can transmit to the SPAD active region, when assuming that the dielectric and silicide layers are partly or entirely removed.

the active region of the SPAD.

Considering the relatively low PDE of SPADs in standard digital CMOS processes and the fact that several dielectric layers exist above their active areas (such as polyimide/nitride/oxide layers in the IBM's 130nm digital CMOS technology used in this work, and their thicknesses are not optically optimized), we had a calculation using the transfer matrix method [23], [175] to obtain the light transmittance into the silicon. This will help to understand how these layers impact the optical detection process. The thicknesses of the dielectric layers and their refractive indexes used in the calculations are from the fabrication process manual and literature [175], [176].

The calculated results are shown in Figure 3-14. The solid red and dashed blue lines present the light transmission improvement when only the silicide layer is removed. We can see that 2 to 5 times more photons are transmitted now. The dashed black line is when all dielectric and silicide layers are removed, while the pink line still keeps the

Table 3-1, Performance summary and comparison of the SPADs

	V_{ex}/V_{Brk} (V/V)	DCR @RT (cps/μm^2)	After-pulsing	Peak PDE (% @ nm)
SPAD-A	0.4/10.5	10	7%	0.9 @ 600
SPAD-B	0.4/11.5	80	3%	0.2 @ 500

silicide layer. Since less layers above silicon, much smoother transmittance profiles are obtained for both cases. Our calculations indicate that the light transmittance will be greatly improved by around 5 times when the dielectric layers and the silicide layer are removed. And this does not taking into consideration the increase of the quantum efficiency of the non-silicide SPAD yet. Thus, we believe that when a specialized CMOS imaging sensor (CIS) process that allows doping profile optimization is used for fabrication, or specialized post processes can be employed for the surface treatment, the PDEs of our SPADs can be significantly improved.

In Table 3-1, a summary of SPAD performance, is provided. By removing the silicide layer (but still keeping the other dielectric layers), the DCR and PDE are improved by factors of 4.5 (minimum) and 5, respectively. The after-pulsing performance becomes about two times worse due to more shallow-level traps in the silicon band-gap. It is noted that the peak PDE of non-silicided SPAD shifts towards the red range (600nm) because the light transmission property is changed.

3.6. Conclusions

We have studied the impact of the silicide layer on the performance of the SPAD in a standard 130nm digital CMOS process. A specialized mask layer available from this technology is used to block the formation of the silicide layer above the active region of the SPAD. We observed that DCR and PDE performances of the SPAD are improved by a factor of 5. The after-pulsing phenomenon becomes more prominent (an increase from 3% to 7%) in the non-silicided device. Since the silicidation process has been intensively employed in the DSM CMOS technologies to provide reliable and reduced source/drain contact resistance for devices with scaled-down geometries, the impact of silicide layer is expected to be similar despite that SPADs can be fabricated in different CMOS technologies. We believe that the work in this chapter can provide guidance and reference for SPAD design targeting improved performance when silicidation blocking is feasible.

Chapter 4

TIME-TO-DIGITAL CONVERTER DESIGN IN CMOS TECHNOLOGY[§]

As we mentioned in Chapter 1, a specialized circuit is needed to measure the time interval between two coincidence gamma ray detections. High resolution, high precision and low power consumption will be the desired features for a TDC to be employed in the SPAD/TDC integration system for PET imaging applications. In this chapter, a TDC will be designed to meet these specifications. First, the background of TDCs in CMOS technology will be reviewed. Then, the architecture and operation principles of the proposed TDC will be discussed. Next, the detailed circuit implementations will be given, which will be followed by the prototype characterization at last.

4.1. CMOS TDC Background

With the digitization function which is analogous to an ADC, a TDC is used to precisely measure the time intervals between two events and to quantize them into digital codes. It can be implemented in various technologies, such as ECL Gates [78], [80], CMOS [68]–[70], FPGA [92], [94]–[97], [101], [104], [105], [107], [108], and BiCMOS [81] . However, in order to facilitate system integration and reduce cost, on-

[§] Part of this work was published as: Z. Cheng, X. Zheng, M. J. Deen and H. Peng, "Recent developments and design challenges of high-performance ring oscillator CMOS time-to-digital converters," IEEE Transactions on Electron Devices, vol. 63, no. 1, pp 235-251, 2016. and Z. Cheng, M. J. Deen, and H. Peng, "A low-power gateable Vernier ring oscillator time-to-digital converter for biomedical imaging applications," IEEE Transactions on Biomedical Circuits and Systems, vol. 10, no. 2, pp 445-454, 2016. Copyright granted by IEEE.

chip TDC in standard CMOS technology is the preferred choice. For typical low-light biomedical imaging (e.g. ToF PET [59], [179] and FLIM [180]–[183]) and ranging applications (e.g. Laser Radar [61], [73]), TDCs are often coupled with SPAD arrays to build the CISs with single photon sensitivity [111], [112], [142], [146], [184].

4.1.1. TDC in Biomedical Imaging Applications

TDCs have been widely adopted in biomedical imaging applications to provide timing information, such as the ToF PET [66], [185], and FLIM [26], [62]. In the case of ToF PET imaging, many research efforts are devoted to integrating TDCs with an array of SPADs on the same substrate and to build the novel dSiPM in standard CMOS technology [142], [186], [187].

As the conceptual diagram of a ToF PET system in Figure 4-1 shows, the TDC collects the ToF information, that is, the time interval between two coincidence gamma rays (the *Start* and *Stop* signals) that come from one annihilation event [188], [189]. Compact size, low power and high resolution are among the most demanding performance requirements of the TDC used in ToF PET imaging. Further, miniaturization of the peripheral circuitry helps to improve the fill factor of the SPADs in the dSiPM, which results in enhanced photon detection efficiency. The low power feature is important for array design. The superior time resolution achieved with a TDC greatly improves the SNR of the reconstructed images. This improvement in SNR is given by Equations (1-1) and (1-2). Equation (1-3) presents the effective sensitivity increase in PET imaging when the ToF information is used in the image reconstruction process.

An example about the SNR improvement calculation was provided in [11]. With 500ps time resolution, the SNR improvement was about 2.3 times and the effective sensitivity increase was about 5.3 times when implemented in a whole body ToF PET system. The image's SNR improvement is increased due to the following reason: in a non-ToF PET, the noise from all pixels along a given LoR is correlated; while in a ToF PET, the data back-projected along a LoR statistically contribute to a reduced number of adjacent image pixels [11]–[13]. This is also depicted in Figure 1-2 in Chapter 1. From Equations (1-1) and (1-2), we also note that the benefit of ToF increases as the timing resolution improves. In terms of the final reconstructed images, this will greatly

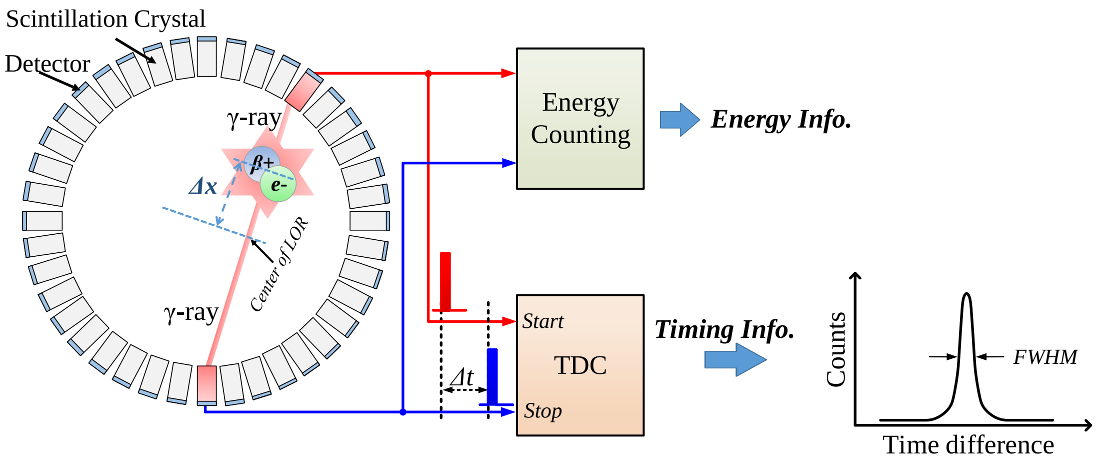


Figure 4-1, TDC utilization in biomedical imaging applications. (a) Conceptual diagram illustrating a TDC collecting the arrival time difference in a ToF PET system. Here, the timing information measured by the TDC (Δt) is used to locate the annihilation position ($\Delta x=\Delta t*c/2$) with respect to the center of the line-of-response.

improve the accuracy of diagnosis, especially for early-stage cancer detection, which eventually improves the survival rate of patients.

Besides the TDC, the noise of photodetectors and the delay time of scintillators also contribute to the ToF PET system's timing resolution. With the superior resolution available from advanced TDC approaches, the system's time resolution will be largely limited by these two noise sources - photodetector and scintillation crystal. A high resolution TDC in the ToF PET imaging system would help to enhance contrast recovery for small lesions and improve the accuracy of diagnosis. This, in turn, allows for earlier tumor detection and improves the survival rate of patients [11], [26].

Another example of a biomedical imaging application is FLIM, which has been extensively used in biological, biochemical or biophysical processes for functional imaging and environmental monitoring [49], [60], [180], [182], [183], [190]–[195]. A diagram of such a system with TDC is shown in Figure 4-2. The decay time of the fluorophore is measured with respect to the external excitation pulse, which typically comes from a laser source. After repeated measurements, a decay-time histogram is obtained and fitted using the mono-exponential or multi-exponential functions. Then the lifetime constants, depending on the types of fluorophores, can be extracted to present the intrinsic property of the fluorophore and its local chemical and physical environment. The time resolution and dynamic range of the TDC need to be sufficient

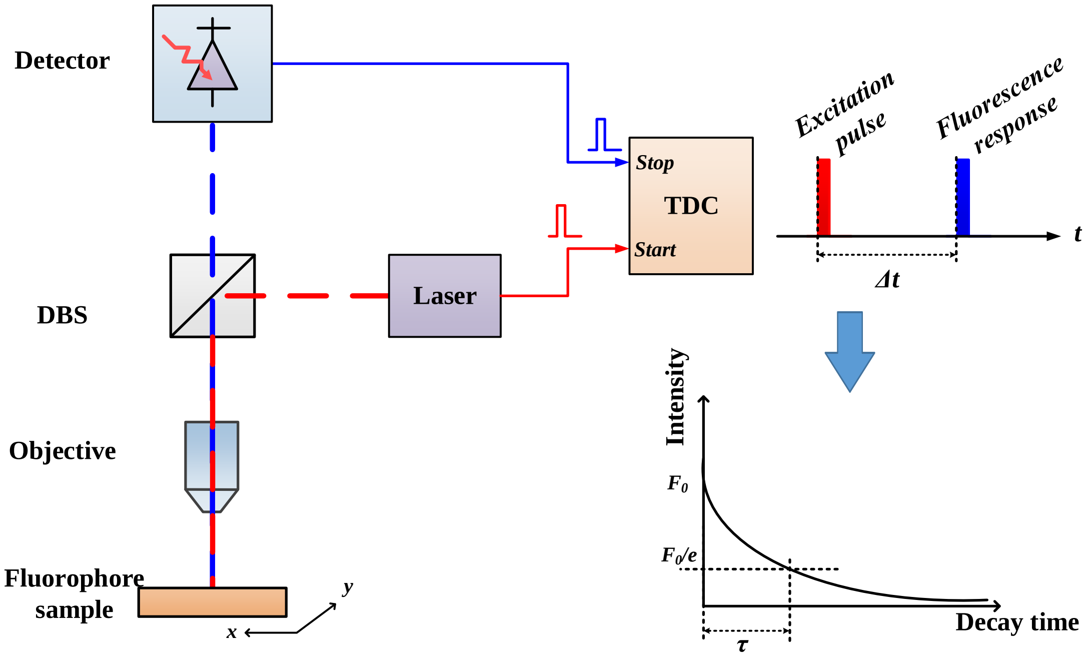


Figure 4-2, Diagram of a FLIM measurement system. A TDC used in a time-correlated single photon counting set up. The TDC is used to measure the fluorescence decay time with respect to the excitation pulse, which can be extracted from the histogram using exponential fitting.

to measure the shortest and longest lifetimes in the experiment, respectively. Typical lifetimes of fluorophores used in single molecule spectroscopy vary from a few nanoseconds (dyes) to hundreds of nanoseconds (quantum dots), even microseconds to milliseconds (lanthanides) [196]. However, the fluorescence lifetime is extremely sensitive to the environmental condition and is decreased by any non-radiative process. In this case, the TDC needs to be able to measure a shorter lifetime. Typically, a TDC suitable for FLIM application should have ~50ps resolution and 50ns dynamic range [180], [183]. In addition to the timing requirements, high conversion speed (or counting rate, frame rate) is also required. When the probability of detecting more than one photon per excitation cycle is non-negligible, distortions of the lifetime histogram will be introduced by the “pile-up” effect, which eventually degrades the SNR in FLIM [196].

For state-of-the-art SPAD, the dead time has been reduced down to nanoseconds [110], [149], [150], [173], [197]–[199], so the intrinsic dead time of TDC becomes the limiting factor for a high-speed acquisitions system [26], [49], [67]. To overcome this problem, an interleaved TDC-pair timing technique was proposed in [49], yielding real-time fluorescence lifetime estimations. Before diving into the detailed implementations of different TDCs, a good understanding of TDC performance metrics will help to understand, analyze, and compact various TDC designs.

4.1.2. Performance Metrics of TDC

There are several important parameters that are used to evaluate the performance of a TDC. Those performance metrics are similar to those in ADCs, and they apply to any TDC architecture.

(a) Resolution (LSB)

The resolution or least significant bit (LSB) determines the minimum input time interval that can be distinguished by a TDC. It is the step width in the quantization characteristics (input-output, I/O transfer curve) of a TDC, which means that there is a range of continuous time inputs being mapped to single digital word. Ideally, the step widths in the I/O transfer curve are constant. The corresponding output digital word increases by 1 per LSB input increment. Recently, TDCs with 1.76ps and 1.25ps resolution were reported in [200] and [91], respectively.

(b) Nonlinearity

Nonlinearity in a TDC will lead to deviations of its quantization characteristic from the ideal shape. Sources of nonlinearity include delay errors of the delay elements, signal crosstalk, layout mismatch and PVT variations. Basically, the nonlinearity performance is heavily dependent on the circuit architecture and layout implementation. Nonlinearity can be minimized by keeping the delay lines as short as possible [201].

The nonlinearity characteristics can be specified through two parameters, the differential nonlinearity (DNL) and the integral nonlinearity (INL). DNL is the difference between the actual and the ideal step widths in the I/O curve, caused by the propagation delay deviation from their desired values. INL is the integral of DNL along the delay chains up to the current position of calculation. When the INL is given as one number without code bin specification, it is the maximum absolute integral nonlinearity value of the entire measurement range [202]. Usually, DNL and INL are normalized to one LSB.

(c) Power and Area

Coupled to SPAD arrays, TDCs are capable of time stamping incident photons. In order to preserve more active region (i.e. higher fill-factor for photon detection and sensitivity), miniaturization of TDCs is essential [41], [182], [203]. Power is another important factor in array designs. Since high power consumption will generate more heat in the array, which results in more thermal noise in SPADs when the on-chip temperature increases. Thus, it is necessary to reduce the power consumption of the TDC-SPAD coupled imager. These two parameters should to be taken into consideration at an early design stage, because power- and area-saving methodologies are most efficient at the system level, rather than at the transistor level.

(d) Precision

Due to external and internal noise sources (e.g. the jitter from the reference clock and delay lines), the output of a TDC has a spread distribution around its expected value even when it has a constant input time interval. The spread of the TDC's output, generally but not always, follows a Gaussian distribution. The standard deviation (sigma) or the FWHM is then taken as the precision of the TDC. As the input to a TDC is kept fixed during the measurement, the precision is also referred to as the single-shot precision.

Precision indicates the reproducibility of the TDC operating with the presence of internal and external noise sources. Depending on the absolute value of the input intervals, the maximum precision occurs at the boundary of two adjacent steps in its quantization characteristic [68], [90].

Quantization error determines the basic level of the precision. The rms value of the precision, $\sigma_{TDC,rms}$, can be expressed as,

$$\sigma_{TDC,rms} = \sqrt{\sigma_q^2 + \sigma_{INL\_\mathrm{start}}^2 + \sigma_{INL\_stop}^2 + \sigma_{CLK}^2 + \sigma_{Additional}^2}\,, \tag{4-1}$$

where σ_q is the quantization error, σ_{INL_start} and σ_{INL_stop} are the standard deviations of the INL of the *Start* and *Stop* signals, respectively, σ_{CLK} is the jitter of the reference clock, and $\sigma_{Additional}$ is the jitter of signals within the TDC [62], [204]. In practice, off-chip influences such as I/O pads, bonding wiring and input channels crosstalk also limit the precision of a TDC.

(e) Dead Time and Counting Rate

The dead time (DT) of a TDC is the minimum time required to complete one conversion. The reciprocal of DT is the counting rate (CR). Describing how fast a TDC can sample and digitize time intervals, DT and CR are important in high-speed applications such as nuclear imaging instruments because they affect the quality of reconstructed images as well as the detection latency. A state-of-the-art TDC in [200] achieved 300 MSps counting rate.

Since TDCs have non-zero intrinsic dead times, an interleaved topology was proposed in [25], [49] to improve the overall sampling rate and conversion speed. Two TDCs were enabled in an interleaved manner, which meant one was idle and ready to accept new data while the other one was processing a conversion, writing data and resetting for the next conversion.

(f) Quantization Error

Quantization error or quantization noise, is defined as the distortion between the digitized output value and the original input value of a TDC. Usually, quantization errors are bounded from zero to one LSB, as a TDC's equivalent output cannot be larger than its input. Quantization error σ_q can be estimated with,

$$\sigma_q = \sqrt{\frac{LSB_{start}^2}{12} + \frac{LSB_{stop}^2}{12}} = \frac{LSB}{\sqrt{6}}, \tag{4-2}$$

where LSB_{start} and LSB_{stop} are the resolutions of *Start* and *Stop* signals, respectively, and *LSB* is the resolution of the TDC [62]. According to Equation (4-2), we can reduce the quantization error by having a higher resolution (smaller LSB) TDC. However, there might be penalties such as increased area, higher power consumption, deteriorated linearity and longer dead time.

(g) Dynamic Range

The dynamic range is the maximum time interval a TDC can measure before its output saturates. In ToF measurement, the dynamic range determines the maximum detection range. For example, in the Laser Radar described in [205], an excellent dynamic range of 985ns was obtained. In ToF PET, the dynamic range of the TDC

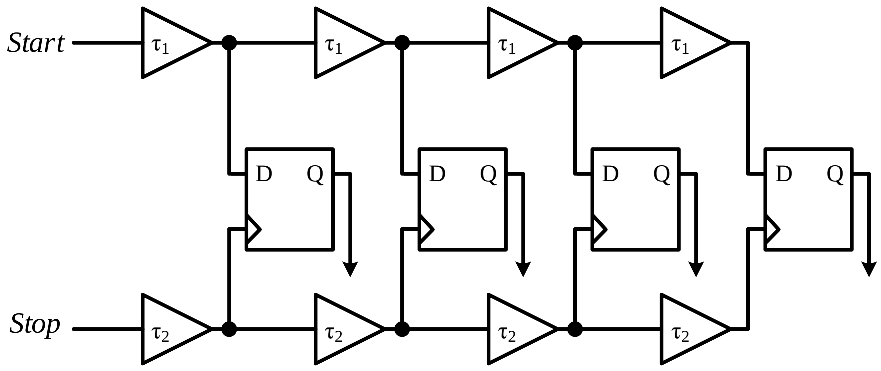


Figure 4-3, Conceptual diagram of the Vernier-delay-line based TDC

should be able to cover all annihilation points in the field-of-view [206]. This corresponds to at least 4ns dynamic range for a whole body ToF PET system with ~0.6m diameter detector ring.

4.1.3. Existing Digital TDC Architectures

One straightforward method to digitize the time interval is with a digital counter, which is only enabled from the arrival of *Start* signal until the arrival of *Stop* signal. However, an important limitation of the counter-based TDC is that its resolution is limited by the period of the reference clock, which turns out to be an impractically high frequency when a high resolution is desired.

One solution that provides sub-nanosecond resolution is with a DLL [66], in which the *Start* signal propagates along a chain of delay elements until the *Stop* signal arrives and triggers the D-flip-flops (DFFs) to record the status of the delay chains to represent the measurement result. However, the resolution of a DLL-based TDC is limited to the propagation delay of an inverter in the chosen technology.

To further improve the time resolution in a standard CMOS technology, sub-gate delays are needed. This is accomplished with a VDL based TDC that uses two chains of different delay elements [68], [90], as shown in Figure 4-3. Here, τ_1 is the delay of elements in one chain and τ_2 is the delay of elements in a second chain. In this case, the resolution is determined by the delay difference between the two chains, ($\tau_1 - \tau_2$). High resolution is obtained by setting τ_2 close to τ_1. Though the VDL-based TDC improves the resolution, the number of delay elements needed to cover a large dynamic range

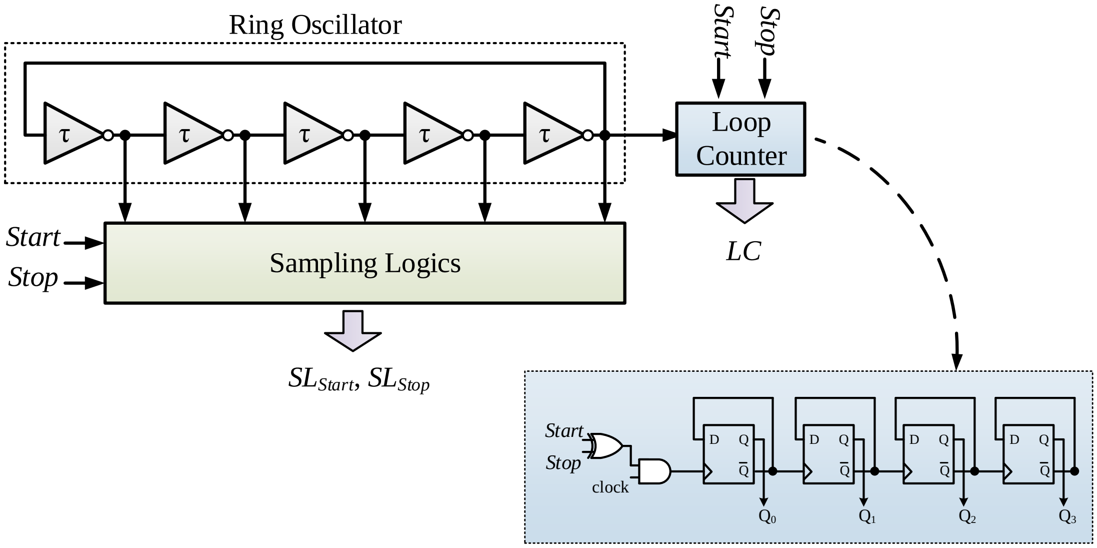


Figure 4-4, Conceptual diagram of the ring-oscillator based TDC.

will increase exponentially [207]. This, in turn, increases the latency time, area, and power consumption of VDL-based TDCs.

In order to simultaneously achieve high resolution and large dynamic range, the multiple interpolation technique [63], [207], [208] was adopted in TDC designs. For example, a coarse-fine TDC [189] amplified the time residue of the first stage, which was then digitized in the second stage, yielding a large interpolation factor. In [208], a two-stage interpolation TDC with a DLL followed by a VDL achieved 10ps resolution and 160ns dynamic range. However, in the multiple interpolation TDC, the full-scale range of the second stage should be equal to the resolution of the first stage. Otherwise, the mismatch will distort the linearity of the TDC [63], and may require extra calibration circuits with more area and power overhead.

Recently, RO-based TDCs, which simultaneously provide both high resolution and large dynamic range, have attracted much interest [209], [210]. Unlike the delay-line based TDCs in which pulses only propagate once per conversion, in RO-based TDCs [205], [209]–[211], the delay line is configured in a ring format such that pulses can propagate inside the delay ring iteratively until the conversion is completed. Figure 4-4 shows the concept of such a TDC. A free-running RO, consisting of an odd number of inverters, is used to generate multiple phases at a relative high frequency. The phase from the last stage is fed to a loop counter. This loop counter is enabled by the arrival of the *Start* signal and sequentially disabled by the arrival of the *Stop* signal. So the loop counter will record the number of clock cycles between the *Start* and *Stop* signals, which is the total iteration number in the RO during each conversion. A possible implementation of the loop counter is also included in Figure 4-4. The sampling logic, being triggered upon the arrival of both *Start* and *Stop* signals, records the status of each delay stage.

