DESIGN OF A TIME-TO-DIGITAL CONVERTER AND MULTI-TIME-GATED SPAD ARRAYS TOWARDS BIOMEDICAL IMAGING APPLICATIONS

DESIGN OF A TIME-TO-DIGITAL CONVERTER AND MULTI-TIME-GATED SPAD ARRAYS TOWARDS BIOMEDICAL IMAGING APPLICATIONS

By
RYAN SCOTT, B. Eng.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree of Master of Applied Science

McMaster University
Hamilton, Ontario

Master of Applied Science (2021)
(Electrical and Computer Engineering)

TITLE:	Design of A Time-to-Digital Converter and Multi-Time-Gated SPAD Arrays Towards Biomedical Imaging Applications
AUTHOR:	Ryan Scott B.Eng. McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. M. Jamal Deen, Distinguished University Professor
NUMBER OF PAGES:	xviii, 137

Lay Abstract

Medical imaging plays a key role in the diagnosis of diseases like cancer, and as such, the optimized performance of medical imaging systems is a large area of research. Recently, highly sensitive photodetectors known as single-photon avalanche diodes (SPADs) were integrated with high-performance timing circuits known as time-to-digital converters (TDCs) to form digital silicon photomultipliers (dSiPMs) and SPAD imagers. DSiPMs and SPAD imagers are capable of timestamping the detection of individual photons with a very high level of accuracy in order to generate biomedical images.

This thesis focuses on the design and measurement of these sensors using standard fabrication processes with the aim of working towards high-performance medical imaging sensors at a low cost. Firstly, we review the results achieved in TDCs and SPAD-based sensors within the recent literature. Following that, we present the design and performance results of a custom-designed TDC that aims to achieve state-of-the-art performance within a small area in order to maintain low-cost and optimal integration with SPADs. Next, the design is described for an array of custom time-gated SPADs with integrated TDCs. Finally, the SPAD is characterized in two different configurations to identify sources of improvement for future design iterations.

Abstract

Digital silicon photomultipliers (dSiPMs) and single-photon avalanche diode (SPAD) imagers are optical sensing systems formed from the integration of time-to-digital converters (TDCs) with arrays of highly sensitive photodetectors known as SPADs. TDCs are high-performance mixed-signal circuits capable of timestamping events with picosecond level resolution. The digital operation mechanisms of SPADs allow for their outputs to be sent to TDCs, where the timestamps of individual photon detections are recorded. In recent years, time-resolved SPAD-based sensors have been a heavily studied topic due to their exceptional performance potential in biomedical imaging applications, including time-of-flight (ToF) positron emission tomography (PET), fluorescence lifetime imaging microscopy (FLIM), and diffuse optical tomography (DOT). This work targets the optimization of these sensors in low-cost standard complementary metal-oxide-semiconductor (CMOS) processes.

Firstly, this thesis provides a detailed review of the work accomplished in CMOS TDCs and their integration in SPAD-based sensors. Next, a feedback time amplification TDC was designed and tested in the TSMC 65 nm process that can achieve < 5 ps timing resolution in a very compact area of 0.016 mm^2. The design is then described for a multi-time-gated array of p+/n-well SPADs that aims to mitigate SPAD dark noise while providing high-speed imaging by applying shifted gate windows simultaneously to an array of SPADs. The p+/n-well SPAD is first characterized in a passive quench configuration where it demonstrated a maximum dark count rate of 44.9 kHz, 18.1% peak PDP at 420 nm, and 0.82 ns timing jitter at a 0.7 V excess bias. While the current multi-time-gated prototype is not fully functional, the measurement results for individual pixels of the multi-time-gated array showed a 3.25 ns median gate window with a 2.2x 10^{-4} dark count probability for a 0.7 V excess bias, with 440 ps timing resolution and ~1 LSB_{rms} timing jitter. Based on the results, limitations of the current design and sources for future improvement are then discussed in detail.