The measured time interval T_M, can be represented as,

$$T_M = (SL_{Stop} - SL_{Start}) \times \tau + \frac{CNT_{LC}}{f_{RO}}, \tag{4-3}$$

where SL_{Start}, SL_{Stop} are the outputs of the triggering logic at the arrival time of the *Start* and *Stop* signals respectively, τ is the propagation delay of the delay cell, CNT_{LC} is the output of the loop counter, and f_{RO} is the oscillation frequency. Theoretically, RO-based TDCs have unlimited dynamic range, but this cannot be accomplished in practice due to the available number of bits of the loop counter and limitations in the counting rate.

TDCs using RO topology usually have a time resolution limited by the propagation delay of the inverter. For example, the TDC in [210] implemented in 90nm CMOS technology had a resolution of 13.6ps. Another disadvantage of a basic RO-based TDC is its high power consumption because the RO is working in free-running mode. For example, the RO-based TDC in [205] consumed 24mW.

To obtain sub-gate resolution in a ring-oscillator based TDC, the multi-path technique can be used [207], [210]. Figure 4-5 shows the concept of the multi-path ring oscillator configuration. This technique is used to reduce the delay per stage. In such a multi-path inverter, the PMOS and NMOS transistors have intentionally asymmetrical connections.

For example, in the conceptual diagram of Figure 4-5, the gate terminals of the PMOS and NMOS are connected to different delay stage outputs, i.e. the 5th-preceding stage (P[m-5]) for the PMOS, and the 3rd-preceding and 1st-preceding stages (P[m-3], P[m-1]) for these two NMOSs. Thus, it allows an earlier arrival of the input transition to the slower PMOS transistor. Moreover, instead of being only tapped from its previous stage, multiple connections to the NMOS transistors are made, because now their speed, rather than that of the slow PMOS transistors, will limit the transition time

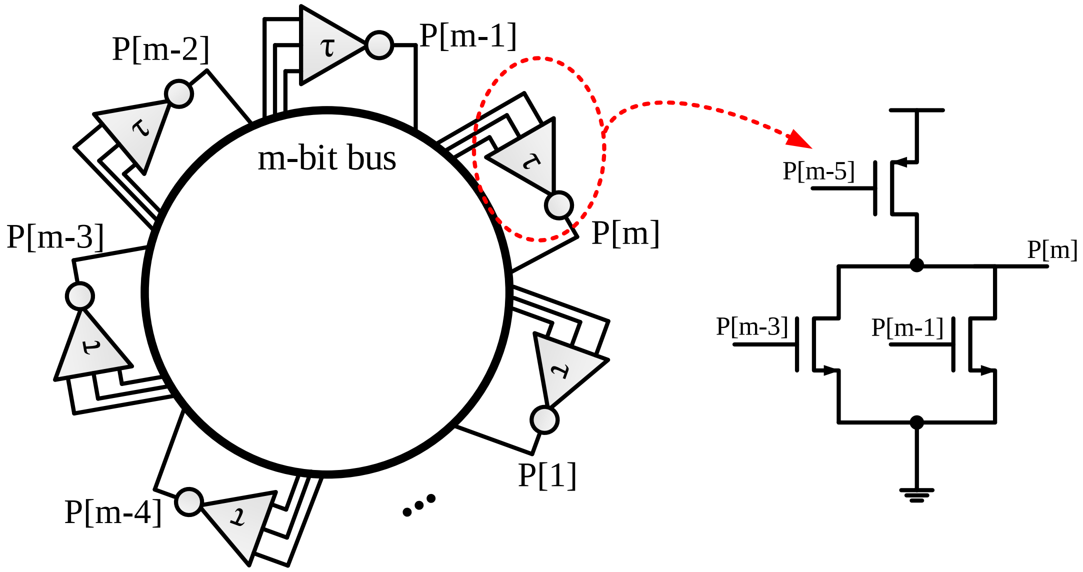


Figure 4-5, Diagram of multi-path oscillator configuration.

of the multi-path inverter. Using this leverage multi-path technique, the transition time of each delay unit can be greatly speeded up. This yields an improvement of the TDC resolution by the same speed-up factor. In the multi-path inverter design, the number of PMOS and NMOS used in the inverter, and the chosen of previous stages are arbitrary. For example, resolution enhancement factors of 5 [212] and 6 [207] were achieved using this approach. In a state-of-the-art example [207], two PMOS and three NMOS transistors (plus one additional NMOS/PMOS pair for gating purpose) were used in one inverter to realize this multi-path technique. It achieved 6ps resolution, 1ps precision with 50Msps conversion rate and 11 bits measurement range in a 0.13µm CMOS process. Another TDC employing this technique successfully reduced the delay per stage from 35ps to 6ps in a 0.13µm CMOS process, with three PMOS and three NMOS transistors in one inverter stage [212].

In order to reduce the power consumption of a RO-based TDC, the gated ring oscillator (GRO) TDC was proposed [207], [213], [214]. Several methods can be used to introduce the gating feature into the delay units. In one method, gating transistors are inserted at both the top and bottom of a conventional inverter to build the gated-inverter delay cell, such as in [207], [213]–[217]. When the gating transistors are closed, the current path in the cell is cut off. Thus, the gated-inverter is off and its output will be maintained by the capacitor at the output node. In another type, some specific logic

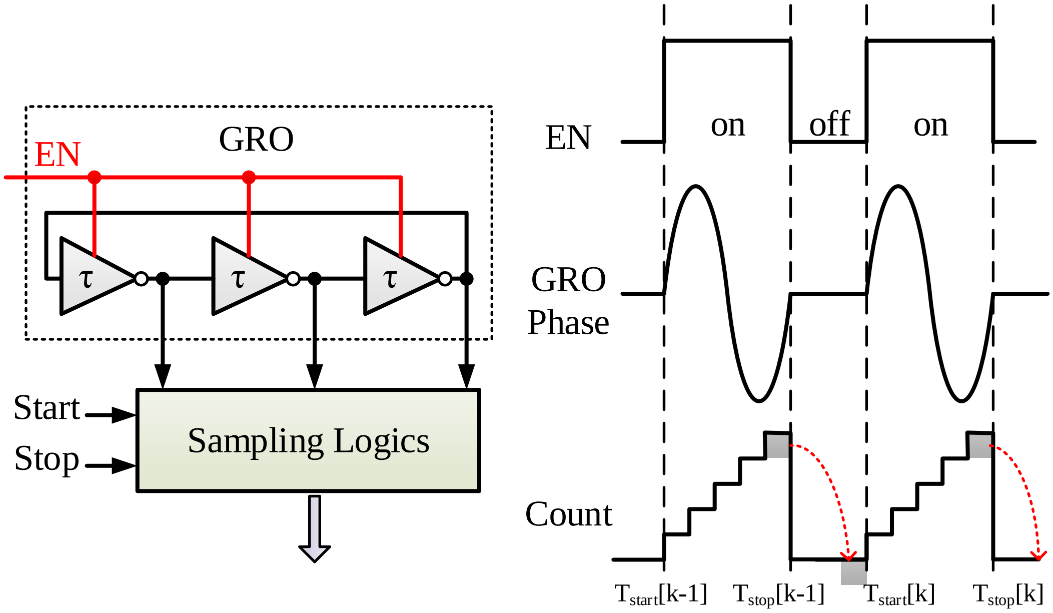


Figure 4-6, Concept of gated ring oscillator based TDC and its timing waveforms.

circuits are added into the RO. For example, in [218], the NAND gate, which had the exactly the same implementation as the inverter gate at the transistor level, was used to add the gating feature and match the propagation delay with other delay units. In another example [16], a pair of multiplexers was inserted between the delay units.

Figure 4-6 shows the principle of the GRO TDC with its timing waveform. An enable signal (EN) turns the RO on and off. When conversion is completed, the RO is off. The status at the output nodes of each inverter, presenting the conversion results, is frozen and kept by either parasitic or intentionally inserted capacitors at each output node in the RO. This is different from the regular RO-based TDC in which information at the output nodes will be reset after every conversion.

In addition to the benefit of power saving, GRO-based TDC introduces a first-order noise shaping mechanism. As mentioned before, noise degrades the resolution. Thus, quantization noise shaping is an alternative approach to obtain higher resolution. In the GRO TDC, a measurement starts from its previous measurement, which means that the residue occurring at the end of one conversion is transferred to the next one. The expression of measured time interval T_M can be written as,

$$T_M[\mathrm{k}] = T_{stop}[k] - T_{start}[k] \;= T_{stop}[k] - T_{stop}[k-1]. \tag{4-4}$$

This discrete-time, first-order difference operation gives rise to a first-order noise shaping in frequency domain. Also, the impact of delay cell mismatch is reduced because of the differential operation.

The remainder of this chapter is organized as follows. The principles of operation of the proposed gateable Vernier ring oscillator TDC are presented in Section 4.2. In Section 4.3, we describe the circuit implementation. Next, the measurement results are presented in Section 4.4. Finally, the conclusions are given in Section 4.5.

4.2. Operation Principles of the Proposed TDC

The goal of the work in this chapter is to develop a TDC suitable for on-chip integration with an array of SPADs targeted at ToF PET imaging. For this application, small footprint and low power consumption are key requirements. In addition, fine time resolution to improve the SNR of reconstructed images and a large dynamic range to cover the field-of-view in the standard whole body PET ring are required.

In order to meet these specifications, a gateable Vernier ring oscillator based architecture is employed and the proposed TDC has been realized in a standard 130nm digital CMOS process. The gateable operation feature of two ring oscillators and the single-transition, end-of-conversion detection arrays help in reducing the power consumption. With a 1.2V power supply, the proposed TDC only consumes 1mA current, which achieves superior state-of-the-art power consumption among all the ring-oscillator based TDCs published in literature.

Figure 4-7 illustrates the principle of the proposed gateable Vernier ring oscillator TDC. The *Start* and *Stop* pulses initiate the oscillation in the slow and fast ROs, respectively. t_M is the time interval to be measured. The delays of each stage in the slow and the fast ROs are τ_s and τ_f, respectively. Because two ROs are arranged in a Vernier configuration, the theoretical resolution of the TDC is (τ_s - τ_f).

In Figure 4-7, there are eight differential delay elements in each RO, so we can have 16 phase pairs, which are *PS*(*1*) to *PS*(*16*) and *PF*(*1*) to *PF*(*16*). The *PF*(*i*) triggers the sampling operation of its corresponding *PS*(*i*), where *i*=1,2,3...16. Two ROs generate phases iteratively, and the time interval between a coincident phase pair will reduce by one LSB after every delay element in the RO during the time-to-digital conversion,

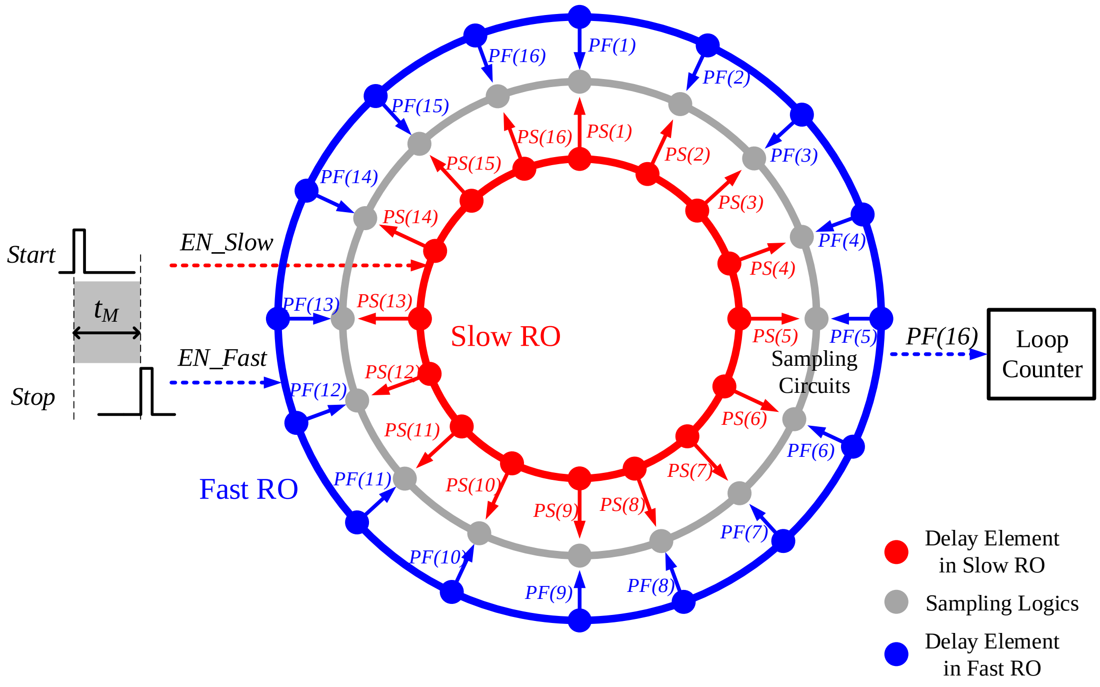


Figure 4-7, Diagram of the operating principles of the proposed gateable Vernier ring oscillator TDC

until a certain *PF*(*i*) arrives prior to its counterpart *PS*(*i*). Then, the conversion is completed and the ROs will be turned off. In order to cover a large dynamic range with the gateable Vernier ring oscillator TDC, a loop counter, being triggered by the phase *PF*(*16*), is implemented to record the circulation number of the fast RO during the conversion.

The timing diagram of the proposed TDC is shown in Figure 4-8. In this example *PF*(*5*) comes prior to *PS*(*5*) in the *N*-th loop, so the conversion is finished there. The measurement result $t_{Measure}$ can be expressed as follows.

$$t_{Measure} = 5 \times LSB + (N-1) \times 16 \times LSB. \tag{4-5}$$

where *LSB* is the resolution of the TDC, which is equal to (τ_s - τ_f).

Generally, assuming that the output from the sampling circuits is OUT_{SC}, and the output from the loop counter is CNT_{LC}, then the measured time can be expressed as

$$t_{Measure} = OUT_{SC} \times LSB + CNT_{LC} \times 16 \times LSB. \tag{4-6}$$

The output of the sampling circuits is recorded and updated at the rising edge of every phase clock in the fast RO, i.e. *PF*(*1*) to *PF*(*16*). The output of the loop counter has an increment every period of *PF*(*16*). Thus, the TDC's output data is valid right

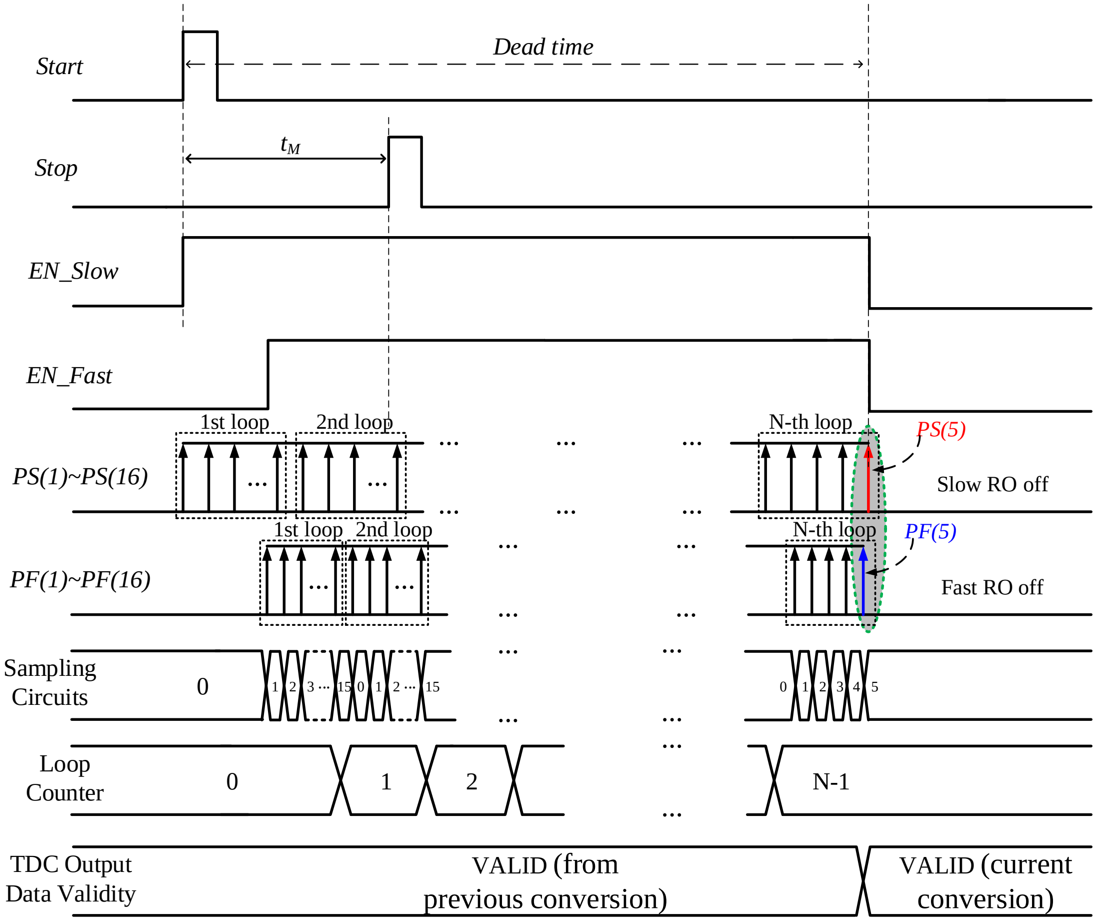


Figure 4-8, Timing diagram of the proposed gateable Vernier ring oscillator TDC.

after the current conversion is completed. Since this prototype does not include a counter to record the circulation number in the slow RO before the *Stop* signal arrives, then its dynamic range is limited by the oscillation period of the slow RO.

The proposed TDC uses differential delay elements and a Vernier configuration of two ROs, so it has an intrinsic first-order immunity towards common-mode noise. This advantage is important in a real ToF PET application since there will be many TDCs in the digital silicon photomultiplier.

4.3. Circuit Implementation

The block diagram of the proposed TDC is shown in Figure 4-9. In this diagram, the enable generation block (**EN_GEN**) generates the enable signals, *EN_Slow* and *EN_Fast*, are to initialize oscillation in slow RO and fast RO, respectively, where the phase pairs are then available. These two enable signals are designed as high-level

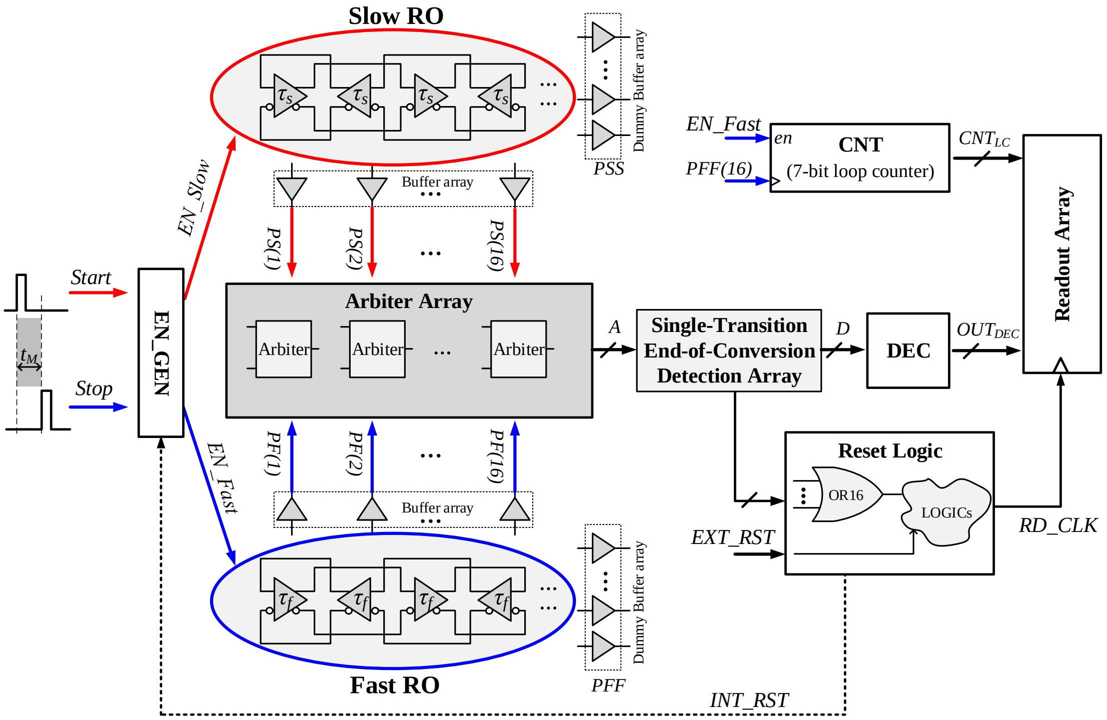


Figure 4-9, Block diagram of the proposed gateable Vernier ring oscillator TDC.

active toggle signals, rather than short pulses. They will be pulled down to ground (zero) when the conversion is finished (i.e., being reset by the *INT_RST* signal from the reset logic). The arbiter array judges which phase, either *PS*(*i*) or *PF*(*i*), comes first. A single-transition end-of-conversion detection array is used to determine the end of conversion. The one-hot to binary decoder (**DEC**) converts the 16-bit one-hot code to a 4-bit binary code. The 7-bit loop counter (**CNT**) is used to cover a large dynamic range. Once the conversion is completed, the reset logic will turn off both ROs. Meanwhile, a flag signal (*RD_CLK*) which has a short pulse duration (about 1ns) is pulled high to trigger the readout array of registers to record all results. This flag signal is also connected to an output pad and used as the indicator of the time conversion completion in the measurement.

Below, we describe in detail the five important circuit blocks – gateable ring oscillator, arbiter, single-transition end-of-conversion detection array, loop counter and the enable generation block.

4.3.1. Gateable Ring Oscillator

To build a ring oscillator, we can either use an odd number of regular inverters, or

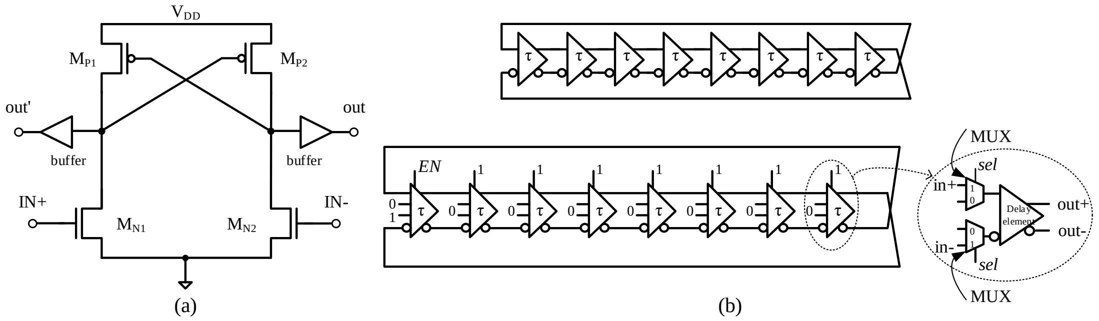


Figure 4-10, (a) Transistor implementation of the differential delay element; (b) Conventional ring oscillator (top) and gateable ring oscillators with regular connections (bottom)

use an even number of differential inverters plus a cross connection. In this prototype, each ring oscillator employs 8 differential delay elements to generate 16 phases. The transistor implementation of the differential delay element is given in Figure 4-10(a). By sizing the transistor (M_{P1} and M_{P2}) in the slow and fast ROs, we can get slightly different propagation delays of the elements, i.e. τ_s and τ_f, respectively, to obtain a fine resolution.

The conventional ring oscillator and gateable ring oscillator, both consisting of 8 stage differential delay elements, are illustrated in Figure 4-10(b). Compared to the conventional free-running ring oscillator, the gateable RO can be gated on and off by the enable signal, *EN*. In this way, as necessary, we can reduce the power consumption. To configure the RO in the gateable mode, important design adjustments of the ring oscillator are needed. We can either introduce enable transistors into the delay element [207], [213], [214], [215], or we can use gating logic [218].

In this work, as depicted in Figure 4-10(b) (bottom), a pair of multiplexers (MUXs) is inserted between the delay elements. The selection bits of all MUXs, except for the first pair, are connected to V_{DD}, while that of the first MUX pair is connected to *EN*, the enable signal. When *EN* equals 0, the inputs 0 and 1 are assigned to the non-inverting and inverting ports of the first delay element, respectively. This means the ring oscillator is gated off and is reset. Otherwise, all the delay elements are connected serially and configure a regular ring oscillator.

In this gateable RO architecture, there are severe mismatches in the connection of the oscillators in Figure 4-10(b). The path between the eighth and the first delay elements is much longer than the rest of the paths. This longer path has more parasitic

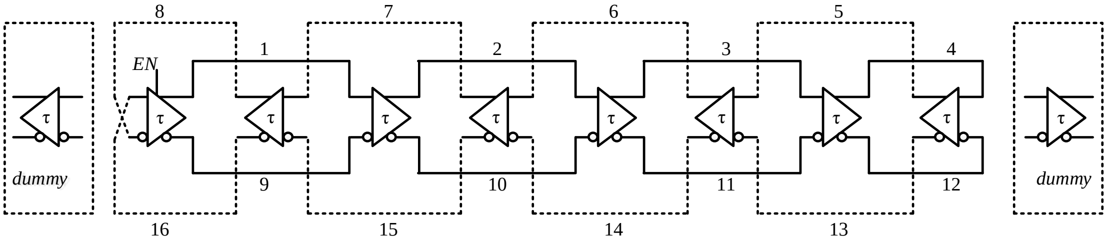


Figure 4-11, Twisted-manner arrangement of the RO.

resistance and capacitance, so there is a load mismatch which introduces non-uniformity in the propagation delays. The different propagation delays eventually contribute to the nonlinearity of the proposed TDC.

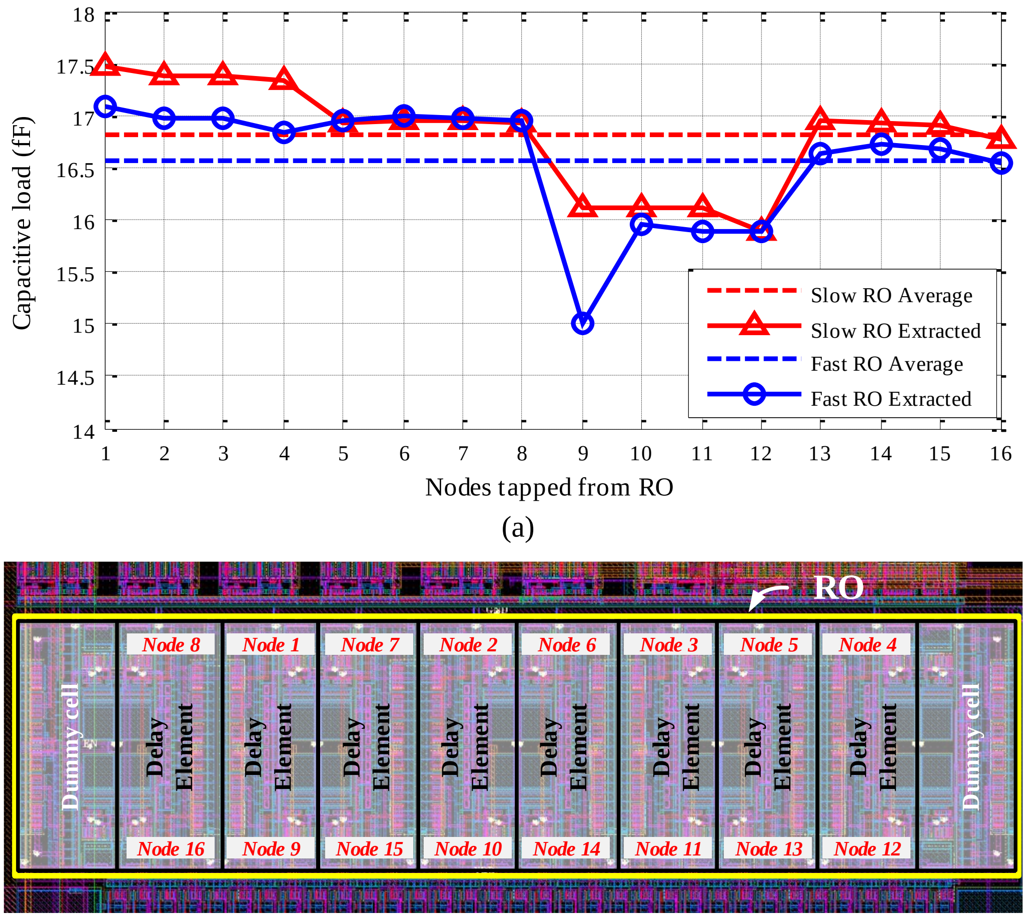
To solve the non-uniform propagation delay problem, as shown in Figure 4-11, a twisted arrangement of the delay cells, adopted from [209], is employed. Special care of the routing in the RO is taken to attain equal lengths between delay elements. The enable signals and the other two inputs are not drawn here to avoid cluttering the schematic. Two dummy cells are employed at both terminals of the RO to match the parasitic loads of the first and the last delay elements. This approach, with the twisted connection and dummy cells, equalizes the loads of each delay elements, resulting in the same propagation delay.

To verify this point, the extracted capacitive loads at every node in the two ROs are shown in Figure 4-12(a). The dips of the extracted capacitive loads around the ninth node in both ROs, are thought to be affected by the imperfect matched parasitic loads in the vertical direction. As shown in the layout of the RO in Figure 4-12(b), we can see that the RO is well matched along the horizontal direction. However, in the vertical direction, the parasitic loads are not the same due to different surrounding circuits. Despite that, the variation of the extracted capacitive load is only 2fF.

4.3.2. Arbiter

In the TDC, a time comparator is needed to compare two input pulses, i.e. *PS*(*i*) and *PF*(*i*), and output a digital value indicating which pulse comes first. The output of the time comparator can be expressed as follows.

$$Out_{Time\_Comparator} = \begin{cases} 1, if\ \ PS(i)\ comes\ prior\ to\ PF(i) \\ 0, if\ \ PF(i)\ comes\ prior\ to\ PS(i) \end{cases}. \tag{4-7}$$



(a)

(b)

Figure 4-12, (a) Post-layout extracted capacitive load uniformity of the ROs; and (b) Layout of the RO and its surrounding circuits. Here the fast RO is given as an example. The two ROs have similar layouts.

Either a DFF [210], or an arbiter [90], [207], [213], [215] can be used to perform this function.

The most important parameter of a time comparator is the offset time, analogous to the voltage offset in the voltage comparator. The offset time is due to the difference in the propagation delays on different paths, which mainly arises from unbalanced loads. A regular DFF has different loads for its clock and data paths, resulting in a large offset time.

In high-resolution TDCs, this offset time becomes comparable to or even larger than one LSB. One can use the fully symmetrical DFF [85], [219]. However, it uses many transistors, thus requires a relatively large area for its implementation. The other option, an arbiter, is suitable for this time interval comparison because the two input paths of

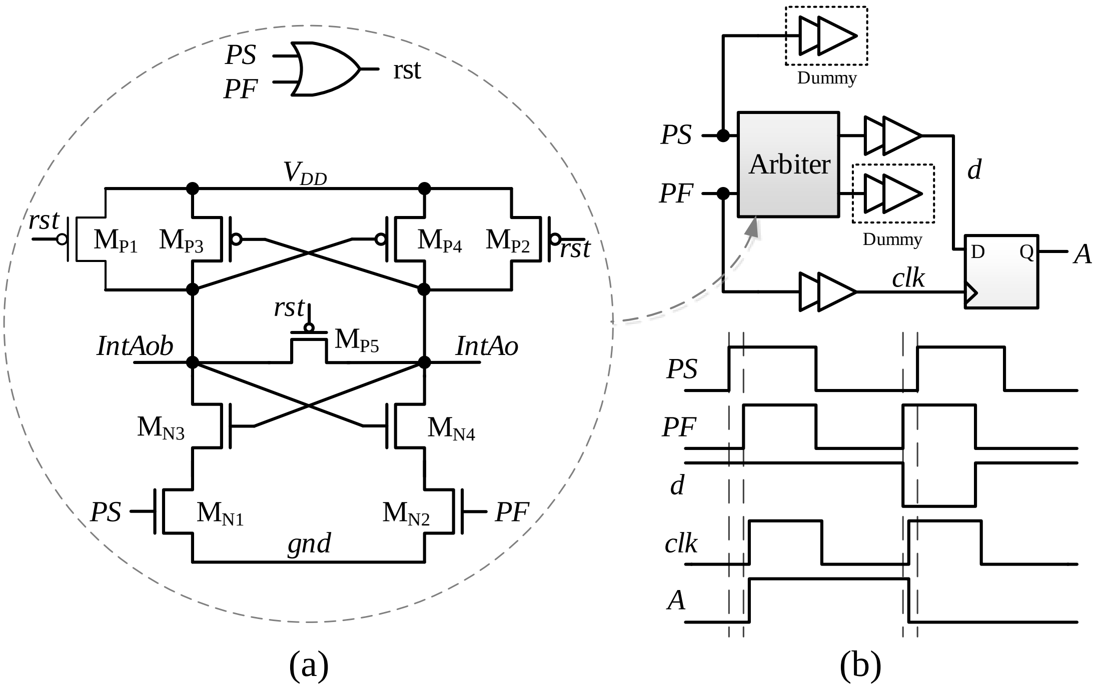


Figure 4-13, Circuit of (a) The arbiter; and (b) The sampling circuit with its timing diagram.

an arbiter can be well matched and it requires less area than the DFF. Therefore, in our TDC design, we used the arbiter as the time comparator to sample the status in the gateable Vernier ROs.

As shown in Figure 4-13(a), the arbiter has identical paths between the *PS* and *PF* inputs. Figure 4-13(b) shows the sampling circuit that utilizes an arbiter and a DFF. The timing diagram is also illustrated in the figure. Whenever both *PS* and *PF* are low, the internal reset signal (*rst*) will turn on M_{P1} and M_{P2} to reset the output of the arbiter. As long as the *PS* phase is leading the *PF* phase, the arbiter output (*IntAo* and its buffered version *d* in Figure 4-13(b)) remains high. Once the *PF* phase arrives prior to the *PS* phase, the output of the arbiter, *IntAo* and *d*, will be pulled Low to indicate the sampling result. Since the arbiter itself is voltage-level sensitive, a regular DFF is used to latch the sampling result at every period of *PF*. Note that dummy cells are inserted to balance the loads of the two input phases (*PS* and *PF*) and the two output ports of the arbiter.

4.3.3. Single-Transition End-of-Conversion Detection Array

From the 16-bit output of the arbiter array, the detection array to determine the end

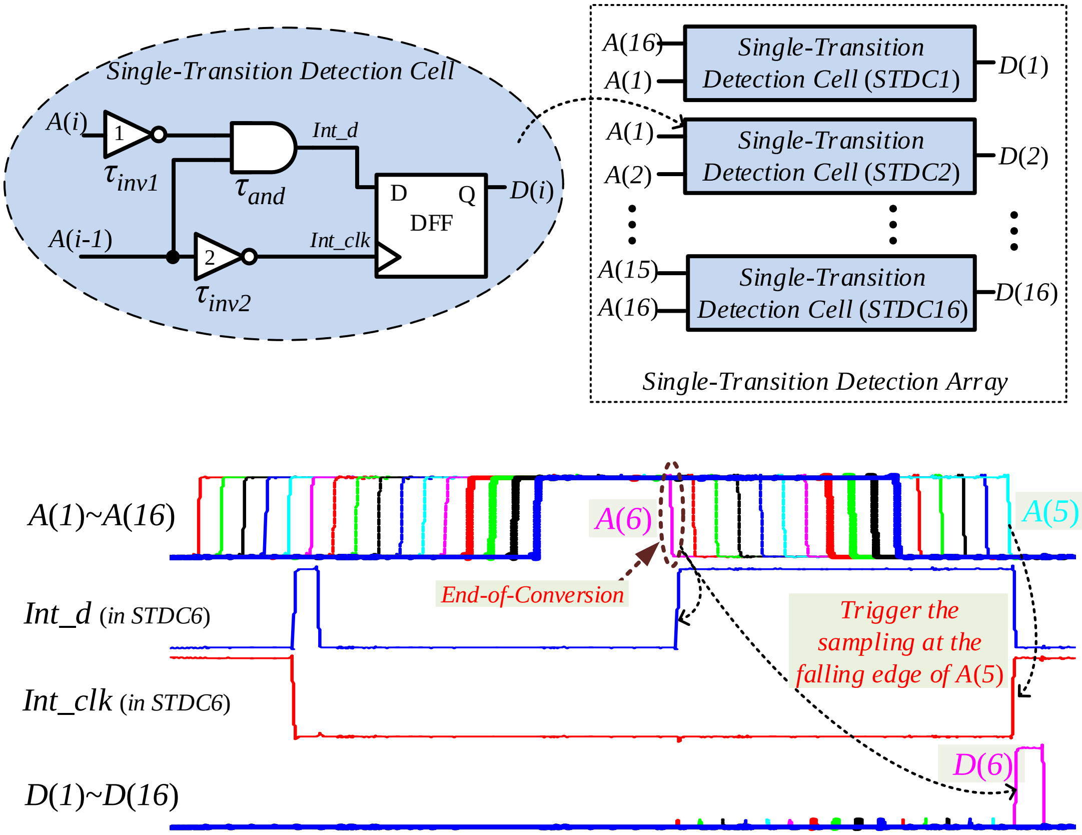

Single-Transition Detection Cell
A(i)
1
τinv1
Int_d
τand
D
Q
DFF
D(i)
A(i-1)
2
τinv2
Int_clk
A(16)
A(1)
Single-Transition Detection Cell (STDC1)
D(1)
A(1)
A(2)
Single-Transition Detection Cell (STDC2)
D(2)
A(15)
A(16)
Single-Transition Detection Cell (STDC16)
D(16)
Single-Transition Detection Array
A(1)~A(16)
A(6)
A(5)
Int_d (in STDC6)
End-of-Conversion
Trigger the sampling at the falling edge of A(5)
Int_clk (in STDC6)
D(6)
D(1)~D(16)

Figure 4-14, Diagram of the single-transition detection circuit with a simulated timing diagram.

of conversion is implemented. Usually this detection module is power hungry as it is toggled at the same frequency of the ring oscillator. In order to reduce the power dissipation of this module, we proposed a detection circuit which only has one transition per conversion.

Figure 4-14 shows the diagram of the single-transition detection circuit. This DFF will only sample the data at the falling edge of the *A*(*i-1*) signal. The output *D*(*i*) will be “1” only when simultaneously *A*(*i*) is “0” and *A*(*i-1*) is “1”. As shown in the simulated timing diagram, a change of DFF’s input signal (*Int_d*) is caused by *A*(*i-1*) and the sampling operation of DFF is also triggered by the falling edge of *A*(*i-1*), i.e. *Int_clk*. The setup timing constraint of the DFF should be satisfied to avoid a potential timing violation. Thus, the delay of the AND gate τ_{and} is designed to be longer than that of the inverter at the bottom τ_{inv2}. With these delay considerations, the DFF can record the previous output *Int_d* without timing violation. In this particular timing example given in Figure 4-14, the conversion is completed at the sixth detection cell as

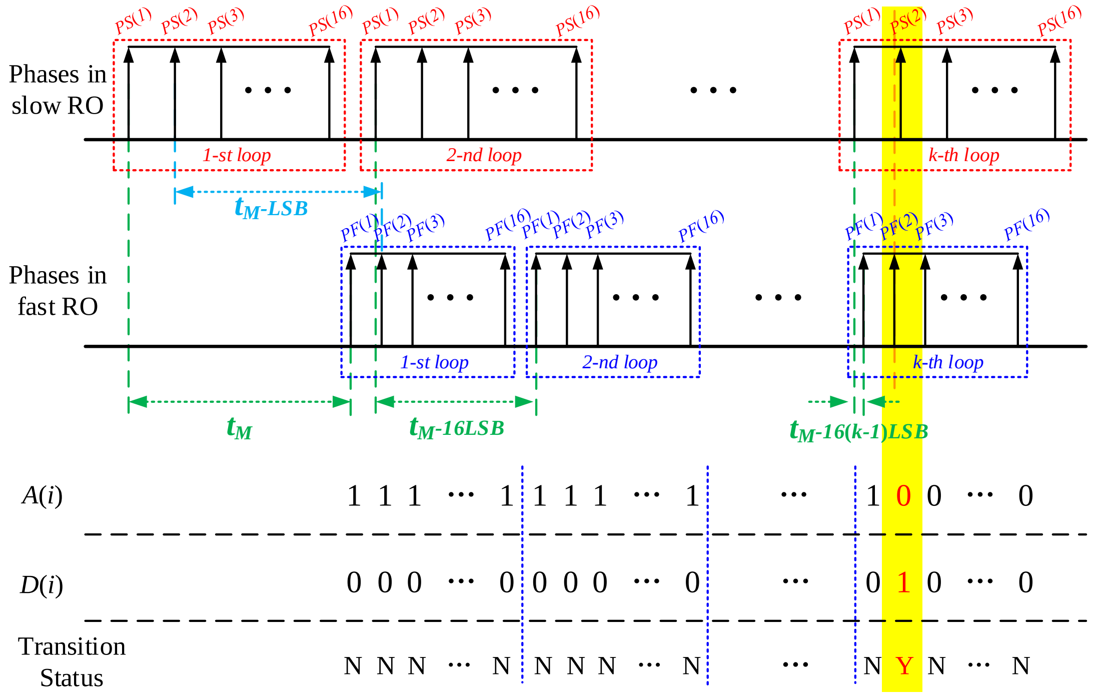


Figure 4-15, Illustration of the timing diagram of the single-transition end-of-conversion detection array. In the "Sampling Status" row, "N" means there is no sampling and "Y" means there is a sampling operation in the detection array.

the output from arbiter *A*(*6*) is pulled down first. At the following falling edge of *A*(*5*), the end-of-conversion result is sampled and latched by the DFF. Consequently, only *D*(*6*) is pulled high and one transition occurs during the conversion period.

A timing diagram with corresponding outputs from the arbiter array and the gated-detection array is illustrated in Figure 4-15. The time interval between the *PS* and *PF* phase pair shrinks by one LSB per stage and sixteen LSBs per loop during the conversion. In this case, *PF*(*2*) comes prior to *PS*(*2*) in the *k-th* loop, thus the conversion is completed there. The output of the gated-detection array, *D*(*i*), only transits from 0 to 1 at the ending point of the conversion. The last row in Figure 4-15 indicates the sampling status of the detection module. It shows that the gated-detection array only transit one time at the end of the conversion.

4.3.4. Loop Counter

To record the number of whole loops in the fast RO, a counter is needed. Using a synchronized counter architecture, the loop counter is enabled by *EN_Fast* and trigged by *PFF*(*16*), as shown in Figure 4-7. In our prototype, we implemented a 7-bit loop

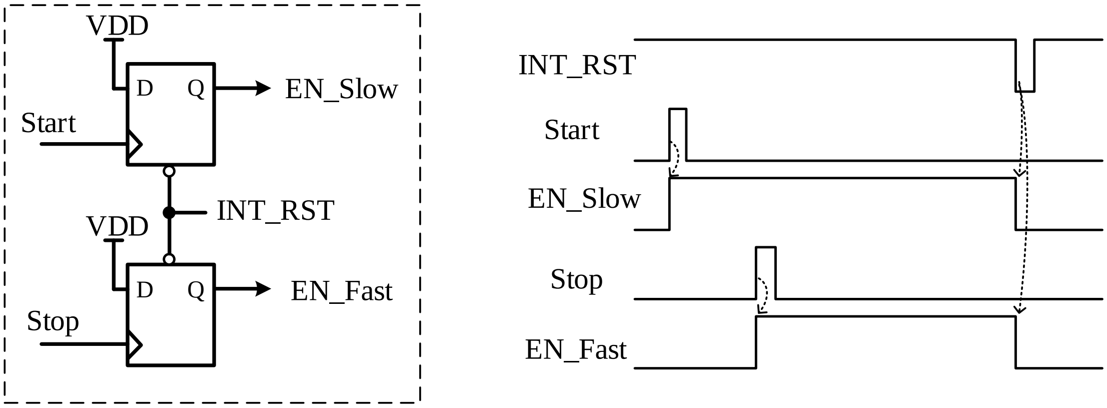
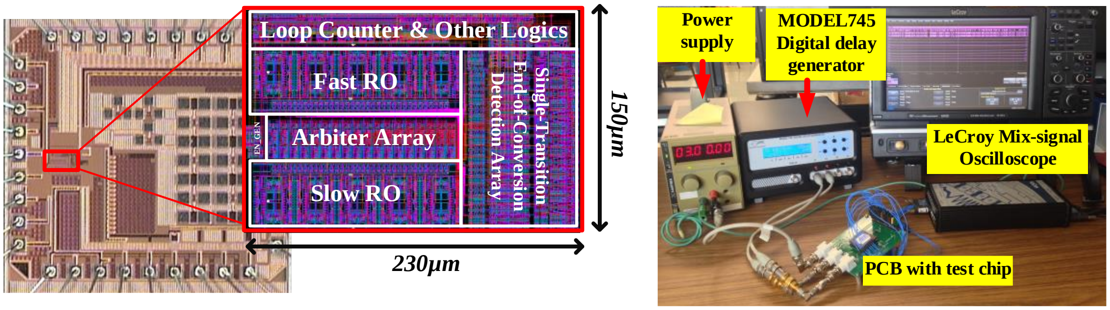


Figure 4-16, Circuit and timing diagram of the enable generation block (EN_GEN)


Loop Counter & Other Logics
Fast RO
Arbiter Array
Slow RO
EN_GEN
Single-Transition End-of-Conversion Detection Array
150μm
230μm
Power supply
MODEL745 Digital delay generator
LeCroy Mix-signal Oscilloscope
PCB with test chip

(a)

(b)

Figure 4-17, (a) Photomicrograph of the test chip (2mm×2mm). The proposed TDC occupies 230μm×150μm area. The major part of the test chip is test structures. (b) Measurement set up.

counter. Also, by using more bits which consumes slightly more area, the proposed TDC can cover a larger dynamic range.

4.3.5. Enable Generation Block

To control the fast and slow ring oscillators, two enable signals are generated from the input Start and Stop pulses. The circuit schematic with its associated timing diagram is shown in Figure 4-16. The flip-flop (with data input pulled up to supply voltage) is used to generate the high-active toggle signal (EN_Slow and EN_Fast). They will be reset when the conversion is finished, which is indicated by the arrival of the internal reset signal (INT_RST).

4.4. Measurement and Results

The proposed gateable Vernier RO TDC has been fabricated in IBM's 130nm digital CMOS process. Figure 4-17(a) shows a photomicrograph of the test chip. The TDC's

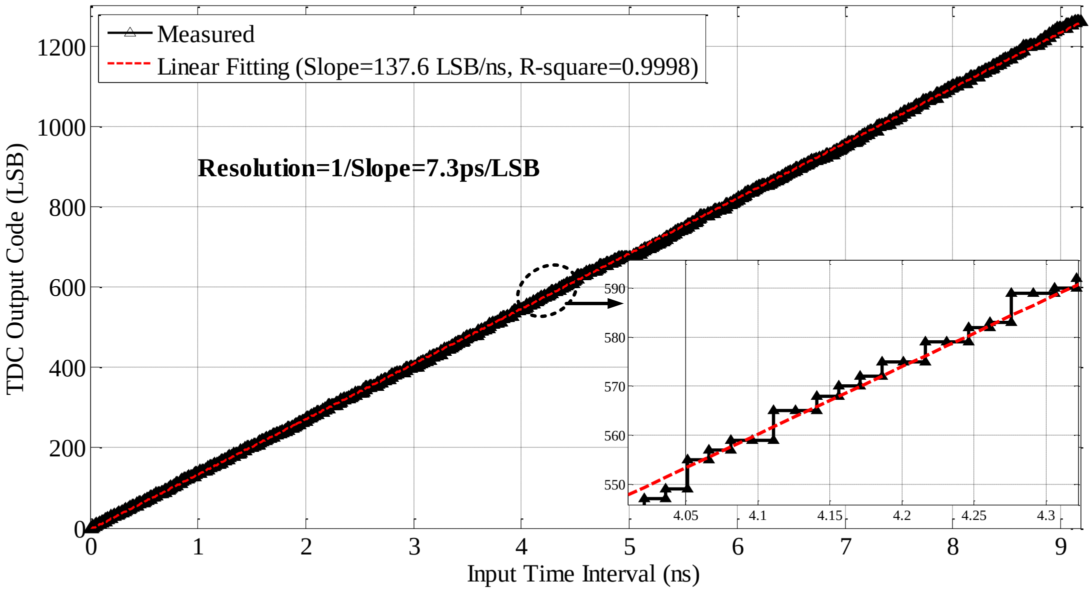


Figure 4-18, Measured input-output characteristic of the TDC.

area is 230μm×150μm (0.03mm^2). When operating at 1 MHz counting rate, its average power consumption is 1.2mW with a 1.2V supply.

To characterize the test chip, a digital delay generator, the Berkeley Nucleonics Corporation MODEL745, is used to generate the adjustable time interval between the two input channels. The TDC outputs are collected by a mixed-signal oscilloscope, the LeCroy 625Zi. Figure 4-17(b) shows the measurement set up in the laboratory. Before the measurement, we measured the jitter performance of the delay generator over the TDC's dynamic range. It has an rms jitter of 17.4ps.