Acknowledgements

Firstly, I would like to express my sincere thanks and appreciation to Dr. M. Jamal Deen for supervising my research. As someone who did not initially plan to pursue graduate studies, meeting Dr. Deen in my 4th-year microelectronics course set me on a new path where I gained many personal and technical skills. My communication with other researchers was greatly improved through our regular research meetings, allowing me to feel more confident in my technical presentations and job interviews. Dr. Deen's technical and theoretical expertise in the field of integrated circuit and photodetector design gave me the chance to develop a strong background so that I can hit the ground running in my transition from research to industry. The advice, guidance, and mentorship throughout my research are greatly appreciated. I would also like to thank my committee members, Dr. Chih-Hung (James) Chen and Dr. Qiyin Fang, for taking the time to review my thesis and providing insightful comments on my research.

I also had the privilege of working with some great researchers in Dr. Deen's team during these two years of study: Wei Jiang, Sumit Majumder, Yamn Chalich, Abu Ilius Faisal, Arif Alam, Junaid Siddiqui, Mahtab Teheri, Sophini Subramaniam, and Mahdi Naghshvarianjahromi. Of my colleagues, Wei Jiang and Sumit Majumder deserve special recognition for being key mentors during my time as a graduate student. Wei Jiang's support and technical expertise for my designs and experiments helped greatly in the completion of my work. When things didn't quite go as expected, Wei took much of his own time to provide his opinions and guidance on how to proceed, and I am very thankful.

The completion of this work was also aided by the McMaster ECE technical staff. For their help in setting up our CAD tools and helping to resolve technology licensing issues, Joe Peric and Dan Manolescu were of great help. Tyler Ackland's assembly of several test boards for our designs was also very important for the timely completion of the work.

Lastly, I would like to thank Zoë Adams, my brothers Kyle and Stephen, my parents, and my grandparents for their support during this work. I dedicate my thesis to them.

Table of Contents

Lay Abstract ..... **iii**
Abstract ..... **iv**
Acknowledgements ..... **v**
Table of Contents ..... **vi**
List of Figures ..... **ix**
List of Tables ..... **xiv**
List of Abbreviations ..... **xv**
List of Symbols ..... **xvii**
Declaration of Academic Achievement ..... **xviii**
Chapter 1 Introduction **1**
1.1. Motivations for Single-Photon Biomedical Imaging 1
1.2. Components of Single-Photon Detectors 7
1.2.1. Single-Photon Avalanche Diodes 7
1.2.2. Time-to-Digital Converters 14
1.3. Research Contributions 19
1.4. Thesis Organization 21
Chapter 2 Review of CMOS Time-to-Digital Converters **23**
2.1. Fundamental Concepts 23
2.2. Results of State-of-the-Art TDCs 26
2.2.1. Vernier TDCs 26
2.2.2. Pulse Shrinking Ring TDCs 30
2.2.3. Multipath Ring Oscillator TDCs 31
2.2.4. ΔΣ TDCs 32
2.2.5. Time Amplification TDCs 33
2.3. TDCs for Biomedical Imaging Applications 36
2.3.1. TDC Topologies 38

2.3.2. TDC Sharing Schemes 42
2.3.3. Data Readout 45
2.4. Conclusions 47
Chapter 3 Time-to-Digital Converter Using Feedback Time-Amplification 49
3.1. Operating Principle 49
3.2. Circuit Design 51
3.2.1. Control Block 51
3.2.2. Digitally-Controlled Gated Delay Line 53
3.2.3. Input Pulse Generator 55
3.2.4. Sampler Array 56
3.2.5. Remainder Generation Logic 57
3.2.6. Time Amplifier 58
3.3. Measurement Results 59
3.3.1. Measurement Setup 60
3.3.2. Results and Discussion 61
3.4. Conclusions 67
Chapter 4 Design of Multi-Time-Gated SPAD Arrays 70
4.1. Operating Principle 70
4.2. SPAD Pixel Design 73
4.2.1. SPAD Structure 73
4.2.2. Front-End Circuitry 75
4.3. Multi-Purpose Delay Line 78
4.4. A 2D Multi-Time-Gated SPAD Array 80
4.5. Layout and Simulation Results 81
4.6. Conclusions 84
Chapter 5 Measurement Results of CMOS SPADs 86
5.1. Fabricated Test Chip and Printed Circuit Board 86
5.2. Passive-Quench SPADs 87
5.2.1. Breakdown Voltage 88