4.4.1. Input-Output Characteristic

In order to measure the I/O characteristic of the TDC, the interval t_M is swept from zero to 9ns in increments of 15ps. Each interval point is measured 5,000 times to minimize statistical error. The measured I/O curve of the TDC is shown in Figure 4-18. The red dashed line gives the best linear fitting curve with an R-square value of 0.9998. The mean value of the measured step widths in the I/O transfer curve is the LSB resolution. The effective resolution can also be presented by the reciprocal of the linear fitting curve. According to the data in Figure 4-18, this prototype TDC yields an effective resolution of 7.3ps.

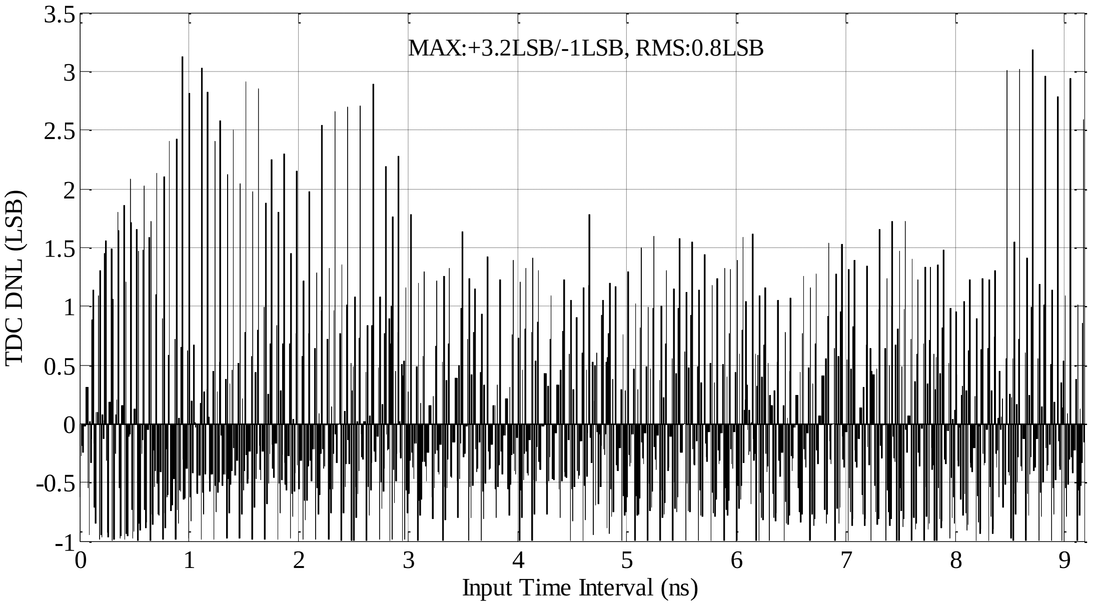


Figure 4-19, DNL of the TDC.

4.4.2. Nonlinearities

DNL of TDC is the deviation of the step widths in the I/O curve from their ideal value and INL is the accumulation of DNLs along the delay chains. The nonlinearities arise from the delay errors of delay elements, signal cross-talk, layout mismatches and PVT variations. The DNL and INL are characterized using the typical code density test [45], [58], [63], [97], [220], [221]. Uniformly distributed input intervals in the entire dynamic range are sent to the TDC. 2,700,000 measurements are cumulated and the hits distribution is collected. When a bin (time slot) gets more hits than the average value, it has a wider step in the I/O transfer curve, a positive DNL occurs, and vice versa. The measured DNL is shown in Figure 4-19. Our TDC has 3.2LSB (maximum) or 0.8LSB (rms value) DNL. Figure 4-20 gives the measured INL performance of the TDC before calibration (in solid black line). The major contributor to this moderately large nonlinearity is the delay error of the delay elements in the ring oscillators. This is because in our TDC, the propagation delays of the delay elements are not locked by the DLL, thus they are subject to PVT variations and mismatches.

Since INL error is caused by the non-uniformity of delay cells' propagation latencies, it can be considered as a systematic characterization of the TDC's nonlinearity performance, we can correct further measurements provided the INL information is

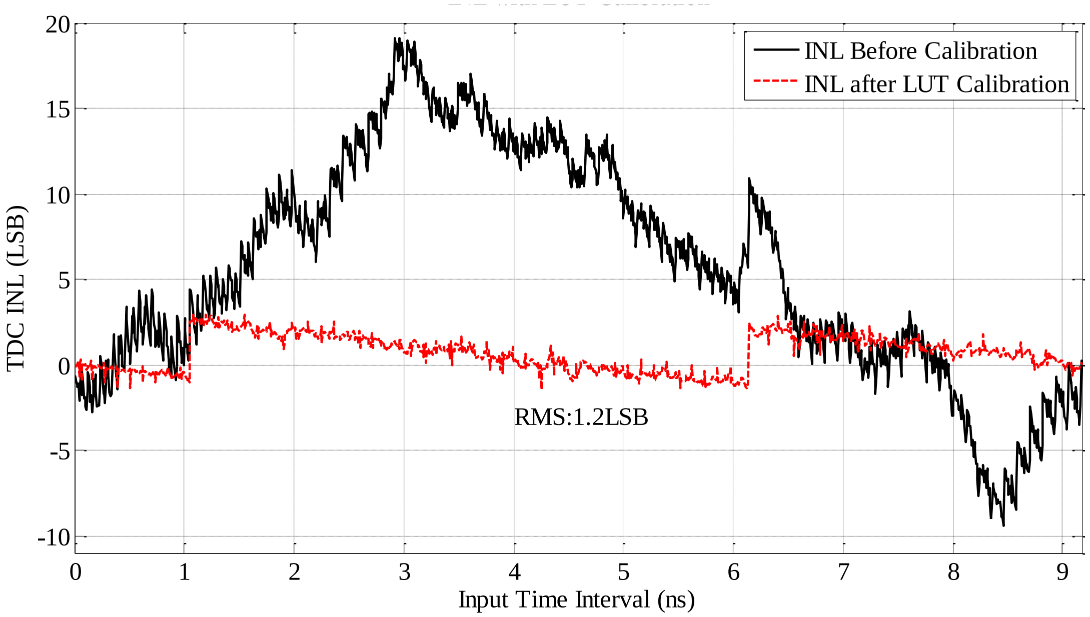


Figure 4-20, INL of the TDC before and after the LUT correction.

known and stored in a look-up-table (LUT) [68],[69]. The INL LUT calibration is exploited in Figure 4-20, with the red dashed line presenting the INL after the LUT correction. With the help of an INL LUT, the INL of the prototype TDC is improved to 1.2LSB rms. Note that the INL LUT needs to be determined for each fabricated test chip as this characteristic is unique for each chip.

4.4.3. Single-shot Precision

The single-shot precision, also called precision, is the standard deviation of the distribution of measurement results around the mean value when a constant time interval is repeatedly measured a large number of times. The distribution is caused by ring oscillator jitter, jitter of the input signals, quantization noise and power supply fluctuations. One channel of the digital delay generator, a T-splitter and cables of different lengths are employed for the single-shot measurement setup, which eliminates the jitters between different channels from the generator. Several time intervals are generated with cables of different lengths, and each measurement is collected 1,000,000 times. The histograms of the measurements with their accompanying statistics are shown in Figure 4-21.

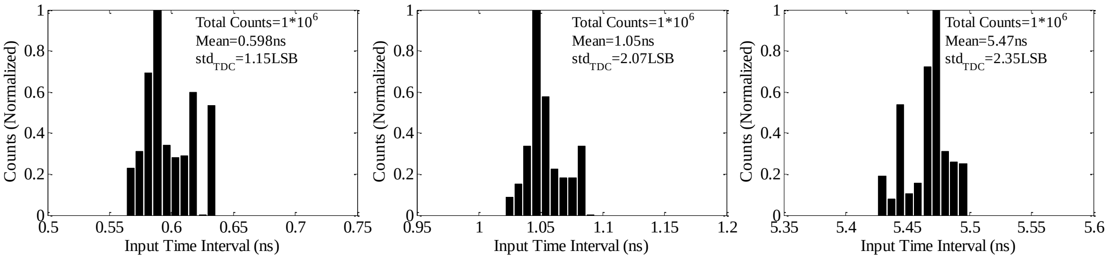


Figure 4-21, Three examples of precision measurement histograms at different positions in the dynamic range.

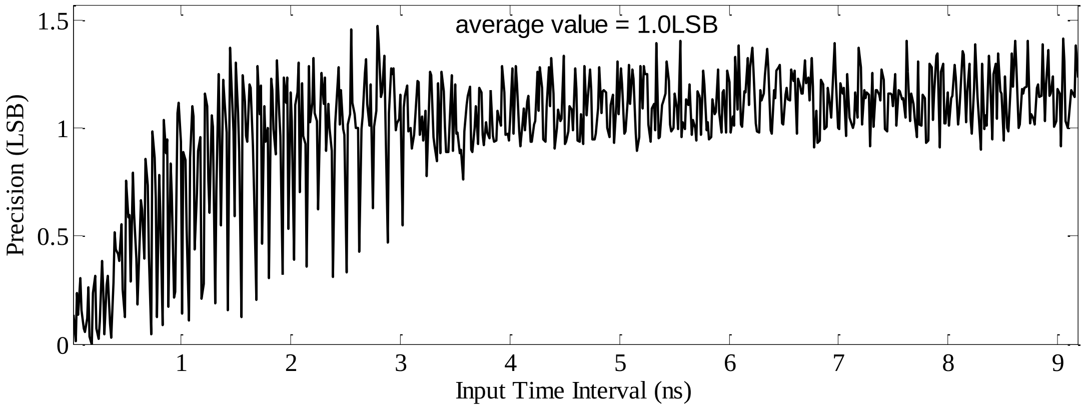


Figure 4-22, Precision of the TDC over the entire dynamic range.

To further study the precision performance of the TDC over its full-scale range, the input time interval is generated using the digital delay generator between two separate channels. The single-shot test is taken over the whole range of 9ns by 15ps increments. Each point is measured 5,000 times. Jitter from the delay generator was corrected from the measurement results using the deconvolution process. The precision measurement result is shown in Figure 4-22. The average value over the whole dynamic range is 1.0LSB, i.e. 7.3ps.

The dynamic range of this prototype is 9ns. As the dynamic range is limited only by the number of bits of the loop counter, it can be easily extended by adding more bits to the loop counter. The dead time of the TDC, defined as the latency time required to complete one time-to-digital conversion during which period the TDC is non-responsive to new inputs, is about 415ns. Thus, when measuring the time interval up to its full-scale range, the maximum sampling rate of the TDC is about 2.4MHz.

The Table 4-1, a summary of the proposed TDC’s performance and a comparison with the state-of-the-art results in the literature is provided. For a consistent comparison, Table 4-1, Performance summary and comparison with the state-of-the-art TDC designs.

	THIS WORK	[218]	[90]	[210]	[213]	[215]	[64]
Architecture	**GVRO**	VGRO	2D VGRO	RO	VGRO	2D VGRO	Switched RO
Technology (nm)	**130**	130	65	90	90	90	90
Supply (V)	**1.2**	1.5	1.2	1.2	1.2	1.2	1
LSB (ps)	**7.3**	8	4.8	13.6	5.8	15	0.315
Dynamic range (ns)	**9**	32	0.6	111	40	40	2-840
Power (mW)	**1.2**	7.5 [a]	1.65	18	4.5	2.1	1.5
Area (mm^2)	**0.03**	0.26	0.02	0.02	0.03	0.04	0.02

[a] Chip-wide power consumption.
[b] Integrated within ADPLL.

the results from publications listed in this table are all based on the ring oscillator architecture. Our proposed TDC features high resolution, good precision, low power consumption and compact size. Compared with previously reported TDCs, our work achieves the lowest power consumption, and uses one of the smallest areas. These features allow the proposed TDC to be integrated with SPAD arrays, to realize single-photon biomedical imaging sensors.

With such compact size and very low power consumption, a dSiPM sensor comprising SPAD arrays and the proposed TDC has great potential to be deployed for both ToF PET and FLIM applications. First, it can significantly reduce the total power consumption as a clinical PET scanner requires up to several thousand dSiPM arrays. Second, due to the superior TDC resolution, the system's time resolution would be largely limited by the other two sources (i.e. the scintillation photon decay and the noise in detectors). Third, due to the relatively low signal rate (~1-2 Mcps) in a clinical PET scanner as limited by both the solid angle coverage and injected radiotracers, the counting rate of the proposed TDC is not expected to suffer from any dead time effect.

4.5. Conclusions

In this chapter, we have successfully designed, fabricated and tested a ring-oscillator based TDC using the gateable Vernier RO architecture. This TDC consumes only 1mA current from a 1.2V supply, achieved the lowest power consumption to the date of writing of this thesis. A key feature of the proposed TDC is that its power hungry modules, the ring oscillators and the end-of-conversion detection circuit, are operating in power-efficient modes. The former only oscillates during the conversion, and the latter module only has one transition per conversion period. Thus, the number of transitions is significantly reduced, resulting in very low power consumption. Since our future work involves the integration of an SPAD array and TDCs on the same silicon substrate to build the dSiPM for biomedical imaging applications, then a standard 130nm CMOS process was chosen to implement the TDC prototype. The 130nm CMOS choice is because the performance of the SPAD, such as dark count rate and photon detection efficiency, degrades in more advanced CMOS technology nodes. Its performance was 7.3ps resolution and 1.0LSB single-shot precision. The integral nonlinearity was 1.2LSB rms with the help of INL LUT calibration. The area of the fabricated TDC is 0.03mm^2. With this compact size and very low power consumption features, an array of the proposed TDC can be integrated with SPADs to build a state-of-the-art dSiPM for biomedical imaging applications.

Chapter 5

PROTOTYPE DEMONSTRATION OF THE SENSING SYSTEM FOR TOF MEASUREMENT

In this chapter, the prototype of the sensing system based on SPAD and TDC that we developed in previous chapters is discussed. The SPAD is able to offer single-photon level sensitivity. The TDC can extract the timing information of the responsive pulse of SPADs with respect to the incident triggering laser photons. For preliminary demonstration purpose, the current sensing system is integrated on a PCB level rather than on a single silicon chip level. The system integration, time-of-flight measurement setup and signal synchronization of the proposed time-resolved photon sensing system will be described in Section 5.1, followed by a system characterization, and time-of-flight measurement to mimic the ToF PET imaging system in Section 5.2 and Section 5.3, respectively.

5.1. System Integration

To build the time-resolved single-photon sensing system, the SPAD and the TDC are integrated onto a single PCB. Taking into consideration that the SPAD and TDC work at two different voltage levels, i.e. 3.3V for SPAD and 1.2V for TDC, level shifters are deployed to shift down the outputs of the SPAD before they are fed into the TDC chip. The diagram of such a PCB is shown in Figure 5-1, in which a photo of the fabricated PCB is also inserted. Since the TDC designed in this thesis requires two input channels (i.e. one for the “Start”, another for the “Stop”), two PCBs will be

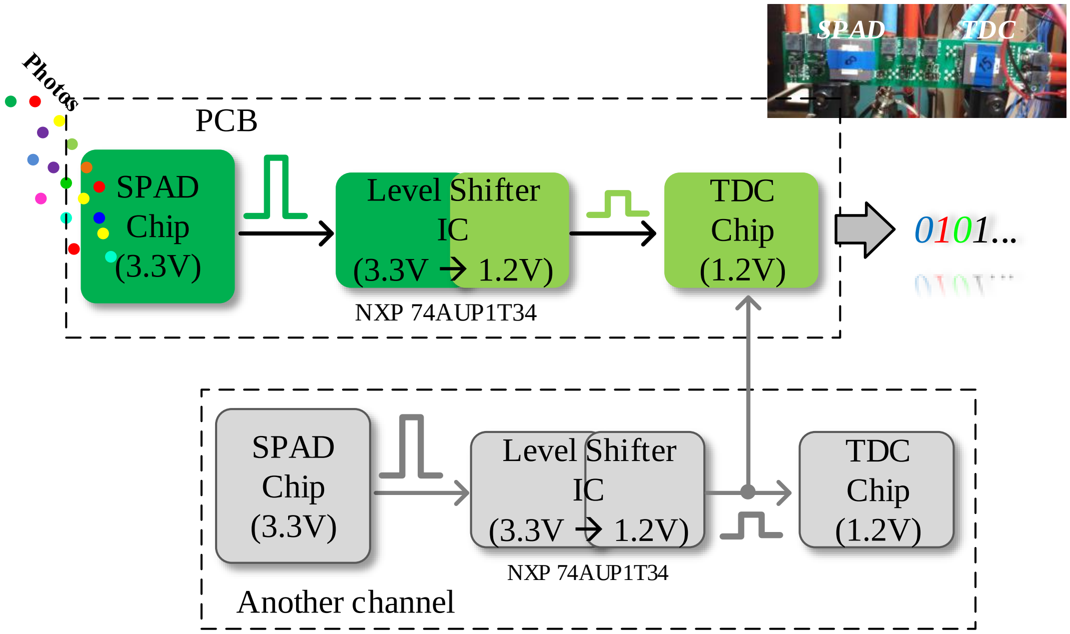


Figure 5-1, PCB designed to integration the SPAD chip and TDC chip together.

utilized. The second PCB is used to provide an additional input channel signal for the TDC chip on the first PCB.

In a real PET imaging system, two PMTs are required at 180° alignment to record the photons corresponding to one annihilation event. Scintillation crystals are typically coupled with these photodetectors in order to convert the high-energy gamma-rays into photons at visible wavelength range. The conceptual diagram is shown in Figure 5-2(a). The output pulses from the PMTs will be fed to a time-stamping block (for example, a TDC here). Then, the digitized timing output will be stored and subsequently used to improve the quality of the reconstructed images.

However, in order to mimic the coincidence detection and timing measurement setup in Figure 5-2(a) using the SPAD and TDC that are developed in this thesis, several challenges arise. First, the SPAD pixels in this thesis are implemented on 1mm×1mm silicon dies. It should be mentioned that the bonding pads are also located in this 1mm×1mm silicon area. Wire bonding technique is used to provide electrical connection between the silicon die and PGA69 ceramic package. Besides, reducing the crystal pitch of scintillation crystals faces several challenges, such as complex and expensive assembly, and reduced scintillation light output. A detailed analysis on scintillation crystal design is beyond the scope of this thesis, but useful information can

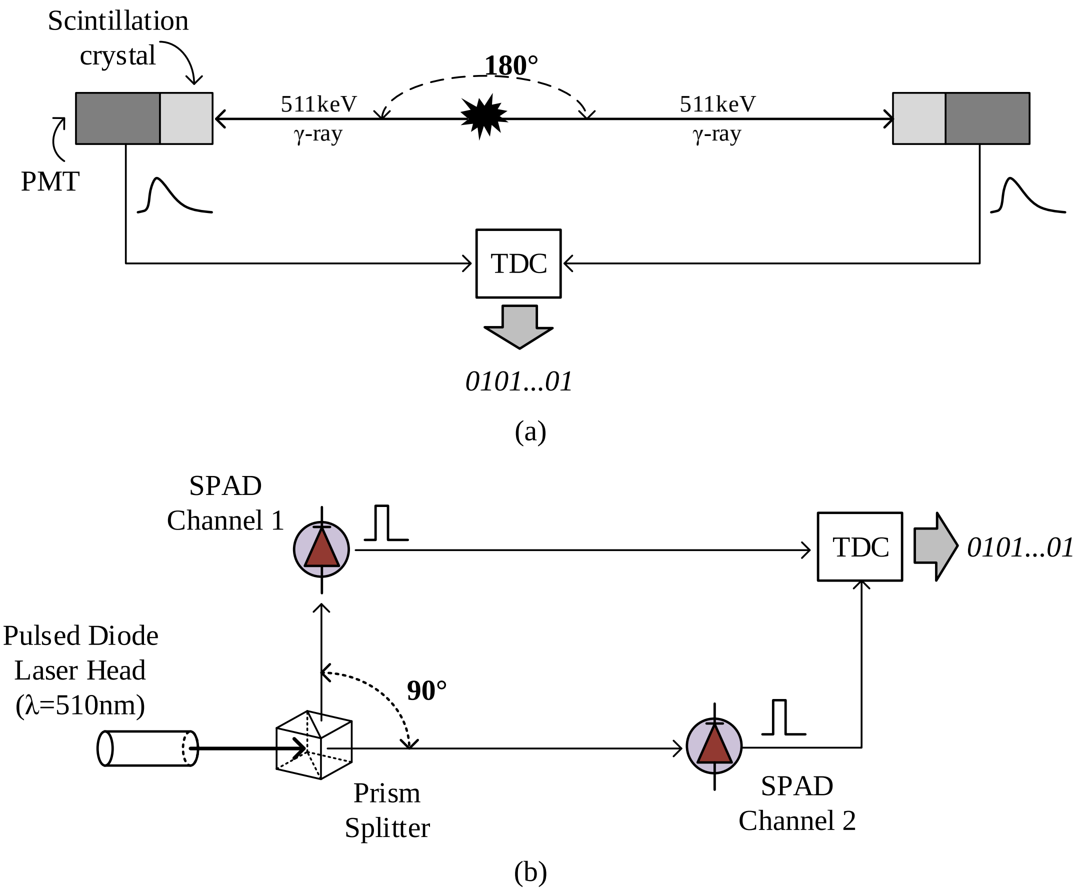


Figure 5-2, Setup difference between (a) the coincidence detection system in real PET imaging application, and (b) the proposed sensing system in this thesis.

be found in [11]. Therefore, at the current step, it is impossible to directly couple 1mm×1mm (the minimum crystal pitch available in our lab) scintillation crystals to the surface of our developed SPAD chips without breaking those bonding wires. The second issue is regarding laboratory safety. Usually, the gamma ray used in the experiment is generated by radioactive sources, such as ^{22}NaI. Nuclear material usage permission and safety training are required in order to use the radioactive isotope sources in the optoelectronic laboratory. And this may also limit other students' access to the laboratory. Last but not least, the gamma rays emitted from an isotope source are fully random and cannot be controlled, in both temporal and spatial manner. This will greatly affect the alignment and prolong the data acquisition time needed to obtain a good statistics in coincidence timing measurement. Considering all the above challenges, we implemented an alternative setup with the developed components to demonstrate the imaging system. The system is shown in Figure 5-2(b).

A pulsed diode laser, PicoQuant laser driver (PDL 800-B) coupled with a 510nm wavelength laser head, is used to generate the trigger photons. By generating low-energy photons at visible wavelength, the laser will mimic the functions of the radioactive source and the scintillation crystal. The laser's repetition rate and power intensity can be configured by the laser driver. Thus, a precise control on the number of photons striking the SPADs and their rates is achieved.

An optical prism beam splitter (polarizing beam splitter cube, WPBS254-VIS, ThorLabs) is chosen to separate the incident laser beam into two beams with equal power. These two beams out of the prism splitter are orthogonal. These two beams mimic the opposing gamma-rays in Figure 5-2(a), but are at 90° rather than 180° in Figure 5-2(b).

Two SPADs are placed after the prism splitter to detect the laser photons. Their distances to the prism splitter can be changed, so the photon travel time will also change. Therefore, the time-of-flight timing value can be easily shifted. Then, the outputs from these two SPADs' channels are connected to the TDC module, and the time-of-flight timing information are digitized and stored by the high-speed LeCroy oscilloscope.

Using this proposed setup, we are able to demonstrate the time-of-flight measurement concept utilizing the components that are developed in previous chapters.

5.2. System Characterization

Before the actual time-of-flight timing measurements, we need to characterize the instrument response function (IRF) of these two SPAD channels in order to determine the proper settings for both the laser power and SPADs' excess bias voltage. Figure 5-3 shows the conceptual setup and the timing diagram. It is noted that we included the prism splitter on the light propagation path from the laser head to the SPAD. Thus, the effect of the prism splitter on the final timing IRF is also considered in this measurement. The measurements are performed at room temperature under ambient conditions.

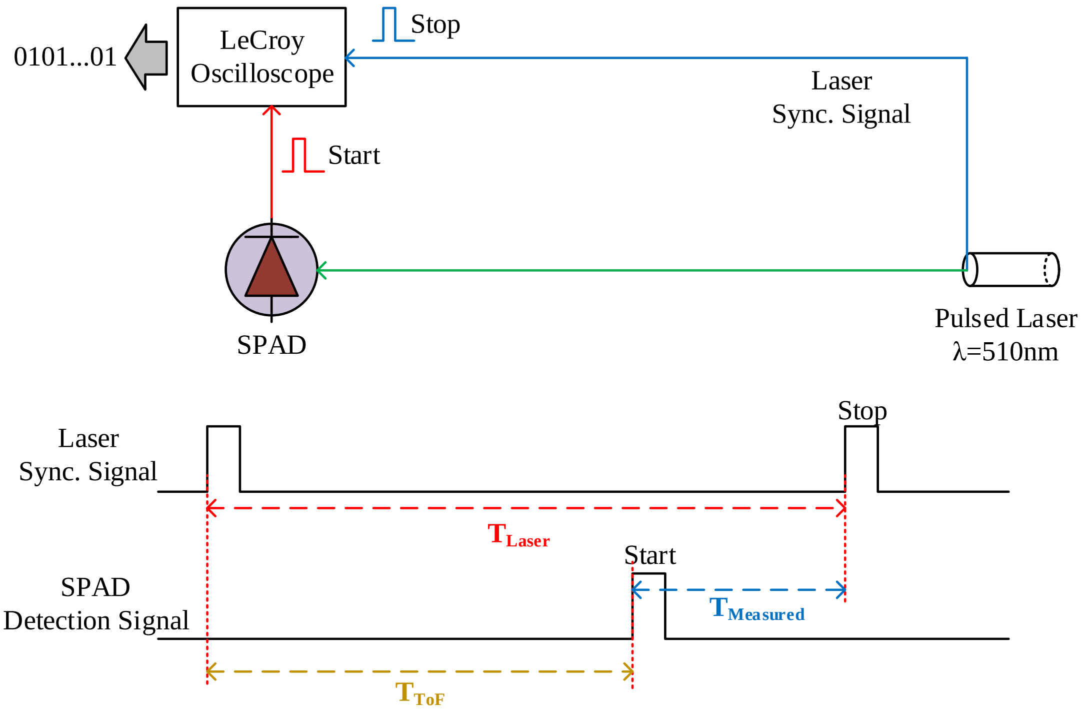


Figure 5-3, Diagram of the setup used to measure the timing response spectrum of the SPAD.

The repetition rate of the laser driver is set to 10MHz. Three intensity settings on the laser driver are used 3.1, 3.7 and 4.3, which corresponds to the laser power of 0.53μW, 0.79μW, and 2.2μW, respectively. The laser power is measured by an optical meter (model 2835c, Newport) with a calibration silicon detector module (model 818-SL, Newport). The laser power of 2.2μW is the maximum power that we can apply to the SPAD in this setup. Otherwise, the SPADs will enter saturation, i.e. they cannot be fully recharged back to Geiger mode before a new photon hits the SPADs. It is also noted that the laser intensity on the driver is not linearly related to the laser power. Five excess bias voltages, from 0.4V to 1.2V with a 0.2V increment, are measured for both SPAD channels.

The IRF measurement results are given in Figure 5-4. The FWMH of the Gaussian fitting on the measured timing spectrum is taken to represent the timing resolution. It can be clearly observed in this figure that the timing resolutions for both SPAD channels are greatly improved when the laser power is increased. The FWMH is also weakly dependent on the excess voltage. For example, the FWMHs of SPAD channel 1 are 765ps and 219ps at 0.4V excess voltage, with laser power of 0.53μW and 2.2 μW,

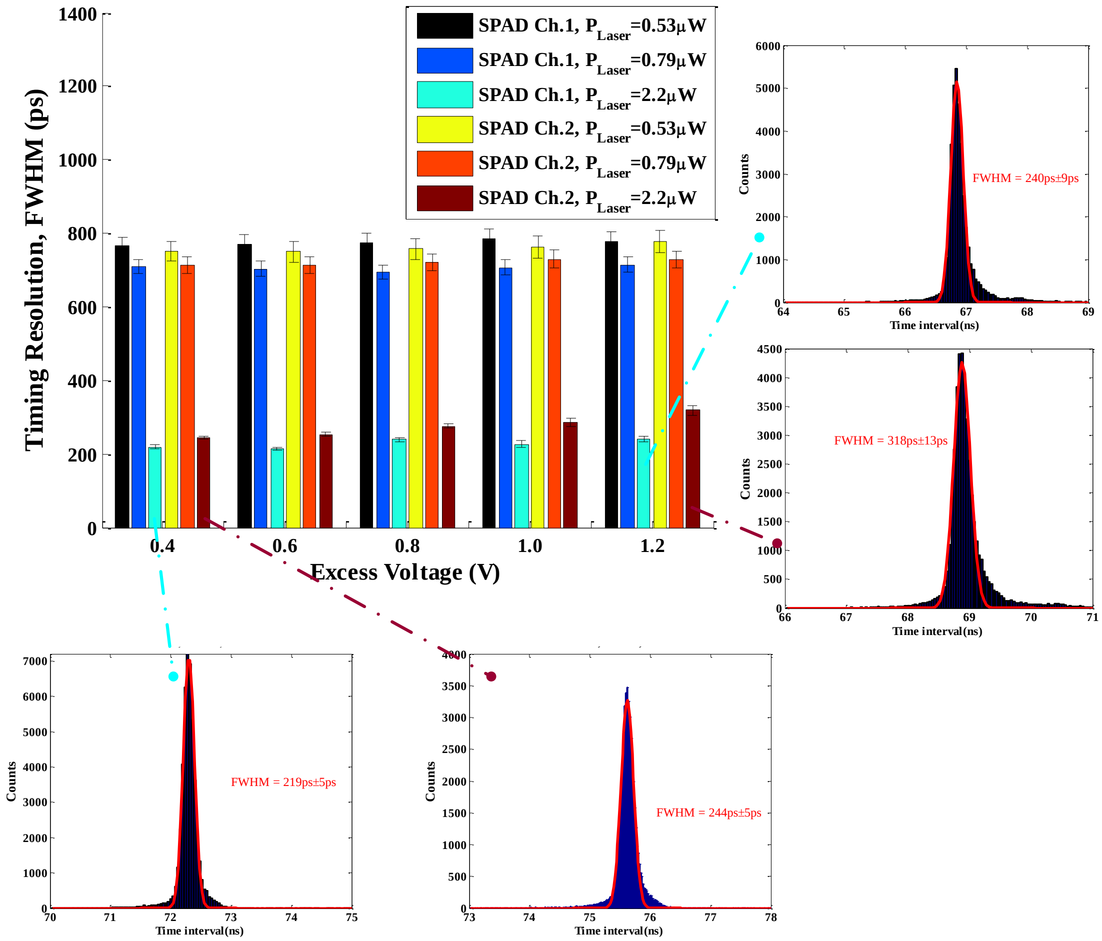


Figure 5-4, FWHM of the measured IRF of two SPAD channels with different laser powers and excess bias voltages.

respectively. When the excess voltage increases to 1.2V, the FWHM value get worse, i.e. 776ps and 240ps with laser powers of 0.53μW and 2.2μW, respectively. It is also noticeable that the timing resolution of SPAD channel 1 is slightly better than that of SPAD channel 2. This is due to the performance variations between SPADs. The absolute value of the laser power is relatively for single photon detection, however, considering the area ratio between the active area of the SPAD pixel (10×10 μm^2) and the spot side of laser between (~2×2 mm^2) is 2.5×10^{-5}, the effective power density that applied to the SPAD is very low.

Four inserts in Figure 5-4 show that the measured IRF of both SPAD channels at 2.2μW laser power, with 0.4V and 1.2V excess bias voltage. Even though the measured IRF of the SPAD channels are slightly worse (i.e. with wider timing spectrum and

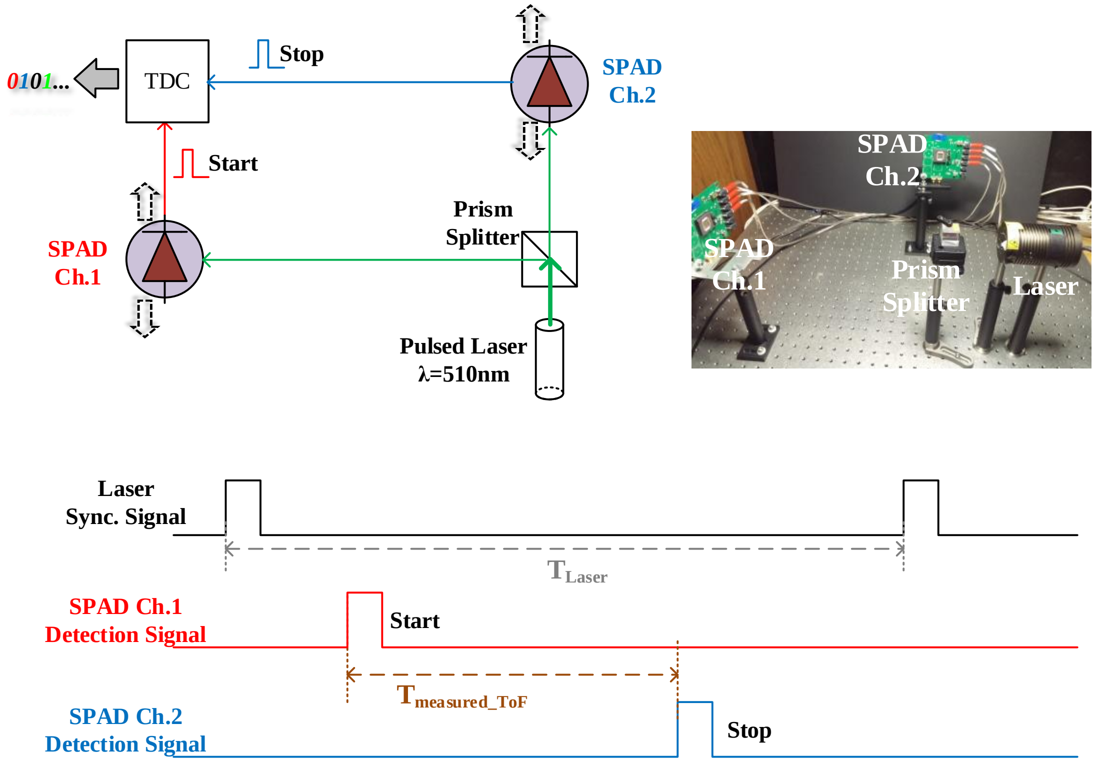


Figure 5-5, Time-of-flight measurement set up.

larger FWHM value) at 1.2V excess voltage than at 0.4V, we also need to take into consideration the increased PDE of the SPAD at higher excess bias voltage. When choosing a low excess voltage in this timing measurement, the detection efficiency of the SPAD will be reduced, and the measurement time will be greatly prolonged before a sufficient number of counts is accumulated. Therefore, as a tradeoff between the timing resolution and detection efficiency, we choose the laser power of 2.2μW with 10MHz repetition rate and the excess voltage of 1.2V, respectively. These configurations are used for the time-of-flight measurements in the rest of this chapter.

5.3. Time-of-Flight Measurement

After the initial characterization of the two SPAD channels, the time-of-flight measurements are implemented using the following setup shown in Figure 5-5. A photography of the actual set up in the laboratory is also included in this figure. The locations of two SPAD channels can be moved forward and backward to change the distances between the SPADs and the prism beam splitter. Thus, the time-of-flight time

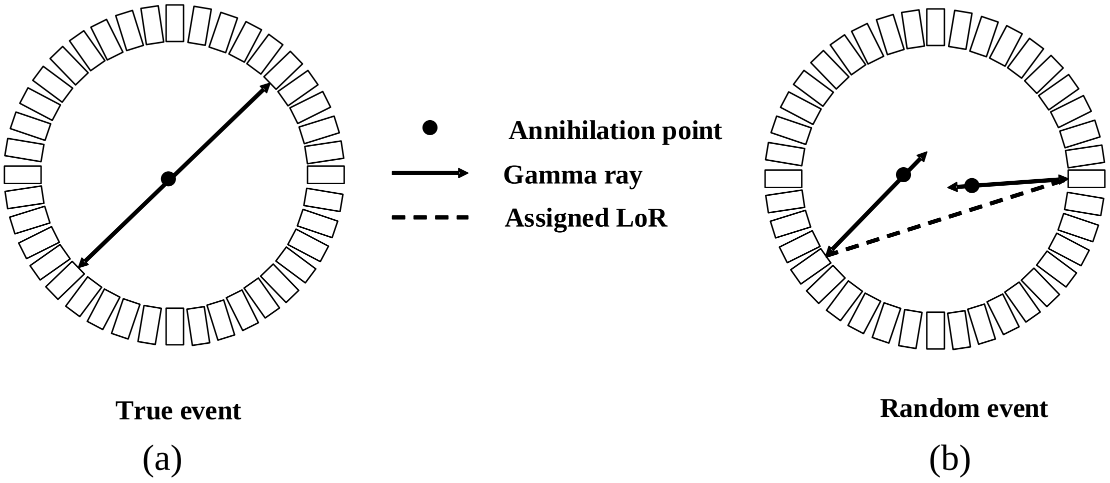


Figure 5-6, Two different types of coincidence events in PET imaging: (a) True event and (b) Random event.

interval can be changed in this setup. An illustration timing diagram of this measurement is shown at the bottom of Figure 5-5. The time intervals between the detection signals of the two SPAD channels ($T_{measured\_ToF}$) are measured with the TDC developed in Chapter 4.

In this setup, the pulse laser is employed to generate the triggering photons for detection purpose at a repetition rate of 10MHz. The prism beam splitter is used to split two laser beams at 90° angle, which then have a certain probability to cause coincidence pulses at the two SPAD channels. As discussed in Figure 5-2, this can emulate the two PMTs placed at 180°to detect the two anti-parallel gamma rays that are coming from a single annihilation event in the PET imaging system [11], [16], [31], [44], [206].

5.3.1. Considerations on False Coincidence Events

It is noted that in PET imaging, there are different types of coincidence detection events. The ideal photon detection condition is called the true coincident event. As shown in Figure 5-6(a), two 511keV gamma-ray photons are detected by a detector pair along the LoR passing through the emission position. In practice, however, this is always not achievable since usually some undesired background types of events are happening. Random coincidence event, as shown in Figure 5-6(b), occurs as a result of detecting two unrelated annihilation events that are emitted from two separate positions and detected by a detector pair along the LoR within the same time window. This results

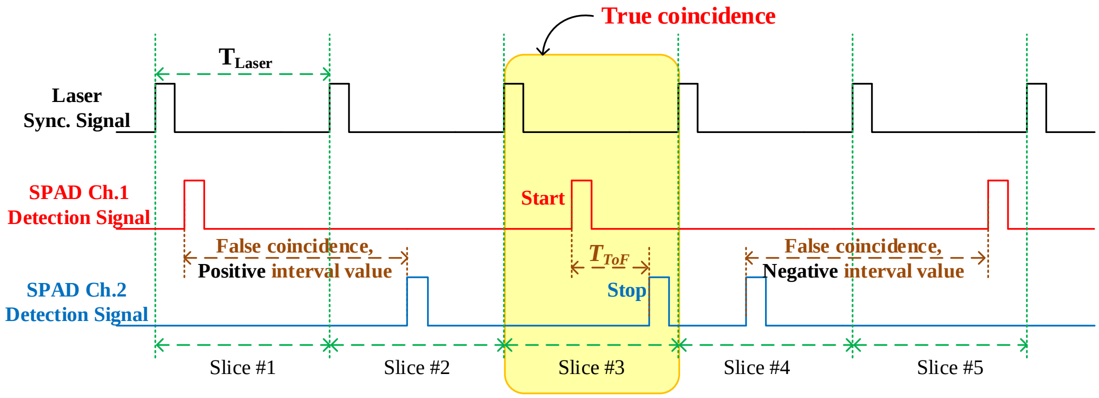


Figure 5-7, False coincidence issue in the time-of-flight measurement using our proposed sensing system.

in a high background image, false position information, and a reduction of image contrast. The random coincidence rate in a PET scanner is estimated [222] as $R_{ij}=2\tau S_a S_b$, where τ is the time resolution of the system, 2τ represents the time window for coincidence selection, S_a and S_b are the signal rates of two detectors a and b respectively, that are defining a given LoR. From this equation, one way to reduce the random coincidence rate is to applying narrow coincidence timing window during data acquisition. Hence, a fast detector with superior timing resolution is necessary. On the other hand, the number of true events is reduced when applying a shorter timing window. Therefore, there is a tradeoff between a reduction of undesired events and the sensitivity [222].

There is another type of false coincidence event in PET imaging, namely the scatter coincidence. It is caused by the Compton scattering effect of 511keV photons which changes the direction of the annihilation photons. This would produce false LoRs, and consequently yields false position information, leading to reduced image contrast as well. The scatter coincidence is usually rejected by applying a narrow energy window around the main photon peak (e.g. the 511keV peak for a ^{22}NaI source). More details on the scatter coincidence and energy window (or energy resolution) are given in [3]–[5], [44], [223]–[225].

An issue analogous to the random event exists in our proposed sensing system. It is shown in the Figure 5-7 below. In this conceptual timing diagram, five detection slices are defined by the repetition period of the laser synchronization signal. Only in one out of five slices (the third one), is a true coincidence detection event observed. Several factors will contribute to a low true coincidence detection rate. The first is that the optical sensitive area of the SPAD is very small (10×10 μm^2) compared to the laser spot size (larger than 1×1 mm^2). So most of the incident photons are not detected by the SPADs. As the second factor, these two SPAD channels are working in a fully asynchronous manner. Therefore, the probability of two SPADs being triggered by photons that come from the same laser pulse is low. Third, the TDC is automatically triggered by the rising edge of its start channel and is reset once the conversion is completed. Thus, all the components in this system are operating asynchronously.

As illustrated in the Figure 5-7, only SPAD channel 1 is responsive to the laser photons in the first slice, while only SPAD channel 2 is responsive in the second slice. This will falsely trigger the TDC and register a false time-of-flight measurement. Another case is illustrated in the fourth and fifth slices. A false coincidence event will be registered with a negative input to the TDC. All these issues need to be considered and calibrated in our time-of-flight measurement.

5.3.2. Measurements and Calibrations

In order to be able to select the true coincidence event in the time-of-flight measurement, we utilized the same concept as in PET imaging, by applying a proper time window to reject the false coincidences. To provide a reference to help select the true coincidence events, we first measured the time-of-flight interval with a high-speed, high-resolution LeCroy oscilloscope (Model 625Zi, 20GS/s). By properly setting the horizontal time scale and triggering conditions, and using the built-in TDC function in the oscilloscope, we can measure the coincidence event when the SPAD channel 1 is triggered prior to the SPAD channel 2, and when the interval is within the laser repetition period.

The measured coincidence timing spectrums with the oscilloscope are given in Figure 5-8. Gaussian fitting is used and the fitting results are indicated by the red lines in the histograms. The FWHM is taken as the coincidence timing resolution. The absolute value of each peak in the spectrum corresponds to absolute time-of-flight information, which can be converted to the photons' travel time differences between these two SPAD channels. The measurements are repeated for three different time-of-flight intervals. For each measurement, 2000 counts are accumulated. The variations of

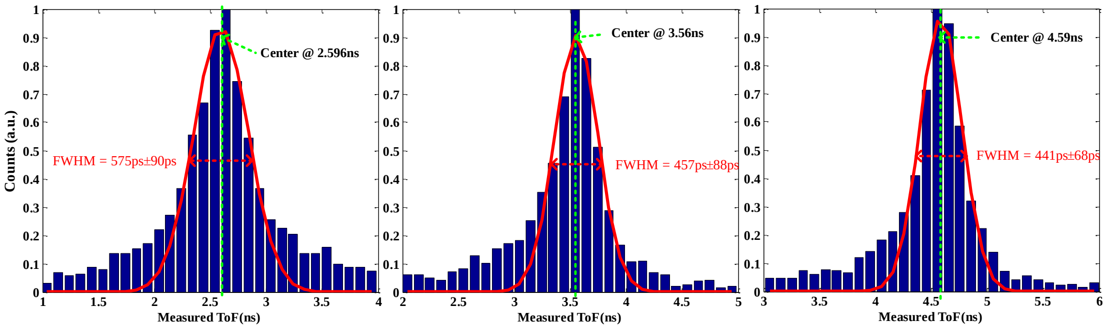


Figure 5-8, Measurement coincidence timing results with LeCroy oscilloscope.

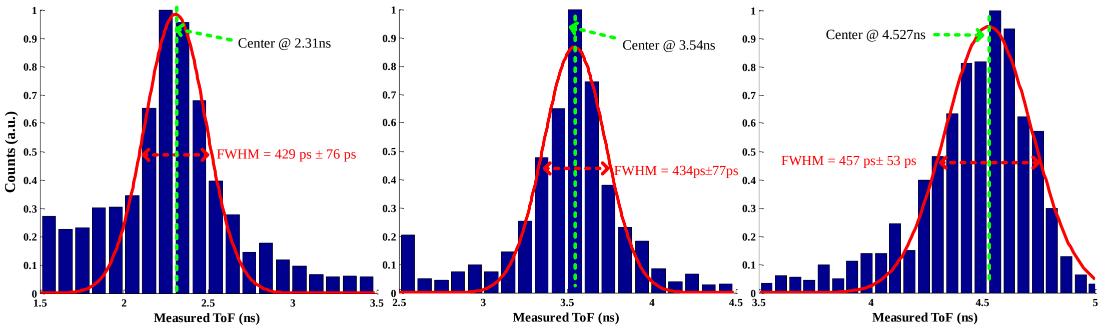


Figure 5-9, Measured coincidence timing results with our proposed system.

FWHMs are due to the fittings and variable IRF of SPADs when their distances to the prism splitter are changed.

According to the center peaks in Figure 5-8, we can set the time windows around the measured center when using our proposed sensing system. The widths of the time windows should be larger than the FWHMs in these time spectrums.

Then, we performed the time-of-flight measurements using the sensing system, which is shown in Figure 5-5. It is noted that since the probability of coincidence detection is low, we need to record a large number of measurement data and then to use the timing information in Figure 5-8 as time windows to filter and select the true coincidence events.

Figure 5-9 presents the measured timing spectrums of three different time-of-flight intervals. On average, the measured coincidence timing resolution with our system is 440ps ± 69ps. When compared with the reference results using the LeCroy oscilloscope (i.e. 491ps ± 82ps, average value), our proposed system achieves slightly better timing resolution and comparable jitter performance.

The actual measured time-of-flight value at three different lengths also differ from the reference measurements. These numbers can be seen from both Figure 5-8 and Figure 5-9, where they are labeled at the center of each peak. The accuracy errors of the time-of-flight interval measurement are from -0.56% to -11%.

We notice that the peak center positions are shifted from the reference measurement with the LeCroy oscilloscope. We consider these shifts are due to the bend of cables that will affect the signal's transmission delay. This effect can be avoid once the components are fully integrated on a single chip. Also, when the position of one SPAD channel is moved to adjust the time-of-flight interval, adjustments in both the optical alignment and the SPAD IRF will change the shape of the final spectrum. This will make the measured timing spectrum non-symmetrical. For example, the first spectrum in Figure 5-9 may suggest that the SPAD channel on the lead edge (i.e. the *Start* channel of the TDC, which is the SPAD channel 1 in this setup) have a larger timing jitter. Besides, as the distance from the SPAD to the prism splitter changes, the IRF of the SPAD will be greatly affected. These effects limit the final coincidence timing resolution on the measured spectra. Nonetheless, this setup successfully demonstrates the concept of single-photon detection and time-of-flight measurement using the photodetectors and TDC that are developed in a standard digital CMOS technology.

5.3.3. Impact of the Timing Resolution on Image Quality

Taking the 440ps coincidence timing resolution achieved with our proposed system, we can calculate its benefits to a ToF PET imaging application. According to equation (1-2), we can calculate the SNR improvement of the ToF PET over the conventional PET.

$$\frac{SNR_{ToF\_PET}}{SNR_{Conv\_PET}} = \sqrt{\frac{D}{\frac{c}{2}\Delta t}} = \sqrt{\frac{2\times 0.4}{3\times 10^{8}\times 440\times 10^{-12}}} \approx 2.5. \qquad (5\text{-}1)$$

Here, we take the body size under imaging *D* as 40cm. Thus, the final images resulting from a ToF PET system can be improved by a factor of 2.5. We can also calculate the effective sensitivity increase factor using the following equation [10],

$$\text{G}=\frac{2D}{c\Delta t} \approx 6.1. \qquad (5\text{-}2)$$

Therefore, a factor of 6.1 increase in the effective sensitivity can be achieved.

Since all the components in this setup will contribute to the final coincidence timing resolution, we can estimate the jitter from the system by the following equation,

$$\begin{aligned}\sigma_{other} &= \sqrt{\sigma_{total}{}^2 - \sigma_{TDC}{}^2 - \sigma_{SPAD.Ch1}{}^2 - \sigma_{SPAD.Ch2}{}^2 - \sigma_{laser}{}^2} \\ &= \sqrt{440^2 - 10^2 - 240^2 - 318^2 - 130^2} \\ &\approx 128\,ps\end{aligned} \tag{5-3}$$

Here, the jitter of the laser is considered as 130ps, which is quoted from the product manual [226], and the jitters of SPAD channels are taken from the measurement results in Figure 5-4.