5.2.2. Dark Count Rate....91
5.2.3. Timing Jitter....96
5.2.4. Photon Detection Probability....100
5.3. Multi-Time-Gated SPADs....103
5.3.1. Time-Gate Window....103
5.3.2. Dark Count Probability....106
5.3.3. Quantization Performance and Jitter....108
5.4. Conclusions....110
Chapter 6 Conclusions and Future Work....114
6.1. Conclusions....114
6.2. Future Work....117
References....124

List of Figures

Figure 1-1: Illustration of some applications of dSiPMs and SPAD imagers in biomedical imaging (© 2021 IEEE). 2
Figure 1-2: Illustration of the pile-up effect. When more than one photon is detected by the same timing unit during a single measurement interval, only the first photon timestamp is recorded (© 2021 IEEE). 4
Figure 1-3: Illustration of the SPAD IV characteristics and stages of operation [9]. 7
Figure 1-4: Simplified illustration of the schematics for passive quench and active quench SPAD pixels. 8
Figure 1-5: Illustration of the main sources of dark counts in SPADs. 10
Figure 1-6: Illustration of the SPAD fill-factor, which is a limiting factor on the pixel's achievable PDE. 12
Figure 1-7: Representation of the typical timing jitter distribution for SPADs, consisting of: a Gaussian component from photons absorbed in the active area; and an exponential tail from avalanches generated by carriers diffusing from outside the active area. Note: the same data is presented in Figure 5-12. 13
Figure 1-8: Simplified schematic for a basic time-to-amplitude converter (© 2021 IEEE). 15
Figure 1-9: Comparison of the ideal and actual TDC quantization characteristics (© 2021 IEEE). 17
Figure 2-1:Illustration of a basic delay line TDC (© 2021 IEEE). 23
Figure 2-2: Illustration of a Vernier delay line TDC (© 2021 IEEE). 24
Figure 2-3: Illustration of a pulse shrinking delay line TDC. (© 2021 IEEE) 25
Figure 2-4: Diagram of a Vernier ring oscillator (VRO) TDC (© 2021 IEEE). 28
Figure 2-5: Diagram of a 2D Vernier ring oscillator (2D-VRO) TDC. The 1D array of samplers from the traditional VDL falls along the main diagonal. The 2D array reduces size

and latency by allowing comparison between all phase differences in the 2D plane (© 2021 IEEE).....29
Figure 2-6: Diagram of a pulse shrinking ring (PSR) TDC (© 2021 IEEE). 31
Figure 2-7:Diagram of a multipath gated ring oscillator (MP-GRO) TDC (© 2021 IEEE). 32
Figure 2-8: Diagram of a ΔΣ TDC (© 2021 IEEE). 33
Figure 2-9: Diagram of feedforward time amplification (TA) TDC (© 2021 IEEE). 35
Figure 2-10: Diagram of a feedback time amplification (TA) TDC. The feedback TA TDC places the TA in the feedback path and therefore requires only a single TDC stage (© 2021 IEEE).....36
Figure 2-11: Diagram of a differential delay-locked loop (DLL) interpolation TDC (© 2021 IEEE).....39
Figure 2-12: Diagram of a differential ring oscillator TDC (© 2021 IEEE). 40
Figure 2-13: Illustration of various methods of sharing TDCs between an array of SPADs (© 2021 IEEE). 43
Figure 2-14: Graphical summary of the general TDC trade-offs (© 2021 IEEE). 47
Figure 3-1: Block diagram of the proposed feedback time amplification TDC. 50
Figure 3-2: Timing diagram of the proposed feedback time amplification TDC. 51
Figure 3-3: Simplified schematic of the control block. 52
Figure 3-4: Schematic of the current-starved gated delay cell that is replicated to form the gated delay lines.....54
Figure 3-5: The digitally-controlled delay line (DCDL) biasing circuit.....55
Figure 3-6: Schematic of the input pulse generator. 56
Figure 3-7: Schematic of the sampler, consisting of an arbiter and DFF, that is replicated to sample the state of each gated delay line. 57
Figure 3-8: Schematic of the remainder generation logic.....58
Figure 3-9: Schematic of the pulse-train time amplifier. 59
Figure 3-10: Annotated layout of the complete TDC in the TSMC 65 nm CMOS process. 59