Now, we are able to calculate the intrinsic coincidence timing resolution of our system if it is deployed for a ToF PET imaging application. The jitter of SPAD developed in this work was previously characterized in our group using a solid-state 7ps pulsed laser, which was 60ps [227]. The jitter of the scintillation crystal (e.g. LYSO) is estimated as 300ps [13], [228], [229]. Thus, the system's intrinsic coincidence timing resolution will be,

$$\begin{aligned}\sigma_{total}' &= \sqrt{\sigma_{TDC}{}^2 + \sigma_{SPAD1}{}^2 + \sigma_{LYSO1}{}^2 + \sigma_{SPAD2}{}^2 + \sigma_{LYSO2}{}^2 + \sigma_{other}{}^2} \\ &= \sqrt{10^2 + 60^2 + 300^2 + 60^2 + 300^2 + 128^2} \\ &\approx 451\,ps\end{aligned} \tag{5-4}$$

Therefore, similar SNR gain and effective sensitivity improvement of the reconstructed images in the ToF PET system are expected.

5.4. Conclusions

In this chapter, a prototype of a low-cost and compact sensing system was proposed and implemented. This system deployed SPADs for single-photon optical detection, and a TDC for time interval digitization. The SPADs and TDCs were both designed and fabricated in the standard digital CMOS process (130nm, IBM) that features a low fabrication cost and good performance. The current version was integrated on a PCB. However, the system integration on a single chip could be implemented easily in a future version of this work. A time-of-flight measurement setup was demonstrated using the proposed prototype to perfom the coincidence timing measurements. The measurements were repeated with four different time-of-flight values and calibrated with a high-speed oscillacope. The peak positions of measured timing spectrum also matched with the distances of two SPAD channels with respect to the prism splitter. From calculations, the prototype was expected to improve the SNR of PET images by a factor of 2.5, and improve the effective sensitivity in PET imaging by a factor of 6.1.

Chapter 6

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

In this thesis, a compact and low-cost single-photon sensing system was designed and fabricated based on a mainstream standard digital CMOS technology (130nm, IBM). A prototype was implemented and demonstrated for time-of-flight measurement as to emulate the ToF PET imaging applications. This chapter summarizes the research work and provides recommendations for future improvements.

6.1. Summary and Discussion

PET imaging is an attractive non-invasive nuclear scanning modality. Compared with conventional anatomic medical imaging methods such as CT and MRI, PET imaging offers a unique advantage, which is its high sensitivity to biological activity at molecular level. Thus, PET imaging can distinguish lesion cells from normal cells. It is widely used for early-stage tumor detection, cancer treatment planning and monitoring. An advanced innovation of PET imaging is the ToF PET. In ToF PET, the arrival time difference between two gamma-rays in a coincidence event is employed to localize the annihilation point. This technique helps improve the signal-to-noise ratio in the reconstructed images and subsequently results in better image contrast. Another important innovation is the PET/MR combined imaging modality. The fusion images taken through PET/MR can take both advantages of functional information obtained with PET and anatomic information obtained with MRI.

However, these medical imaging systems bring numerous challenges to current detection technologies. First, PMT, the most commonly used photodetector, is very expensive, bulky and requires over 1000V operating voltage. Second, timing circuits are needed to record the arrival time difference of coincidence detections. Third, since PMT is a vacuum tube device, its performance will be dramatically degraded in high magnetic fields (from 0.3T to 10T in MRI system).

Therefore, the purposes of this work were to prove concepts and build a compact, low-cost single-photon sensing system towards PET imaging applications. To resolve the cost issue, a mainstream standard digital CMOS process was chosen to design and implement both detectors and time conversion circuits. The 130nm process node was selected because the photodetector's performance is greatly dependent on CMOS process and degrades in advanced technology nodes. Thus, 130nm process was considered as a tradeoff between photodetector's optical properties and digital circuits' performances.

The single photon avalanche diode was used as the photodetector technology in this work. Because the SPAD was operating above its breakdown voltage, a large gain that was comparable to PMT's (~10^6) was achievable. Single-photon level optical sensitivity was obtained with SPADs. The SPAD design in this work was fully compatible with the 130nm digital CMOS process. Thus, it has the potential to be easily integrated with complex circuits on the same silicon substrate.

Time-to-digital conversion was done with a novel TDC in this work. Noting that we need to integrate the TDC with the SPAD on the same chip, there were two important factors to be considered in the design of the TDC. The first one was compact size, because one would like to reserve as much of the silicon area for optical detection as possible. The second one was low power consumption, since the thermal generated noise of SPAD was heavily dependent on temperature. Therefore, a gateable architecture was proposed to implement the TDC. The delay elements in the TDC were arranged in a ring format and worked as a ring oscillator to use these delay elements repeatedly. This helped to reduce the total area.

In Chapter 1, the application background on PET imaging was introduced firstly. The concept, principle and requirements of a PET system were discussed. Then, two advanced innovations of PET imaging were reviewed. One was the ToF PET and another was the PET/MR imaging. Their advantages over the conventional PET imaging were analyzed. Moreover, some challenges were discussed in terms of the photodetectors and electronics targeting these applications. Some of the current competing photodetector technologies used in PET imaging applications, such as PMTs, APDs, SPADs and SiPMs, were reviewed and both their advantages and disadvantages were assessed. An important functional electronic component was the TDC. There different types of TDCs, i.e. analog-type, CMOS-based digital-type and FPGA-based digital-type, were reviewed and their state-of-the-art designs were summarized. Starting from the PET imaging application, the research motivation of this work was presented as well. Particular research focuses were given to low-cost and compact SPAD and TDC designs in a mainstream standard digital CMOS process. The research contributions were summarized, and followed by the outline of this thesis.

In Chapter 2, a comprehensive and accurate model for the CMOS SPAD was developed. A circuit model that included all the behaviors (i.e. static, dynamic and statistical) would be useful for circuit designer to elaborate SPAD with complex functional circuits to build a large sensor system. Thus, this model was implemented in Verilog-A HDL code, which was fully compatible with mainstream commercial simulation EDA tools, such as HSpice and Cadence Spectre. Especially, in order to improve the modeling accuracy, nonlinear variable diode resistor, stray capacitors and temperature dependences of parameters were considered. Moreover, detailed steps on how to extract the physical parameters from the inter-avalanche time measurements were discussed. With a passive quenching resistor, the SPAD model was validated to show the dark noises due to difference noise contributors. The simulation results were compared with measurement results and a good match was obtained to demonstrate its modeling accuracy.

In Chapter 3, the silicide layer in the SPAD design was studied. Two test structures were designed and implemented to investigate the impact of silicide layer on SPAD. Detailed measurements and comparison were performed for both silicided and non-silicided SPADs. The non-silicided SPAD had a breakdown voltage 1V lower than the silicided SPAD. At different excess bias voltages and temperatures, the non-silicided

SPAD showed about 5 times improvement in both DCR and PDE. But the after-pulsing probability was 2 times higher than the silicided SPAD. Possible reasons that may result in these changes were discussed.

In Chapter 4, the research effort was devoted to develop a TDC in 130nm CMOS technology. Targeting the ToF PET imaging application, compact size, low power consumption and high resolution were the main performance metrics that were highlighted. A gateable Vernier ring oscillator architecture was chosen for its unique features. First of all, fine time resolution was guaranteed through its Vernier working principle. Second, since the delay elements were arranged in a ring format and were iteratively used, the total area was reduced. Third, a low power consumption was obtained with the gateable logic design. The fabricated prototype chips were fully characterized. Its main performance metrics included 7.3ps resolution, 1LSB precision, 9ns dynamic range, 0.03mm^2 area, 1.2mW power consumption and 1.2LSB integral nonlinearity after off-line calibration.

In Chapter 5, the SPAD and TDC developed in this work were integrated on a PCB level. With a pulse diode laser, the IRF of two SPAD channels were first investigated to obtain the proper settings for both the laser power and the SPAD excess bias voltage. Then, a time-of-flight setup was built. This setup could emulate the coincidence photon detection and ToF timing measurement as in a real ToF PET imaging system. The coincidence timing resolution was obtained with Gaussian fitting on the measured timing spectrums. The ToF measurement was repeated with several different time intervals, which were realized by moving the location of one SPAD channel. The measured FWHM coincidence timing resolution was 440ps. This superior timing resolution would help to improve the SNR by a factor of 2.5, and to increase the effective sensitivity by a factor of 6.1, if this prototype sensing system was employed in a ToF PET imaging system.

6.2. Recommendations for Future Work

The main goal in this research was to develop a low-cost, low-power and miniaturized sensing system that is capable to detect light down to single-photon level with sub-nanosecond coincidence timing resolution towards the ToF PET imaging application. Thus, the work conducted in this thesis started with the design, analysis and characterization of a single-pixel SPAD in a mainstream standard digital CMOS process. High-performance TDC circuit was also developed in the same CMOS technology. Therefore, the feasibility of integrating the CMOS SPAD and TDC on one silicon die is possible. A low-cost and compact sensing system based on these components was built and preliminary experiments were conducted to demonstrate the time-of-flight measurement. Good coincidence timing resolution was obtained with the prototype.

However, some drawbacks of the current sensing system exist, such as low photon detection efficiency, PCB-level system integration, and small pixel area and pixel number. In order to realize the fully integrated CMOS single-photon sensing system with time stamping capability, the following improvements and recommendations should be considered in the future.

1) The PDE of SPADs in standard digital CMOS process needs to be improved. In chapter 3, we conducted a rough calculation showing that the light transmittance would be greatly improved by eliminating the dielectric layers above the active area of the SPAD. This is because PDE is highly related to the light transmittance in the passivation layers. Because of the presence of polyimide/nitride/oxide/silicide layers above the SPAD active area in the chosen technology, most of the incident photons are reflected. To improve the light transmittance, a proper post-process is required to fully remove or thin the thickness of these passivation layers [230]. As a further surface treatment, anti-reflection coating can be applied to the SPAD chip to maximize the light transmittance, thus yielding an improved PDE. An increase in the fabrication cost should be expected if these post-process steps are employed.
2) DCR needs to be reduced for SPADs in standard CMOS. Since the doping profile cannot be modified in a standard digital CMOS process, it is hardly possible to improve the noise performance. The DCR is reduced with lower excess bias voltage, but the PDE is also decreased as a side effect. So it is not recommended to reduce the DCR by simply applying a lower excess

voltage. One possible solution is to employ a diffident CMOS technology. For example, the high-voltage process (0.35μm HV CMOS) could be a better choice to implement SPADs with better DCR and PDE. Again, the tradeoff between optical detection and digital circuit performance would be raised in this case.

3) Future effort could be devoted to the design and optimization of the front-end circuits that are associated with the SPAD. In this work, the passive quenching/reset approach is adopted. Active quenching and active reset techniques can remarkably reduce the dead time of the SPAD pixel. The down side is that with the increased number of transistors in the front-end circuits, the fill factor of the pixel will be reduced.
4) The prototype system in this thesis is based on a PCB implementation. The SPAD cells and TDCs developed in this work can be fully integrated on a single chip in a future version. On-chip integration will help to reduce the system jitter, simplify the measurement setup, data synchronization and transport.
5) The current TDC prototype is based on a gateable Vernier ring oscillator design. Its oscillation frequency is subject to PVT variations. This is the reason for a large integral nonlinearity before calibration. Thus, future improvement on TDC design should take advantage of the DLL technique. Two DLLs should be used to lock the delay per element in two ring oscillators against PVT variations. Then, the nonlinearity of the TDC can be greatly improved. Also, considering that DLLs are only used to generate control voltage signals for the voltage controlled delay elements, they can be shared among a large array of TDCs that are integrated on the same chip. This will help to keep the silicon area occupied by TDCs to a minimal level.
6) When the SPADs and TDCs are integrated together to build a compact single-photon sensor, the dead time of TDC becomes important. As every TDC has its intrinsic dead time to accomplish the time-to-digital conversion, the photons that are detected during this period will not be time stamped. Thus, some timing information will be lost. A possible solution is to arrange

two TDCs operating in a interleave manner. While the first TDC is taking measurement, the second one can be reset and ready for next measurement period. When the second TDC works, the first one can transfer, store the results and reset.

7) It is noted that the silicon area occupied by electronics cannot be used for optical detection, thus reducing the fill factor of the sensor. 3-D integration technique would be an attractive approach, because it enables the three dimensional placement of electronics and photodetectors, thus, most of the silicon area on the top layer can be assigned to optical sensing [231], [232]. For example, the SPADs can be placed on the top part of the substrate (layer-A), and all functional circuits can be implemented on another substrate (layer-B). Then, the layer-B is placed underneath the layer-A. The through-silicon-via (TSV) technique can be used to provide electrical connections between these two layers. There are already commercial products utilizing TSV boding technology to realize vertically integrated sensor system. The 6mm×6mm SiPM sensors available from SensL can achieve a fill factor of greater than 90% [233].
8) The dark noises of SPAD will trigger false operations of the TDC, thus making it possible to miss the true photon detection. Thus, algorithms or circuit blocks should be implemented and integrated on-chip with the SPAD/TDC system, in order to distinguish the noises triggering from the true detection. One possible method is to record the number of triggering counts in a short period. The dark noise is randomly distributed in time domain, while the photons generated by scintillation crystals from one gamma ray will strike the SPAD in a very narrow time window. Thus, by setting a proper threshold for the registered photon counts, it is possible to separate the false triggering from the real gamma ray detection in PET imaging applications.
9) It has been proven that the multiple time stamping method combined with multi-channel TDCs can improve the coincidence timing resolution in a PET

imaging system [45]. Thus, multiple channel TDC designs should be considered in the future work.

The work presented in this thesis was a pilot research to investigate a single-photon sensitivity, low-cost and compact sensing system, targeting the ToF PET imaging application. It is demonstrated that the mainstream standard digital CMOS process could be the candidate to implement both the high-sensitivity photodetector and advanced functional circuit blocks, not to mention its greatest potential, i.e. remarkably reducing the fabrication cost and allowing miniaturization and system integration on the same silicon substrate.

REFERENCES

[1] J. Langner, “Development of a parallel computing optimized head movement correction method in positron-emission-tomography,” Ph.D. thesis, University of Applied Sciences, Dept. Computer Science, Dresden, Germany, 2003.

[2] Wikipedia, “Positron emission tomography,” 2016. [Online]. Available: https://en.wikipedia.org/wiki/Positron_emission_tomography.

[3] G. B. Saha, *Basics of PET Imaging: Physics, Chemistry, and Regulations*, 3rd ed. Cleveland, OH, USA: Springer International Publishing, 2015.

[4] S. R. Cherry, J. A. Sorenson, and M. E. Phelps, *Physics in Nuclear Medicine*, 4th ed. Philadelphia, PA, USA: Elsevier, 2012.

[5] G. F. Knoll, *Radiation Detection and Measurement*, 4th ed. Ann Arbor, Michigan, USA: Wiley, 2011.

[6] E. M. Rohren, T. G. Turkington, and R. E. Coleman, “Clinical Applications of PET in Oncology,” *Radiology*, vol. 231, no. 2, pp. 305–332, May 2004.

[7] F. Keng, “Clinical applications of positron emission tomography in cardiology,” *Ann Acad Med Singapore*, vol. 33, no. 2, pp. 175–182, 2204.

[8] A. Kadir, O. Almkvist, A. Forsberg, A. Wall, H. Engler, B. Långström, and A. Nordberg, “Dynamic changes in PET amyloid and FDG imaging at different stages of Alzheimer’s disease,” *Neurobiol. Aging*, vol. 33, no. 1, pp. 198.e1–198.e14, Jan. 2012.

[9] N. A. Mullani, J. Markham, and M. M. Ter-Pogossian, “Feasibility of Time-of-Flight Reconstruction in Positron Emission Tomography,” *J. Nucl. Med.*, vol. 21, no. 11, pp. 1095–1097, 1980.

[10] T. F. Budinger, “Time-of-Flight Positron Emission Tomography: Status Relative to Conventional PET,” *J. Nucl. Med.*, vol. 24, no. 1, pp. 73–79, 1983.

[11] H. Peng and C. S. Levin, “Recent Developments in PET Instrumentation,” *Curr. Pharm. Biotechnol.*, vol. 11, no. 6, pp. 555–571, Sep. 2010.

[12] V. C. Spanoudaki and C. S. Levin, “Photo-detectors for time of flight positron emission tomography (ToF-PET),” *Sensors*, vol. 10, no. 11, pp. 10484–10505, Jan. 2010.

[13] W. W. Moses, “Recent advances and future advances in time-of-flight PET,” *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*,

vol. 580, no. 2, pp. 919–924, 2007.

[14] M. Conti, “State of the art and challenges of time-of-flight PET,” *Phys. Medica*, vol. 25, no. 1, pp. 1–11, 2009.

[15] Z. Cheng, X. Zheng, M. J. Deen, and H. Peng, “Recent Developments and Design Challenges of High-Performance Ring Oscillator CMOS Time-to-Digital Converters,” *IEEE Trans. Electron Devices*, vol. 63, no. 1, pp. 235–251, Jan. 2016.

[16] Z. Cheng, M. J. Deen, and H. Peng, “A Low-Power Gateable Vernier Ring Oscillator Time-to-Digital Converter for Biomedical Imaging Applications,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 2, pp. 445–454, Apr. 2016.

[17] S. Vandenberghe and P. K. Marsden, “PET-MRI: a review of challenges and solutions in the development of integrated multimodality imaging,” *Phys. Med. Biol.*, vol. 60, no. 4, pp. R115–R154, Feb. 2015.

[18] B. J. Pichler, A. Kolb, T. Nägele, and H.-P. Schlemmer, “PET/MRI: Paving the Way for the Next Generation of Clinical Multimodality Imaging Applications,” *J. Nucl. Med.*, vol. 51, no. 3, pp. 333–336, 2010.

[19] C. Plathow and W. A. Weber, “Tumor Cell Metabolism Imaging,” *J. Nucl. Med.*, vol. 49, no. Suppl 2, p. 43S–63S, 2008.

[20] Wikipedia, “Photomultiplier,” 2016. [Online]. Available: https://en.wikipedia.org/wiki/Photomultiplier.

[21] Hamamatsu, “Micro PMT module H12402, H12403,” 2016. [Online]. Available: http://www.hamamatsu.com/jp/en/product/category/3100/3065/3067/H12402_H12403/index.html.

[22] K. Iniewski, *Electronics for radiation detection*. Boca Raton, FL,USA: CRC Press, 2011.

[23] M. J. Deen and P. K. Basu, *Silicon Photonics: Fundamentals and Devices*. West Sussex, United Kingdom: Wiley, 2012.

[24] D. Bronzi, F. Villa, S. Tisa, A. Tosi, and F. Zappa, “SPAD Figures of Merit for Photon-Counting, Photon-Timing, and Imaging Applications: A Review,” *IEEE Sens. J.*, vol. 16, no. 1, pp. 3–12, Jan. 2016.

[25] L. H. C. Braga, L. Gasparini, L. Grant, R. K. Henderson, N. Massari, M. Perenzoni, D. Stoppa, and R. Walker, “A Fully Digital 8x16 SiPM Array for PET Applications With Per-Pixel TDCs and Real-Time Energy Output,” *IEEE J. Solid-State Circuits,* vol. 49, no. 1, pp. 301–314, 2014.

[26] D. Palubiak and M. J. Deen, “CMOS SPADs : Design Issues and Research Challenges

for Detectors , Circuits , and Arrays," *IEEE J. Sel. Top. Quantum Electron.*, vol. 20, no. 6, pp. 6000718:1–6000718:18, 2014.

[27] F. Villa, D. Bronzi, Y. Zou, C. Scarcella, G. Boso, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, U. Paschen, and W. Brockherde, "CMOS SPADs with up to 500 μm diameter and 55% detection efficiency at 420 nm," *J. Mod. Opt.*, vol. 61, no. 2, pp. 102–115, 2014.

[28] A. Gulinatti, I. Rech, F. Panzeri, C. Cammi, P. Maccagnani, M. Ghioni, and S. Cova, "New silicon SPAD technology for enhanced red-sensitivity, high-resolution timing and system integration," *J. Mod. Opt.*, vol. 59, no. 17, pp. 1489–1499, Oct. 2012.

[29] A. Giudice, G. Simmerle, D. Veronese, R. Biasi, A. Gulinatti, I. Rech, M. Ghioni, and P. Maccagnani, "High-detection efficiency and picosecond timing compact detector modules with red-enhanced SPADs," in *Proceedings of SPIE*, Baltimore, Maryland, USA, 2012, pp. 83750–83758.

[30] Z. Cheng, X. Zheng, M. J. Deen, and H. Peng, "Development of a High Performance Digital Silicon Photomultiplier (dSiPM) for ToF PET Imaging," *J. Nucl. Med.*, vol. 56, no. supplement 3, p. 603, May 2015.

[31] Z. Cheng, H. Peng, and M. J. Deen, "High Performance Integrated Circuits for Biomedical Imaging Applications," in *225th ECS Meeting*, Orlando, FL, USA, 2014, vol. 12, no. 91, p. 1497.

[32] D. P. Palubiak, Z. Li, and M. J. Deen, "Afterpulsing Characteristics of Free-Running and Time-Gated Single-Photon Avalanche Diodes in 130-nm CMOS," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3727–3733, Nov. 2015.

[33] Z. Cheng, D. Palubiak, X. Zheng, M. J. Deen, and H. Peng, "Impact of silicide layer on SPAD in a 130nm CMOS process," *J. Phys. D. Appl. Phys.*, vol. 49, no. 34, pp. #345105-1-11, 2016.

[34] SensL, "4-side scaleable arrays of C-series sensors," 2016. [Online]. Available: http://sensl.com/products/sipmarrays/arrayc.

[35] D. Renker, "Geiger-mode avalanche photodiodes, history, properties and problems," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 567, no. 1, pp. 48–56, Nov. 2006.

[36] SensL, "C-Series Low noise, blue-sensitive silicon photomultipliers Datasheet," 2014. [Online]. Available: http://sensl.com/products/sipmarrays/arrayc/.

[37] SensL, "J-Series High PDE and Timing Resolution SiPM Sensors in a TSV Package Datasheet," 2015. [Online]. Available: http://sensl.com/products/sipmarrays/arrayj/.

[38] B. Berube, V. Rheaume, A. C. Therrien, S. A. Charlebois, and J. Pratte, "Development of a Single Photon Avalanche Diode (SPAD) Array in High Voltage CMOS 0.8 μm dedicated to a 3D Integrated Circuit (3DIC)," in *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC),* Anaheim, CA, USA, 2012, pp. 1835–1839.

[39] F. Acerbi, A. Ferri, A. Gola, M. Cazzanelli, L. Pavesi, N. Zorzi, and C. Piemonte, "Characterization of Single-Photon Time Resolution: From Single SPAD to Silicon Photomultiplier," *IEEE Trans. Nucl. Sci.*, vol. 61, no. 5, pp. 2678–2686, Oct. 2014.

[40] A. C. Therrien, B.-L. Berube, C. Thibaudeau, S. A. Charlebois, R. Lecomte, R. Fontaine, and J.-F. Pratte, "Modeling of Single Photon Avalanche Diode Array Detectors for PET Applications," *IEEE Trans. Nucl. Sci.*, vol. 61, no. 1, pp. 14–22, Feb. 2014.

[41] S. Mandai and E. Charbon, "Multi-Channel Digital SiPMs : Concept , Analysis and Implementation," in *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC),* Anaheim, CA, USA, 2012, pp. 1840 – 1844.

[42] A. Carimatto, S. Mandai, E. Venialgo, T. Gong, G. Borghi, D. R. Schaart, and E. Charbon, "A 67,392-SPAD PVTB-compensated multi-chiannel Digital SiPM with 432 Column-Parallel 48ps 17b TDCs for endoscopic time-of-flight PET," in *IEEE International Solid-State Circuits Conference*, San Francisco, CA, USA, 2015, pp. 202–204.

[43] S. Seifert, G. van der Lei, H. T. van Dam, and D. R. Schaart, "First characterization of a digital SiPM based time-of-flight PET detector with 1 mm spatial resolution.," *Phys. Med. Biol.,* vol. 58, no. 9, pp. 3061–3074, 2013.

[44] F. R. Schneider, K. Shimazoe, I. Somlai-Schweiger, and S. I. Ziegler, "A PET detector prototype based on digital SiPMs and GAGG scintillators," *Phys. Med. Biol.*, vol. 60, no. 4, pp. 1667–1679, Feb. 2015.

[45] S. Mandai, V. Jain, and E. Charbon, "A 780×800 μm^2 multichannel digital silicon photomultiplier with column-parallel time-to-digital converter and basic characterization," *IEEE Trans. Nucl. Sci.*, vol. 61, no. 1, pp. 44–52, 2014.

[46] H. T. van Dam, G. Borghi, S. Seifert, and D. R. Schaart, "Sub-200 ps CRT in monolithic scintillator PET detectors using digital SiPM arrays and maximum likelihood interaction time estimation.," *Phys. Med. Biol.*, vol. 58, no. 10, pp. 3243–3257, 2013.

[47] R. J. Walker, E. a G. Webster, J. Li, N. Massari, and R. K. Henderson, "High fill factor digital Silicon Photomultiplier structures in 130nm CMOS imaging technology," in *IEEE Nuclear Science Symposium and Medical Imaging Conference Record*

(NSS/MIC), Anaheim, CA, USA, 2012, pp. 1945–1948.

[48] S. Mandai and E. Charbon, “A 4×4×416 digital SiPM array with 192 TDCs for multiple high-resolution timestamp acquisition,” *J. Instrum.*, vol. 8, pp. P05024–P05024, 2013.

[49] D. Tyndall, B. R. Rae, D. D.-U. Li, J. Arlt, A. Johnston, J. A. Richardson, and R. K. Henderson, “A High-Throughput Time-Resolved Mini-Silicon Photomultiplier With Embedded Fluorescence Lifetime Estimation in 0.13µm CMOS,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 6, no. 6, pp. 562–570, 2012.

[50] Y. Haemisch, T. Frach, C. Degenhardt, and A. Thon, “Fully Digital Arrays of Silicon Photomultipliers (dSiPM) – a Scalable Alternative to Vacuum Photomultiplier Tubes (PMT),” *Phys. Procedia*, vol. 37, pp. 1546–1560, 2012.

[51] C. Degenhardt, G. Prescher, T. Frach, A. Thon, and R. De Gruyter, “The Digital Silicon Photomultiplier – A Novel Sensor for the Detection of Scintillation Light,” in *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, Orlando, FL, USA, 2009, pp. 2383–2386.

[52] T. Frach, G. Prescher, C. Degenhardt, and B. Zwaans, “The digital silicon photomultiplier — System architecture and performance evaluation,” in *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, Knoxville, TN, USA, 2010, pp. 1722–1727.

[53] T. Frach, G. Prescher, C. Degenhardt, R. de Gruyter, A. Schmitz, and R. Ballizany, “The digital silicon photomultiplier — Principle of operation and intrinsic detector performance,” in *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, Orlando, FL, USA, 2009, pp. 1959–1965.

[54] E. Venialgo, S. Mandai, T. Gong, D. R. Schaart, and E. Charbon, “Time estimation with multichannel digital silicon photomultipliers,” *Phys. Med. Biol.*, vol. 60, no. 6, pp. 2435–2452, 2015.

[55] Philips, “DPC Sensor Typers,” 2016. [Online]. Available: http://www.digitalphotoncounting.com.

[56] T.-H. Tsai, M.-S. Yuan, C.-H. Chang, C.-C. Liao, C.-C. Li, and R. B. Staszewski, “A 1.22ps integrated-jitter 0.25-to-4GHz fractional-N ADPLL in 16nm FinFET CM0S,” in *IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2015, pp. 1–3.

[57] K. Waheed, R. B. Staszewski, F. Dülger, M. S. Ullah, and S. D. Vamvakos, “Spurious-free time-to-digital conversion in an ADPLL using short dithering sequences,” *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 58, no. 9, pp. 2051–2060, 2011.

[58] M. Lee and A. A. Abidi, “A 9b, 1.25ps resolution coarse-fine time-to-digital Converter in 90 nm CMOS that Amplifies a Time Residue,” *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 769–777, 2008.

[59] F. Villa, B. Markovic, S. Bellisai, D. Bronzi, A. Tosi, F. Zappa, S. Tisa, D. Durini, S. Weyers, U. Paschen, and W. Brockherde, “SPAD smart pixel for time-of-flight and time-correlated single-photon counting measurements,” *IEEE Photonics J.*, vol. 4, no. 3, pp. 795–804, 2012.

[60] C. Veerappan, J. Richardson, R. Walker, D. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, “A 160x128 Single-Photon Image Sensor with On-Pixel 55ps 10b Time-to-Digital Converter,” in *IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2011, pp. 311–313.

[61] Y. Maruyama, J. Blacksberg, and E. Charbon, “A 1024x8 , 700-ps Time-Gated SPAD Line Sensor for Planetary Surface Exploration With Laser Raman Spectroscopy and LIBS,” *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 179–189, 2014.

[62] B. Markovic, S. Tisa, F. A. Villa, A. Tosi, and F. Zappa, “A High-Linearity, 17 ps Precision Time-to-Digital Converter Based on a Single-Stage Vernier Delay Loop Fine Interpolation,” *IEEE Trans. Circuits Syst. I*, vol. 60, no. 3, pp. 557–569, Mar. 2013.

[63] J. Jansson, V. Koskinen, A. Mäntyniemi, and J. Kostamovaara, “A Multichannel High-Precision CMOS Time-to-Digital Converter for Laser-Scanner-Based Perception Systems,” *IEEE Trans. Instrum. Meas.*, vol. 61, no. 9, pp. 2581–2590, 2012.

[64] A. Elshazly, S. Rao, B. Young, P. Kumar Hanumolu, and P. K. Hanumolu, “A Noise-Shaping Time-to-Digital Converter Using Swithed-Ring Oscillators - Analysis, Design, and Measurement Techniques,” *IEEE J. Solid-State Circuits*, vol. 49, no. 5, pp. 1184–1197, 2014.

[65] Y. Qin, H.-J. Kwon, M. M. R. Howlader, and M. J. Deen, “Microfabricated electrochemical pH and free chlorine sensors for water quality monitoring: recent advances and research challenges,” *RSC Adv.*, vol. 5, no. 85, pp. 69086–69109, 2015.

[66] O. Bourrion and L. Gallin-Martel, “An integrated CMOS time-to-digital converter for coincidence detection in a liquid xenon PET prototype,” *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 563, no. 1, pp. 100–103, Jul. 2006.

[67] D. Palubiak, M. M. El-Desouki, O. Marinov, M. J. Deen, and Q. Fang, “High-speed, single-photon avalanche-photodiode Imager for Biomedical Applications,” *IEEE Sens.*

J., vol. 11, no. 10, pp. 2401–2412, 2011.

[68] P. Dudek, S. Szczepan, and J. V Hatfield, "A High-Resolution CMOS Time-to-Digital Converter Utilizing a Vernier Delay Line," *IEEE J. Solid-State Circuits*, vol. 35, no. 2, pp. 240–247, 2000.

[69] A. Mantyniemi, T. Rahkonen, and J. Kostamovaara, "A CMOS Time-to-Digital Converter (TDC) Based On a Cyclic Time Domain Successive Approximation Interpolation Method," *IEEE J. Solid-State Circuits*, vol. 44, no. 11, pp. 3067–3078, Nov. 2009.

[70] C. Hwang, P. Chen, and H. Tsao, "A High-Precision Time-to-Digital Converter Using a Two-Level Conversion Scheme," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 4, pp. 1349–1352, 2004.

[71] M. Lee, M. E. Heidari, and A. A. Abidi, "A Low-Noise Wideband Digital Phase-Locked Loop Based on a Coarse–Fine Time-to-Digital Converter With Subpicosecond Resolution," *IEEE J. Solid-State Circuits*, vol. 44, no. 10, pp. 2808–2816, Oct. 2009.

[72] I. Vornicu, R. Carmona-Galan, and A. Rodriguez-Vazquez, "A CMOS 0.18μm 64x64 Single Photon Image Sensor with In-Pixel 11b Time-to-Digital Converter," in *International Semiconductor Conference (CAS)*, Sinaia, Romania, 2014, pp. 131–134.

[73] C. Niclass, M. Soga, H. Matsubara, and S. Kato, "A 100-m range 10-frame/s 340x96-pixel time-of-flight depth sensor in 0.18μm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 2, pp. 559–572, 2013.

[74] S. Henzler, *Time-to-Digital Converters*. Munchen, Germany: Springer, 2010.

[75] J. Kalisz, "Review of methods for time interval measurements with picosecond resolution," *Metrologia*, vol. 41, no. 1, pp. 17–32, Feb. 2004.

[76] D. I. Porat, "Review of Sub-Nanosecond Time-Interval Measurements," *IEEE Trans. Nucl. Sci.*, vol. NS-20, no. 5, pp. 36–51, 1973.

[77] J. Kostamovaara and R. Myllylä, "Time-to-digital converter with an analog interpolation circuit," *Rev. Sci. Instrum.*, vol. 57, no. 11, pp. 2880–2885, 1986.

[78] K. Park and J. Park, "Time-to-digital converter of very high pulse stretching ratio for digital storage oscilloscopes," *Rev. Sci. Instrum.*, vol. 70, no. 2, pp. 1568–1574, 1999.

[79] M. Tanveer, I. Nissinen, J. Nissinen, J. Kostamovaara, J. Borg, and J. Johansson, "Time-to-digital converter based on analog time expansion for 3D time-of-flight cameras," in *Proceedings of SPIE*, San Francisco, CA, USA, 2014, vol. 9022, p. 90220A.

[80] K. Määttä and J. Kostamovaara, "A high-precision time-to-digital converter for pulsed time-of-flight laser radar applications," *IEEE Trans. Instrum. Meas.*, vol. 47, no. 2, pp.

521–536, 1998.

[81] E. Räisänen-Ruotsalainen, T. Rahkonen, and J. Kostamovaara, “An Integrated time-to-digital converter with 30-ps single-shot precision,” *IEEE J. Solid-State Circuits*, vol. 35, no. 10, pp. 1507–1510, Oct. 2000.

[82] P. Chen, C. Chen, and Y. Shen, “A Low-Cost Low-Power CMOS Time-to-Digital Converter Based on Pulse Stretching,” *IEEE Trans. Nucl. Sci.*, vol. 53, no. 4, pp. 2215–2220, 2006.

[83] M. Kim, H. Lee, J. K. Woo, N. Xing, M. O. Kim, and S. Kim, “A low-cost and low-power time-to-digital converter using triple-slope time stretching,” *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 58, no. 3, pp. 169–173, 2011.

[84] T. E. Rahkonen and J. T. Kostamovaara, “The use of stabilized CMOS delay lines for the digitization of short time intervals,” *IEEE J. Solid-State Circuits*, vol. 28, no. 8, pp. 887–896, 1993.

[85] R. B. Staszewski, S. Vemulapalli, P. Vallur, J. Wallberg, and P. T. Balsara, “1.3V 20ps time-to-digital converter for frequency synthesis in 90-nm CMOS,” *IEEE Trans. Circuits Syst. II*, vol. 53, no. 3, pp. 220–224, Mar. 2006.

[86] J. G. Maneatis, “Low Jitter Process-Independent DLL and PLL Based on Self-Biased Techniques,” *IEEE J. Solid-State Circuits*, vol. 31, no. 11, pp. 1723–1732, 1996.

[87] J. Jansson, A. Mäntyniemi, and J. Kostamovaara, “A Delay Line Based CMOS Time Digitizer IC with 13 ps Single-shot Precision,” in *IEEE International Symposium on Circuits and Systems*, Kobe, Japan, 2005, pp. 4269–4272.

[88] J. Jansson, A. Mäntyniemi, and J. Kostamovaara, “Synchronization in a Multilevel CMOS Time-to-Digital Converter,” *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 56, no. 8, pp. 1622–1634, 2009.

[89] K. Nose, M. Kajita, and M. Mizuno, “A 1-ps resolution jitter-measurement macro using interpolated jitter oversampling,” *IEEE J. Solid-State Circuits*, vol. 41, no. 12, pp. 2911–2920, Dec. 2006.

[90] L. Vercesi, A. Liscidini, and R. Castello, “Two-Dimensions Vernier Time-to-Digital Converter,” *IEEE J. Solid-State Circuits*, vol. 45, no. 8, pp. 1504–1512, Aug. 2010.

[91] Y. Seo, J. Kim, H. Park, and J. Sim, “A 1.25 ps Resolution 8b Cyclic TDC in 0.13μm CMOS,” *IEEE J. Solid-State Circuits*, vol. 47, no. 3, pp. 736–743, 2012.

[92] J. Wang, S. Liu, Q. Shen, H. Li, and Q. An, “A fully fledged TDC implemented in field-programmable gate arrays,” *IEEE Trans. Nucl. Sci.*, vol. 57, no. 2, pp. 446–450, 2010.

[93] J. Wu, “An FPGA wave union TDC for time-of-flight applications,” in *IEEE Nuclear*

Science Symposium and Medical Imaging Conference (NSS/MIC), Orlando, FL, USA, 2009, pp. 299–304.

[94] C. Hervé, J. Cerrai, and T. Le Caër, “High resolution time-to-digital converter (TDC) implemented in field programmable gate array (FPGA) with compensated process voltage and temperature (PVT) variations,” *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 682, pp. 16–25, 2012.

[95] J. Y. Won, S. Il Kwon, H. S. Yoon, G. B. Ko, J. Son, and J. S. Lee, “Dual-Phase Tapped-Delay-Line Time-to-Digital Converter With On-the-Fly Calibration Implemented in 40 nm FPGA,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 1, pp. 231–242, 2016.

[96] S. S. Junnarkar, P. O’Connor, P. Vaska, and R. Fontaine, “FPGA-based self-calibrating time-to-digital converter for time-of-flight experiments,” *IEEE Trans. Nucl. Sci.*, vol. 56, no. 4, pp. 2374–2379, 2009.

[97] R. Szplet and K. Klepacki, “An FPGA-Integrated Time-to-Digital Converter Based on Two-Stage Pulse Shrinking,” *IEEE Trans. Instrum. Meas.*, vol. 59, no. 6, pp. 1663–1670, 2010.

[98] J. Song, Q. An, and S. Liu, “A high-resolution time-to-digital converter implemented in field-programmable-gate-arrays,” *IEEE Trans. Nucl. Sci.*, vol. 53, no. 1, pp. 236–241, 2006.

[99] J. Zhang and D. Zhou, “A new delay line loops shrinking time-to-digital converter in low-cost FPGA,” *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 771, pp. 10–16, Jan. 2015.

[100] M. W. Fishburn, L. H. Menninga, C. Favi, and E. Charbon, “A 19.6 ps, FPGA-based TDC with multiple channels for open source applications,” *IEEE Trans. Nucl. Sci.*, vol. 60, no. 3, pp. 2203–2208, 2013.

[101] C. Favi and E. Charbon, “A 17ps time-to-digital converter implemented in 65nm FPGA technology,” in *Proceeding of the ACM/SIGDA international symposium on FPGA*, New York, NY, USA, 2009, pp. 113–120.

[102] R. Narasimman, A. Prabhakar, and N. Chandrachoodan, “Implementation of a 30 ps resolution time to digital converter in FPGA,” in *International Conference on Electronic Design, Computer Networks & Automated Verification (EDCAV)*, Shillong, India, 2015, pp. 12–17.

[103] R. Szplet, J. Kalisz, and R. Szymanowski, “Interpolating time counter with 100 ps resolution on a single FPGA device,” *IEEE Trans. Instrum. Meas.*, vol. 49, no. 4, pp. 879–883, 2000.

[104] J. Kalisz, R. Szplet, J. Pasierbinski, and A. Poniecki, “Field-programmable-gate-array-based time-to-digital converter with 200-ps resolution,” *IEEE Trans. Instrum. Meas.*, vol. 46, no. 1, pp. 51–55, 1997.

[105] J. Wu, “On-chip processing for the wave union TDC implemented in FPGA,” in *16th IEEE-NPSS Real Time Conference*, Beijing, China, 2009, pp. 279–282.

[106] J. Wu, “Several key issues on implementing delay line based TDCs using FPGAs,” *IEEE Trans. Nucl. Sci.*, vol. 57, no. 3, pp. 1543–1548, 2010.

[107] J. Wu and Z. Shi, “The 10-ps wave union TDC: Improving FPGA TDC resolution beyond its cell delay,” in *IEEE Nuclear Science Symposium Conference Record*, Dresden, Germany, 2008, pp. 3440–3446.

[108] J. Martos, J. Soret, J. M. Benlloch, P. Conde, A. J. González, and F. Sánchez, “Time-to-Digital Converter Based on FPGA With Multiple Channel Capability,” *IEEE Trans. Nucl. Sci.*, vol. 61, no. 1, pp. 107–114, 2014.

[109] J. Kalisz, T. Orzanowski, and R. Szplet, “Delay-locked loop technique for temperature stabilisation of internal delays of CMOS FPGA devices,” *Electron. Lett.*, vol. 36, no. 14, p. 1184, 2000.

[110] Z. Li and M. J. Deen, “Towards a portable Raman spectrometer using a concave grating and a time-gated CMOS SPAD,” *Opt. Express*, vol. 22, no. 15, pp. 18736–18747, Jul. 2014.

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

[112] D. Tamborini, B. Markovic, F. Villa, and A. Tosi, “16-Channel Module Based on a Monolithic Array of Single-Photon Detectors and 10-ps Time-to-Digital Converters,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 20, no. 6, pp. 3802908:1–3802908:8, Nov. 2014.

[113] S. Cova, A. Longoni, and A. Andreoni, “Towards picosecond resolution with single-photon avalanche diodes,” *Rev. Sci. Instrum.*, vol. 52, no. 3, pp. 408–412, 1981.

[114] W. G. Oldham, R. R. Samuelson, and P. Antognetti, “Triggering phenomena in avalanche diodes,” *IEEE Trans. Electron Devices*, vol. ED-19, no. 9, pp. 1056–1060, Sep. 1972.

[115] Q. He, Y. Xu, and F. Zhao, “An accurate simulation model for single-photon avalanche diodes including important statistical effects,” *J. Semicond.*, vol. 34, no. 10, pp. 104007–1–6, Oct. 2013.

[116] A. Dalla Mora, A. Tosi, S. Tisa, and F. Zappa, "Single-Photon Avalanche Diode Model for Circuit Simulations," *IEEE Photonics Technol. Lett.*, vol. 19, no. 23, pp. 1922–1924, 2007.

[117] F. Zappa, A. Tosi, A. Dalla Mora, and S. Tisa, "SPICE modeling of single photon avalanche diodes," *Sensors Actuators A Phys.*, vol. 153, no. 2, pp. 197–204, Aug. 2009.

[118] R. Mita, G. Palumbo, and P. G. Fallica, "Accurate model for single-photon avalanche diodes," *IET Circuits, Devices Syst.*, vol. 2, no. 2, pp. 207–212, 2008.

[119] G. Giustolisi, R. Mita, and G. Palumbo, "Behavioral modeling of statistical phenomena of single-photon avalanche diodes," *Int. J. Circuit Theory Appl.*, vol. 40, no. 7, pp. 661–679, 2012.

[120] M. Anti, A. Tosi, F. Acerbi, and F. Zappa, "Modeling of afterpulsing in Single-Photon Avalanche Diodes," in *Proceedings of SPIE*, San Francisco, CA, USA, 2011, vol. 7933, pp. 79331R–1–8.

[121] M. A. Itzler, X. Jiang, and M. Entwistle, "Power law temporal dependence of InGaAs/InP SPAD afterpulsing," *J. Mod. Opt.*, vol. 59, no. 17, pp. 1472–1480, Oct. 2012.

[122] D. B. Horoshko, V. N. Chizhevsky, and S. Y. Kilin, "Full-response characterization of afterpulsing in single-photon detectors," *arXiv:1409.6752 [quant-ph]*, pp. 1–4, Sep. 2014.

[123] B. Korzh, T. Lunghi, K. Kuzmenko, G. Boso, and H. Zbinden, "Afterpulsing studies of low-noise InGaAs/InP single-photon negative-feedback avalanche diodes," *J. Mod. Opt.*, vol. 62, no. 14, pp. 1151–1157, Nov. 2015.

[124] W. J. Kindt, "Geiger Mode Avalanche Photodiode Arrays- For spatially resolved single photon counting," Ph.D. thesis, Dept. Electrical Engineering, Technology University of Delft, Delft, The Netherland, 1999.

[125] S. M. Sze and K. K.NG, *Physics of Semiconductor Devices*, 3rd ed. Hoboken, New Jersey, USA: Wiley, 2007.

[126] G. A. M. Hurkx, "On the modelling of tunnelling currents in reverse-biased P-N junctions," *Solid. State. Electron.*, vol. 32, no. 8, pp. 665–668, 1989.

[127] R. H. Haitz, "Mechanisms Contributing to the Noise Pulse Rate of Avalanche Diodes," *J. Appl. Phys.*, vol. 36, no. 10, pp. 3123–3131, 1965.

[128] G. Vincent, A. Chantre, and D. Bois, "Electric field effect on the thermal emission of traps in semiconductor junctions," *J. Appl. Phys.*, vol. 50, no. 8, pp. 5484–5487, 1979.

[129] S. R. Forrest, R. F. Leheny, R. E. Nahory, and M. A. Pollack, "$In_{0.53}Ga_{0.47}As$

photodiodes with dark current limited by generation recombination and tunneling," *Appl. Phys. Lett.*, vol. 37, no. 3, pp. 322–325, 1980.

[130] N. Tabatabaie, G. E. Stillman, R. Chin, and P. D. Dapkus, "Tunneling in the reverse dark current of GaAlAsSb avalanche photodiodes," *Appl. Phys. Lett.*, vol. 40, no. 5, pp. 415–417, 1982.

[131] G. Karve, S. Wang, F. Ma, X. Li, J. C. Campbell, R. G. Ispasoiu, D. S. Bethune, W. P. Risk, G. S. Kinsey, J. C. Boisvert, T. D. Isshiki, and R. Sudharsanan, "Origin of dark counts in $In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As$ avalanche photodiodes operated in Geiger mode," *Appl. Phys. Lett.*, vol. 86, no. 6, pp. 063505–1–3, 2005.

[132] W. J. Kindt and H. W. Van Zeijl, "Modelling and fabrication of Geiger mode avalanche photodiodes," *IEEE Trans. Nucl. Sci.*, vol. 45, no. 3, pp. 715–719, Jun. 1998.

[133] M. Ghioni, A. Giuduce, S. Cova, and F. Zappa, "High-rate quantum key distribution at short wavelength: performance analysis and evaluation of silicon single photon avalanche diodes," *J. Mod. Opt.*, vol. 50, no. 14, pp. 2251–2269, Sep. 2003.

[134] M. Ghioni, S. Cova, F. Zappa, and C. Samori, "Compact active quenching circuit for fast photon counting with avalanche photodiodes," *Rev. Sci. Instrum.*, vol. 67, no. 10, pp. 3440–3448, 1996.

[135] H. Dautet, P. Deschamps, B. Dion, A. D. Macgregor, D. Macsween, R. J. McIntyre, C. Trottier, and P. P. Webb, "Photon counting techniques with silicon avalanche photodiodes.," *Appl. Opt.*, vol. 32, no. 21, pp. 3894–3900, Jul. 1993.

[136] S. Cova, A. Lacaita, G. Ripamonti, A. Lacaita, and G. Ripamonti, "Trapping phenomena in avalanche photodiodes on nanosecond scale," *IEEE Electron Device Lett.*, vol. 12, no. 12, pp. 685–687, 1991.

[137] G. Ripamonti, F. Zappa, and S. D. Cova, "Effects of trap levels in single-photon optical time-domain reflectometry: evaluation and correction," *J. Light. Technol.*, vol. 10, no. 10, pp. 1398–1402, 1992.

[138] M. Ghioni, A. Gulinatti, I. Rech, and F. Zappa, "Progress in Silicon Single-Photon Avalanche Diodes," *IEEE J. Sel. Top. Quantum Electron.*, vol. 13, no. 4, pp. 852–862, 2007.

[139] A. C. Giudice, M. Ghioni, S. Cova, and F. Zappa, "A process and deep level evaluation tool: afterpulsing in avalanche junctions," in *33rd Conference on European Solid-State Device Research (ESSDERC)*, Estoril, Portugal, 2003, pp. 347–350.

[140] E. A. Gutiérrez-D., M. J. Deen, and C. L. Claeys, *Low temperature electronics: physics, devices, circuits, and applications*. San Diego, CA, USA: Academic Press, 2000.

[141] W. Liu, *Handbook of III-V Heterojunction Bipolar Transistors*. New York, NY, USA: Wiley, 1998.

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

[143] X. Zheng, Z. Cheng, M. J. Deen, and H. Peng, "Improving the spatial resolution in CZT detectors using charge sharing effect and transient signal analysis: Simulation study," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 808, pp. 60–70, Feb. 2016.

[144] C. S. Bamji, P. O. Connor, T. Elkhatib, S. Mehta, B. Thompson, L. A. Prather, D. Snow, O. C. Akkaya, A. Daniel, A. D. Payne, T. Perry, M. Fenton, and V. Chan, "A 0.13μm CMOS System-on-Chip for a 512x424 Time-of-Flight Image Sensor With Multi-Frequency Photo-Demodulation up to 130MHz and 2GS/s ADC," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 303–319, 2015.