Figure 3-11: Block diagram of the measurement setup used for the TDC characterization. 61
Figure 3-12: Delay vs. bias code for the DCDL. 62
Figure 3-13: Results of the DCDL jitter for several input codes. 62
Figure 3-14: Quantization characteristics of the 10-bit TDC, determined using a statistical code density test. 64
Figure 3-15: Nonlinearity performance of the 10-bit TDC. 64
Figure 3-16: Response of the 4-bit TDC obtained by truncating the 6 LSBs. Effectively, this uses 1 of the 3 available conversion results. 65
Figure 3-17: Response of the 7-bit TDC obtained by truncating the 3 LSBs. Effectively, this uses 2 of the 3 available conversion results. 66
Figure 3-18: Comparison of the TDC precision in the 4-bit, 7-bit, and 10-bit cases. 67
Figure 4-1: Conceptual diagram of a SPAD pixel with time-gated front-end circuitry. ...70
Figure 4-2: Conceptual block diagram of the proposed 1D multi-time-gated SPAD array. 73
Figure 4-3: Conceptual timing diagram of the proposed 1D multi-time-gated SPAD array. 73
Figure 4-4: Top view of the SPAD layout in Cadence Virtuoso, and the cross-sectional view. The p+ region extends into the lesser doped p-well and pushes the STI away from the high field region. This mitigates premature edge breakdown of the junction. 74
Figure 4-5: TCAD simulation of the proposed SPAD verifying the highest field being within the planar junction. 75
Figure 4-6: Schematic of the SPAD front-end circuit. P1, P2, and P3 are timing signals generated by the pulse generation circuitry within the SPAD pixel that produces the gate window. 76
Figure 4-7: Schematic of the SPAD pulse generation circuit. Delayed replicas of the clock are tapped from a shared multi-purpose delay line located outside the SPAD pixel. The in-pixel pulse generator uses combinational logic to generate the gating signals (i.e., P1, P2, and P3). 77

Figure 4-8: Layout of the SPAD with front-end and pulse generation circuits. 77
Figure 4-9: Schematic of the shared multi-purpose delay line that drives the multi-time-gating and TDC operation. 78
Figure 4-10: Conceptual block diagram of the proposed 2D multi-time-gated SPAD array. 81
Figure 4-11: Layout of the 1D multi-time-gated SPAD array. 82
Figure 4-12: Layout of the 2D multi-time-gated SPAD array. 82
Figure 4-13: Post layout simulation verifying the correct operation of the 2D multi-time-gated SPAD array. 83
Figure 4-14: A scalable multi-time-gated architecture that shares a fine interpolating TDC between several smaller arrays. 85
Figure 5-1: Chip-level layout that was fabricated in the TSMC 65 nm CMOS process. ..86
Figure 5-2: Photograph of the test PCB used to assess the functionality and performance of the design. 87
Figure 5-3: Schematic and layout of the passively quenched p+/n-well SPAD in the TSMC 65 nm process. 88
Figure 5-4: Diagram of the experimental setup for the breakdown voltage measurements. 89
Figure 5-5: Results of the breakdown voltage measurement at different temperatures for 4 SPADs. 90
Figure 5-6: Diagram of the experimental setup for the excess voltage and temperature-dependent dark count rate measurements. 91
Figure 5-7: Dark count rate vs. excess voltage for different temperatures. 93
Figure 5-8: A sample of the interarrival time distributions for a 0.3 V excess bias at -30 °C, 0 °C, and 30 °C. 94
Figure 5-9: The Arrhenius plot obtained from the temperature dependent DCR measurement. 96
Figure 5-10: Experimental setup used for the timing jitter measurement of the passively quenched SPAD. 97

Figure 5-11: Timing jitter histograms for excess biases of 0.3 V, 0.5 V, and 0.7 V. 98
Figure 5-12: Illustration of the timing jitter histogram components obtained with 0.3 V excess bias. 99
Figure 5-13: Experimental setup for the PDE measurement. 101
Figure 5-14: Results of the PDE measurement for a range of wavelengths. 102
Figure 5-15: Measurement results for the peak PDP and DCR variation with excess voltage on a single plot. 103
Figure 5-16: Experimental setup used to determine the time-gate windows and the system timing performance. 104
Figure 5-17: Results of the time-gate window measurement for 3 time-gated pixels 105
Figure 5-18: Experimental setup for the dark count probability per gate window measurement. 107
Figure 5-19: Experimental results of the dark count probability per gate window. 108
Figure 5-20: Illustration of the achieved timing performance for individual pixels in the multi-time-gated array of SPADs. 109
Figure 5-21: Illustration of the constant time-gated pixel output resulting from the insufficient pre-charge during the P1 phase of the gate window or leakage through the reset MOSFET during the gate window. 112
Figure 6-1: Illustration of a 3D pixel structure for a SPAD-based sensor. (© 2021 IEEE) 120

List of Tables

Table 1-1: Typical requirements for PET, FLIM, and DOT/NIROT imaging systems.....5
Table 2-1: Results of State-of-the-Art TDCs (© 2021 IEEE). 27
Table 2-2: Summary of TDC integrated with SPADs to form dSiPMs and SPAD imagers (© 2021 IEEE). 37
Table 3-1: Comparison of TDC performance to published works.....68
Table 4-1: Summary of the results of time-gated SPAD arrays. 71

List of Abbreviations

2DV	Two-Dimensional Vernier
2DV-GRO	Two-Dimensional Vernier Gated Ring Oscillator
ADC	Analog-to-Digital Converter
AP	Afterpulsing Probability
AQR	Active Quench and Reset
ASIC	Application Specific Integrated Circuit
BLUE	Best Linear Unbiased Estimator
BTBT	Band-to-Band Tunneling
CDF	Cumulative Distribution Function
CMM	Center-of-Mass Method
CMOS	Complementary Metal-Oxide-Semiconductor
DCDL	Digitally-Controlled Delay Line
DCP	Dark Count Probability
DCR	Dark Count Rate
DEMUX	Demultiplexer
DFF	D Flip-Flop
DL	Delay Line
DLL	Delay-Locked Loop
DNL	Differential Nonlinearity
DOT	Diffuse Optical Tomography
DR	Dynamic Range
dSiPM	Digital Silicon Photomultiplier
DTC	Digital-to-Time Converter
DToF	Distribution Time-of-Flight
FCS	Fluorescence Correlation Spectroscopy
FF	Fill Factor
FLIM	Fluorescence Lifetime Imaging Microscopy
fNIRS	Functional Near-Infrared Spectroscopy
FoM	Figure-of-Merit
FPGA	Field-Programmable Gate Array
FRET	Förster Resonance Energy Transfer
FWHM	Full-Width at Half-Maximum

GDL	Gated Delay Line
IAT	Inter-Arrival Time
INL	Integral Nonlinearity
IRF	Instrument Response Function
LSB	Least Significant Bit
MP-GRO	Multipath Gated Ring Oscillator
MSB	Most Significant Bit
MUX	Multiplexer
NIROT	Near Infrared Optical Tomography
PCB	Printed Circuit Board
PDE	Photon Detection Efficiency
PET	Positron Emission Tomography
PH	Partial Histogramming
PQR	Passive Quench and Reset
PS	Pulse Shrinking
PSR	Pulse Shrinking Ring
PVT	Process, Voltage, and Temperature
RO	Ring Oscillator
SiPD	Silicon Photodetector
SPAD	Single-Photon Avalanche Diode
TA	Time Amplifier
TAC	Time-to-Amplitude Converter
TAT	Trap-Assisted Tunneling
TCSPC	Time-Correlated Single Photon Counting
TDC	Time-to-Digital Converter
TG	Time-Gated
ToF	Time-of-Flight
TSPC	True Single-Phase Clock
TSV	Through-Silicon Via
V-GRO	Vernier Gated Ring Oscillator
VDL	Vernier Delay Line
VRO	Vernier Ring Oscillator
VTC	Voltage-to-Time Converter