[145] D. Bronzi, F. Villa, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, and W. Brockherde, "100000 Frames/s 64X32 Single-Photon Detector Array for 2-D Imaging and 3-D Ranging," *IEEE J. Sel. Top. Quantum Electron.*, vol. 20, no. 6, pp. 3804310:1–3804310:10, 2014.

[146] F. Villa, R. Lussana, D. Bronzi, S. Tisa, A. Tosi, F. Zappa, A. D. Mora, D. Contini, D. Durini, S. Weyers, and W. Brockherde, "CMOS Imager With 1024 SPADs and TDCs for Single-Photon Timing and 3-D Time-of-Flight," *IEEE J. Sel. Top. Quantum Electron.*, vol. 20, no. 6, pp. 364–373, 2014.

[147] S. P. Poland, N. Krstajić, J. Monypenny, S. Coelho, D. Tyndall, R. J. Walker, V. Devauges, J. Richardson, N. Dutton, P. Barber, D. D. Li, K. Suhling, T. Ng, R. K. Henderson, and S. M. Ameer-Beg, "A high speed multifocal multiphoton fluorescence lifetime imaging microscope for live-cell FRET imaging," *Biomed. Opt. Express*, vol. 6, no. 2, p. 277, Feb. 2015.

[148] G. Giustolisi, A. D. Grasso, and G. Palumbo, "Integrated Quenching-and-Reset Circuit for Single-Photon Avalanche Diodes," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 1, pp. 271–277, Jan. 2015.

[149] S. Chick, R. Coath, R. Sellahewa, R. Turchetta, T. Leitner, and A. Fenigstein, "Dead Time Compensation in CMOS Single Photon Avalanche Diodes With Active Quenching and External Reset," *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2725–2731, Aug. 2014.

[150] D. Bronzi, S. Tisa, F. Villa, S. Bellisai, A. Tosi, and F. Zappa, "Fast Sensing and Quenching of CMOS SPADs for Minimal Afterpulsing Effects," *IEEE Photonics Technol. Lett.*, vol. 25, no. 8, pp. 776–779, Apr. 2013.

[151] M. Liu, C. Hu, J. C. Campbell, Z. Pan, and M. M. Tashima, "Reduce Afterpulsing of Single Photon Avalanche Diodes Using Passive Quenching With Active Reset," *IEEE J. Quantum Electron.*, vol. 44, no. 5, pp. 430–434, May 2008.

[152] S. Tisa, F. Zappa, A. Tosi, and S. Cova, "Electronics for single photon avalanche diode arrays," *Sensors Actuators A Phys.*, vol. 140, no. 1, pp. 113–122, Oct. 2007.

[153] F. Zappa, S. Tisa, A. Tosi, and S. Cova, "Principles and features of single-photon avalanche diode arrays," *Sensors Actuators A Phys.*, vol. 140, no. 1, pp. 103–112, Oct. 2007.

[154] Bing-Yue Tsui, Ming-Da Wu, Tian-Choy Gan, B. Tsui, M. Wu, and T. Gan, "Impact of silicide formation on the resistance of common source/drain region," *IEEE Electron Device Lett.*, vol. 22, no. 10, pp. 463–465, Oct. 2001.

[155] K.-I. Goto, A. Fushida, J. Watanabe, T. Sukegawa, Y. Tada, T. Nakamura, T. Yamazaki, and T. Sugii, "A new leakage mechanism of Co salicide and optimized process conditions," *IEEE Trans. Electron Devices*, vol. 46, no. 1, pp. 117–124, 1999.

[156] Hi-Deok Lee and H. D. Lee, "Characterization of shallow silicided junctions for sub-quarter micron ULSI technology. Extraction of silicidation induced Schottky contact area," *IEEE Trans. Electron Devices*, vol. 47, no. 4, pp. 762–767, Apr. 2000.

[157] S.-L. Zhang and U. Smith, "Self-aligned silicides for Ohmic contacts in complementary metal–oxide–semiconductor technology: $TiSi_2$, $CoSi_2$, and NiSi," *J. Vac. Sci. Technol. A Vacuum, Surfaces, Film.*, vol. 22, no. 4, pp. 1361–1370, 2004.

[158] D. Codegoni, G. P. P. Carnevale, C. De Marco, I. Mica, and M. L. L. Polignano, "Leakage current and deep levels in CoSi2 silicided junctions," *Mater. Sci. Eng. B*, vol. 124, pp. 349–353, Dec. 2005.

[159] K. Goto, I. Fushida, J. Watanabe, T. Sukegawa, K. Kawamura, T. Yamazaki, and T. Sugii, "Leakage mechanism and optimized conditioms of Co salicide process for deep-submicron CMOS devices," in *International Electron Devices Meeting (IEDM)*, Washington, DC, USA, 1995, pp. 449–452.

[160] J. Y. Dai, Z. R. Guo, S. F. Tee, C. L. Tay, E. Er, and S. Redkar, "Formation of cobalt silicide spikes in 0.18 μm complementary metal oxide semiconductor process," *Appl. Phys. Lett.*, vol. 78, no. 20, pp. 3091–3093, 2001.

[161] K. Banerjee, C. Hu, A. Amerasekera, and J. A. Kittl, "High current effects in silicide

films for sub-0.25μm VLSI technologies," in *IEEE International Reliability Physics Symposium*, Reno, NV, USA, 1998, pp. 284–292.

[162] Shou-Gwo Wuu, Dun-Nian Yaung, Chien-Hsien Tseng, Ho-Ching Chien, C. S. Wang, Yean-Kuen Hsiao, Chin-Kung Chang, and B. J. Chang, "High performance 0.25μm CMOS color imager technology with non-silicide source/drain pixel," in *International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2000, pp. 705–708.

[163] Shou-Gwo Wuu, Ho-Ching Chien, Dun-Nian Yaung, Chien-Hsien Tseng, C. S. Wang, Chin-Kung Chang, and Yu-Kung Hsaio, "A high performance active pixel sensor with 0.18μm CMOS color imager technology," in *International Electron Devices Meeting (IEDM)*, Washington, DC, USA, 2001, pp. 24.3.1–24.3.4.

[164] M.-J. Lee and W.-Y. Choi, "Effects of Parasitic Resistance on the Performance of Silicon Avalanche Photodetectors in Standard CMOS Technology," *IEEE Electron Device Lett.*, vol. 37, no. 1, pp. 60–63, Jan. 2016.

[165] U. Akgun, A. S. Ayan, G. Aydin, F. Duru, J. Olson, and Y. Onel, "Afterpulse timing and rate investigation of three different Hamamatsu Photomultiplier Tubes," *J. Instrum.*, vol. 3, no. 01, pp. T01001–T01009, Jan. 2008.

[166] Z. Cheng, X. Zheng, D. Palubiak, M. J. Deen, and H. Peng, "A Comprehensive and Accurate Analytical SPAD Model for Circuit Simulation," *IEEE Trans. Electron Devices*, vol. 63, no. 5, pp. 1940–1948, May 2016.

[167] W. Füssel, M. Schmidt, H. Angermann, G. Mende, and H. Flietner, "Defects at the Si/SiO2 interface: Their nature and behaviour in technological processes and stress," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 377, no. 2–3, pp. 177–183, Aug. 1996.

[168] J. Albohn, W. Füssel, N. D. Sinh, K. Kliefoth, and W. Fuhs, "Capture cross sections of defect states at the Si/SiO$_2$ interface," *J. Appl. Phys.*, vol. 88, no. 2, pp. 842–849, 2000.

[169] K. Kutsuki, T. Ono, and K. Hirose, "First-principles study on electronic structure of Si/SiO$_2$ interface—Effect of interface defects on local charge density," *Sci. Technol. Adv. Mater.*, vol. 8, no. 3, pp. 204–207, Jan. 2007.

[170] K. Kotani, "Improvement of power conversion efficiency in photovoltaic-assisted UHF rectifiers by non-silicide technique applied to photovoltaic cells," *Jpn. J. Appl. Phys.*, vol. 54, no. 4S, pp. 04DE02:1–04DE02:5, Apr. 2015.

[171] G. Humer, M. Peev, C. Schaeff, S. Ramelow, M. Stipcevic, and R. Ursin, "A simple and robust method for estimating afterpulsing in single photon detectors," *J. Light. Technol.*, vol. 33, no. 14, pp. 3098–3107, 2015.

[172] E. Charbon, H.-J. Yoon, and Y. Maruyama, “A Geiger mode APD fabricated in standard 65nm CMOS technology,” in *International Electron Devices Meeting (IEDM)*, Washington, DC, USA, 2013, pp. 27.5.1–27.5.4.

[173] C. Niclass, M. Gersbach, R. Henderson, L. Grant, and E. Charbon, “A Single Photon Avalanche Diode Implemented in 130-nm CMOS Technology,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 13, no. 4, pp. 863–869, 2007.

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

[175] G. F. Burkhard, E. T. Hoke, and M. D. McGehee, “Accounting for Interference, Scattering, and Electrode Absorption to Make Accurate Internal Quantum Efficiency Measurements in Organic and Other Thin Solar Cells,” *Adv. Mater.*, vol. 22, no. 30, pp. 3293–3297, Aug. 2010.

[176] M. N. Polyanskiy, “Refractive index database,” 2016. [Online]. Available: http://refractiveindex.info.

[177] J. Richardson, R. Walker, L. Grant, D. Stoppa, F. Borghetti, E. Charbon, M. Gersbach, and R. K. Henderson, “A 32x32 50ps Resolution 10 bit Time to Digital Converter Array in 130nm CMOS for Time Correlated Imaging,” in *IEEE Custom Integrated Circuits Conference (CICC)*, San Jose, CA, USA, 2009, no. 029217, pp. 77–80.

[178] C. Niclass, K. Ito, M. Soga, H. Matsubara, I. Aoyagi, S. Kato, and M. Kagami, “Design and characterization of a 256x64-pixel single-photon imager in CMOS for a MEMS-based laser scanning time-of-flight sensor,” *Opt. Express*, vol. 20, no. 11, pp. 11863–11881, May 2012.

[179] A. Kuhn, S. Surti, J. S. Karp, P. S. Raby, K. S. Shah, A. E. Perkins, and G. Muehllehner, “Design of a lanthanum bromide detector for time-of-flight PET,” *IEEE Trans. Nucl. Sci.*, vol. 51, no. 5, pp. 2550–2557, 2004.

[180] M. Gersbach, Y. Maruyama, R. Trimananda, M. W. Fishburn, D. Stoppa, J. A. Richardson, R. Walker, R. Henderson, and E. Charbon, “A time-resolved, low-noise single-photon image sensor fabricated in deep-submicron CMOS technology,” *IEEE J. Solid-State Circuits*, vol. 47, no. 6, pp. 1394–1407, 2012.

[181] J. Guo and S. Sonkusale, “A 65 nm CMOS digital phase imager for time-resolved fluorescence imaging,” *IEEE J. Solid-State Circuits*, vol. 47, no. 7, pp. 1731–1742, 2012.

[182] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, “A 128x128 Single-Photon

Image Sensor with column-level 10-bit time-to-digital Converter Array," *IEEE J. Solid-State Circuits*, vol. 43, no. 12, pp. 2977–2989, 2008.

[183] D. E. Schwartz, E. Charbon, and K. L. Shepard, "A Single-Photon Avalanche Diode Array for Fluorescence Lifetime Imaging Microscopy," *IEEE J. Solid-State Circuits*, vol. 43, no. 11, pp. 2546–2557, 2008.

[184] H. Alhemsi, Zhiyun Li, and M. J. Deen, "Time-resolved near-infrared spectroscopic imaging systems," in *Saudi International Electronics, Communications and Photonics Conference (SIECPC)*, Riyadh, Saudi Arabia, 2013, pp. 1–6.

[185] A. S. Yousif and J. W. Haslett, "A Fine Resolution TDC Architecture for Next Generation PET Imaging," *IEEE Trans. Nucl. Sci.*, vol. 54, no. 5, pp. 1574–1582, Oct. 2007.

[186] C. Degenhardt, B. Zwaans, T. Frach, and R. de Gruyter, "Arrays of digital Silicon Photomultipliers - Intrinsic performance and application to scintillator readout," in *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, Knoxville, TN, USA, 2010, pp. 1954–1956.

[187] M. M. El-Desouki, D. Palubiak, M. J. Deen, Q. Fang, and O. Marinov, "A Novel, High-Dynamic-Range, High-Speed, and High-Sensitivity CMOS Imager Using Time-Domain Single-Photon Counting and Avalanche Photodiodes," *IEEE Sens. J.*, vol. 11, no. 4, pp. 1078–1083, 2011.

[188] X. Fang, D. Brasse, C. Hu-Guo, and Y. Hu, "Design and integration of a high accuracy multichannel analog CMOS peak detect and hold circuit for APD-based PET imaging," *IEEE Trans. Biomed. Circuits Syst.*, vol. 6, no. 2, pp. 179–187, 2012.

[189] N. Ollivier-Henry, W. Gao, X. Fang, N. A. Mbow, D. Brasse, B. Humbert, C. Hu-Guo, C. Colledani, and Y. Hu, "Design and characteristics of a multichannel front-end ASIC using current-mode CSA for small-animal PET imaging," *IEEE Trans. Biomed. Circuits Syst.*, vol. 5, no. 1, pp. 90–99, 2011.

[190] B. Markovic, A. Tosi, F. Zappa, and S. Tisa, "Smart-pixel with SPAD detector and Time-to-Digital Converter for Time-Correlated Single Photon Counting," in *23rd Annual Meeting of the IEEE Photonics Society*, Denver, CO, USA, 2010, pp. 181–182.

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

[192] D. Stoppa, F. Borghetti, J. Richardson, R. Walker, L. Grant, R. K. Henderson, M. Gersbach, E. Charbon, I. Bruno, and K. Fbk, "A 32x32-Pixel Array with In-Pixel

Photon Counting and Arrival Time Measurement in the Analog Domain," in *Proceedings of ESSCIRC*, Athens, Greece, 2009, pp. 6–9.

[193] M. Kfouri, O. Marinov, P. Quevedo, N. Faramarzpour, S. Shirani, L. W.-C. Liu, Q. Fang, and M. J. Deen, "Toward a Miniaturized Wireless Fluorescence-Based Diagnostic Imaging System," *IEEE J. Sel. Top. Quantum Electron.*, vol. 14, no. 1, pp. 226–234, 2008.

[194] M. El-Desouki, M. Jamal Deen, Q. Fang, L. Liu, F. Tse, and D. Armstrong, "CMOS Image Sensors for High Speed Applications," *Sensors*, vol. 9, no. 1, pp. 430–444, Jan. 2009.

[195] B. Jaggi, M. J. Deen, and B. Palcic, "Quantitative light microscope using a solid state detector in the primary image plane," UP patent 4845552, 1989.

[196] X. Michalet, O. H. W. Siegmund, J. V. Vallerga, P. Jelinsky, J. E. Millaud, and S. Weiss, "Detectors for single-molecule fluorescence imaging and spectroscopy," *J. Mod. Opt.*, vol. 54, no. 2–3, pp. 239–281, 2007.

[197] C. Niclass and M. Soga, "A Miniature Actively Recharged Single-Photon Detector Free of Afterpulsing Effects with 6ns Dead Time in a 0.18μm CMOS Technology," in *International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2010, pp. 340–343.

[198] A. Vila, E. Vilella, O. Alonso, and A. Dieguez, "Crosstalk-free single photon avalanche photodiodes located in a shared well," *IEEE Electron Device Lett.*, vol. 35, no. 1, pp. 99–101, 2014.

[199] E. Sciacca, a. C. Giudice, D. Sanfilippo, F. Zappa, S. Lombardo, R. Consentino, C. Di Franco, M. Ghioni, G. Fallica, G. Bonanno, S. Cova, and E. Rimini, "Silicon planar technology for single-photon optical detectors," *IEEE Trans. Electron Devices*, vol. 50, no. 4, pp. 918–925, Apr. 2003.

[200] J. Kim, Y. Seo, Y. Suh, H. Park, and J. Sim, "A 300-ms/s, 1.76-ps-resolution, 10-b asynchronous Pipelined Time-to-Digital Converter With on-Chip Digital Background Calibration in 0.13μm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 2, pp. 516–526, 2013.

[201] S. Henzler, S. Koeppe, D. Lorenz, W. Kamp, R. Kuenemund, and D. Schmitt-landsiedel, "A local passive time interpolation concept for variation-tolerant high-resolution Time-to-Digital Conversion," *IEEE J. Solid-State Circuits*, vol. 43, no. 7, pp. 1666–1676, 2008.

[202] IEEE, "IEEE Standard for Terminology and Test Methods for Analog-to-Digital

Converters," *IEEE Stand. 1241-2010*, 2000.

[203] L. H. C. Braga, M. Perenzoni, R. Walker, R. K. Henderson, D. Stoppa, L. Pancheri, and L. Gasparini, "A CMOS mini-SiPM detector with in-pixel data compression for PET applications," in *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC),* Valencia, Spain, 2011, pp. 548–552.

[204] J. Jansson, A. Mäntyniemi, and J. Kostamovaara, "A CMOS Time-to-Digital Converter With Better Than 10ps single-shot precision," *IEEE J. Solid-State Circuits*, vol. 41, no. 6, pp. 1286–1296, 2006.

[205] I. Nissinen, A. Mantyniemi, and J. Kostamovaara, "A CMOS Time-to-Digital Converter based on a Ring Oscillator for a Laser Radar," in *29th European Solid-State Circuits Conference*, Estoril, Portugal, 2003, pp. 469 – 472.

[206] H. Peng and C. S. Levin, "Design study of a high-resolution breast-dedicated PET system built from cadmium zinc telluride detectors.," *Phys. Med. Biol.*, vol. 55, no. 9, pp. 2761–2788, May 2010.

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

[208] K. Kim, Y. Kim, W. Yu, and S. Cho, "A 7 bit, 3.75ps resolution two-step time-to-digital Converter in 65 nm CMOS Using Pulse-Train Time Amplifier," *IEEE J. Solid-State Circuits*, vol. 48, no. 4, pp. 1–9, 2013.

[209] W. Liu, W. R. Li, P. Ren, C. Lin, S. Zhang, and Y. Wang, "A PVT Tolerant 10 to 500 MHz All-Digital Phase-Locked Loop With Coupled TDC and DCO," *IEEE J. Solid-State Circuits*, vol. 45, no. 2, pp. 314–321, Feb. 2010.

[210] K. Choi, S. Lee, B. Lee, and W. Choi, "A Time-to-Digital Converter Based on a Multiphase Reference Clock and a Binary Counter With a Novel Sampling Error Corrector," *IEEE Trans. Circuits Syst. II*, vol. 59, no. 3, pp. 143–147, 2012.

[211] I. Nissinen and J. Kostamovaara, "On-Chip Voltage Reference-Based Time-to-Digital Converter for Pulsed Time-of-Flight Laser Radar Measurements," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 6, pp. 1938–1948, Jun. 2009.

[212] C.-M. Hsu, M. Z. Straayer, and M. H. Perrott, "A Low-Noise Wide-BW 3.6-GHz Digital delta-sigma Fractional-N Frequency Synthesizer With a Noise-Shaping Time-to-Digital Converter and Quantization Noise Cancellation," *IEEE J. Solid-State Circuits*, vol. 43, no. 12, pp. 2776–2786, 2008.

[213] P. Lu, A. Liscidini, and P. Andreani, "A 3.6mW, 90nm CMOS Gated-Vernier Time-to-

Digital Converter With an Equivalent Resolution of 3.2 ps," *IEEE J. Solid-State Circuits*, vol. 47, no. 7, pp. 1626–1635, 2012.

[214] C. Jiang, Y. Huang, and Z. Hong, "A multi-path gated ring oscillator based time-to-digital converter in 65 nm CMOS technology," *J. Semicond.*, vol. 34, no. 3, pp. 035004:1–035004:5, Mar. 2013.

[215] P. Lu, A. Liscidini, and P. Andreani, "A 2-D GRO Vernier time-to-digital converter with large input range and small latency," *Analog Integr. Circuits Signal Process.*, vol. 76, no. 2, pp. 195–206, Jun. 2013.

[216] P. Lu, Y. Wu, and P. Andreani, "A 90nm CMOS digital PLL based on Vernier-Gated-Ring-Oscillator Time-to-Digital Converter," in *IEEE International Symposium on Circuits and Systems*, Seoul, Korea, 2012, pp. 2593–2596.

[217] C. Jiang, J. Liu, Y. Huang, and Z. Hong, "A low-noise 8.95-11GHz All-digital frequency Synthesizer with a Metastability-Free Time-to-Digital Converter and a Sleepy Counter in 65nm CMOS," in *Proceedings of ESSCIRC*, Bordeaux, France, 2012, pp. 365–368.

[218] J. Yu, F. F. Dai, and R. C. Jaeger, "A 12-Bit Vernier Ring Time-to-Digital Converter in 0.13μm CMOS Technology," *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 830–842, 2010.

[219] B. Nikolic, V. G. Oklobdzija, V. Stojanovic, W. Jia, J. K. Chiu, and M. M. Leung, "Improved Sense-Amplifier-Based Flip-Flop : Design and Measurements," *IEEE J. Solid-State Circuits*, vol. 35, no. 6, pp. 876–884, 2000.

[220] B. K. Swann, B. J. Blalock, L. G. Clonts, D. M. Binkley, J. M. Rochelle, E. Breeding, and K. M. Baldwin, "A 100-ps Time-Resolution CMOS Time-to-Digital Converter for Positron Emission Tomography Imaging Applications," *IEEE J. Solid-State Circuits*, vol. 39, no. 11, pp. 1839–1852, 2004.

[221] B. Markovic, D. Tamborini, F. Villa, S. Tisa, A. Tosi, and F. Zappa, "10 ps Resolution, 160 ns Full Scale Range and Less Than 1.5% Differential Non-Linearity Time-To-Digital Converter Module for High Performance Timing Measurements.," *Rev. Sci. Instrum.*, vol. 83, no. 7, p. 074703, Jul. 2012.

[222] M. M. Khalil, *Basic Sciences of Nuclear Medicine*. London, UK: Springer, 2011.

[223] T. K. Lewellen, "Recent developments in PET detector technology.," *Phys. Med. Biol.*, vol. 53, no. 17, pp. R287–R317, 2008.

[224] X. Zheng, Z. Cheng, M. J. Deen, and H. Peng, "Investigation of the sub-pixel spatial resolution and charge-sharing effect in CZT detectors for PET imaging," *J. Nucl. Med.*,

vol. 56, no. supplement 3, p. 1873, May 2015.

[225] X. Zheng, M. J. Deen, and H. Peng, “Performance Characteristics of CZT Detectors for PET Imaging Applications,” *ECS Trans.*, vol. 61, no. 35, pp. 7–13, 2014.

[226] PicoQuant, “LDH Series Picosecond Pulsed Diode Laser Heads,” 2016. [Online]. Available: https://www.picoquant.com/products/category/picosecond-pulsed-sources/ldh-series-picosecond-pulsed-diode-laser-heads.

[227] Z. Li, “Miniaturization of Time-Gated Raman Spectrometer with a Concave Grating and a CMOS Single Photon Avalanche Diode,” Ph.D. thesis, Dept. Biomedical Engineering, McMaster University, Hamilton, ON, Canada, 2015.

[228] S. E. Derenzo, W.-S. Choong, and W. W. Moses, “Fundamental limits of scintillation detector timing precision.,” *Phys. Med. Biol.*, vol. 59, no. 13, pp. 3261–86, 2014.

[229] S. Seifert, H. T. van Dam, and D. R. Schaart, “The lower bound on the timing resolution of scintillation detectors,” *Phys. Med. Biol.*, vol. 57, no. 7, pp. 1797–1814, 2012.

[230] Y. Qin, M. M. R. Howlader, M. J. Deen, Y. M. Haddara, and P. R. Selvaganapathy, “Polymer integration for packaging of implantable sensors,” *Sensors Actuators B Chem.*, vol. 202, pp. 758–778, Oct. 2014.

[231] Y. Fu, Y. Qin, T. Wang, S. Chen, and J. Liu, “Ultrafast Transfer of Metal-Enhanced Carbon Nanotubes at Low Temperature for Large-Scale Electronics Assembly,” *Adv. Mater.*, vol. 22, no. 44, pp. 5039–5042, Nov. 2010.

[232] Y. Qin, M. Howlader, and M. Deen, “Low-Temperature Bonding for Silicon-Based Micro-Optical Systems,” *Photonics*, vol. 2, no. 4, pp. 1164–1201, Dec. 2015.

[233] C. Jackson, L. Wall, K. O. Neill, B. Mcgarvey, D. Herbert, and A. B. Park, “Through silicon via developments for silicon photomultiplier sensors,” in *Proceedings of SPIE*, San Francisco, CA, USA, 2015, vol. 9359, pp. 1–8.

[234] Canadian Microsystem Corporation, CMRF8SF-DM 0.13μm CMOS Process Design Manual, 2016.