List of Symbols

SNR_{ToF}	Signal-to-noise ratio for a time-of-flight PET measurement [dB]
$SNR_{Non-ToF}$	Signal-to-noise ratio for a non-time-of-flight PET measurement [dB]
D	Diameter of a PET ring [m]
c	Speed of light [m/s]
Δt	Timing difference between photon detections [s]
G	Sensitivity gain from time-of-flight PET measurements
DCR	Dark count rate of a SPAD [Hz]
DCR_0	Primary dark count rate of a SPAD without afterpulsing [Hz]
N_{linear}	Number of linear bits for a time-to-digital converter
b	Number of bits for a time-to-digital converter
INL	Integral nonlinearity [LSB or ps]
F_s	Sampling rate [Hz]
σ	Precision [LSB or ps]
LSB	Resolution of a time-to-digital converter [ps]
τ_i	Width of the i^{th} step on a TDC quantization characteristics [ps]
DR	Dynamic range [ns or ps]
N_i	Number of counts for i^{th} TDC code from a statistical code density test
N_{total}	Total number of counts from a statistical code density test
$H(n)$	Cumulative distribution function obtained from a code density test
T	Absolute temperature [K]
E_A	Activation energy [eV]
k	Boltzmann's constant [eV/K]
Φ_{IN}	Rate of incident photons on the SPAD [Hz]
P_{SiPD}	Optical power measured by the silicon photodetector [W]
λ	Wavelength of incident light [nm]
h	Planck's constant [J·s]
A_{SPAD}	Active area of the single-photon avalanche diode [μm^2]
A_{SiPD}	Active area of the calibrated silicon photodetector [cm^2]
DCR_{eff}	Effective dark count rate for time gated SPAD [Hz]
T_{ON}	Gate window width for a time gated SPAD [ns]

Declaration of Academic Achievement

This thesis was written by Ryan Scott under the supervision, guidance and mentorship of Dr. M. Jamal Deen from McMaster University.

- **Chapters 1 and 2:** I present an overview of fundamental SPAD principles, and a detailed review of standalone TDCs, and TDCs integrated in dSiPMs and SPAD imagers. The results from the research are summarized.
- **Chapter 3:** I designed a TDC using feedback time-amplification in the TSMC 65 nm process. I designed a PCB and ran detailed measurements to extract the performance of the fabricated TDC. Tyler Ackland was responsible for assembling the PCB.
- **Chapter 4:** I designed 1D and 2D multi-time-gated SPAD arrays in the TSMC 65 nm process. These designs utilized a custom p+/n-well SPAD.
- **Chapter 5:** I designed a PCB for testing the p+/n-well SPAD in a passive quench configuration and multi-time-gated SPAD arrays. I ran measurements to characterize these designs in terms of their most important performance metrics. Tyler Ackland was responsible for assembling the PCB.
- **Chapter 6:** Based on the literature review and design experience, I discussed several key research challenges for the coming years.

Chapter 1
Introduction

1.1. Motivations for Single-Photon Biomedical Imaging

Digital silicon photomultipliers (dSiPMs) and single-photon avalanche diode (SPAD) imagers (here, we use the umbrella term "SPAD-based sensors") are optical sensing systems formed from the integration of time-to-digital converters (TDCs) with arrays of highly sensitive photodetectors known as SPADs. TDCs are high-performance mixed-signal circuits capable of timestamping events with sub-gate delay resolution. The digital operation mechanisms of SPADs allow for their outputs to be sent to TDCs, where the timestamps of individual photon detections are recorded. Since SPADs are capable of detecting the lowest levels of light, and modern TDCs can obtain exceptional timing resolution in the range of picoseconds, SPAD-based sensors have found several uses in biomedical imaging applications, including time-of-flight (ToF) positron emission tomography (PET) [1], [2], fluorescence lifetime imaging microscopy (FLIM) [3], [4], and diffuse optical tomography (DOT) [5]–[7]. In this section, we will provide a brief overview of these applications, as well as the operation principles of single-photon detectors to motivate the remainder of this research.

A simplified illustration for a digital PET system using SPAD-based sensors is shown in Figure 1-1, in which a ring structure is made from multiple PET detectors. Each PET detector is generally comprised of three parts: scintillators, which convert the high-energy gamma rays into visible light; SPADs, which are used to convert light signals into electrical signals; and TDCs, which perform the timestamping of the photon events. To perform a PET scan, a radiotracer is first injected into the subject [8]. The decay of the radiotracer causes annihilation events that generate pairs of high-energy gamma rays with ~180 degrees of separation. The high-energy photons can then be converted by the scintillators

*Part of this work was accepted for publication in IEEE Reviews in Biomedical Engineering as: R. Scott, W. Jiang, M. J. Deen, CMOS Time-to-Digital Converters for Biomedical Imaging Applications.

to lower energy photons that are detectable by the SPADs. Therefore, the sensor can extract the position, energy, and timing information from the gamma events to reconstruct a biological image.

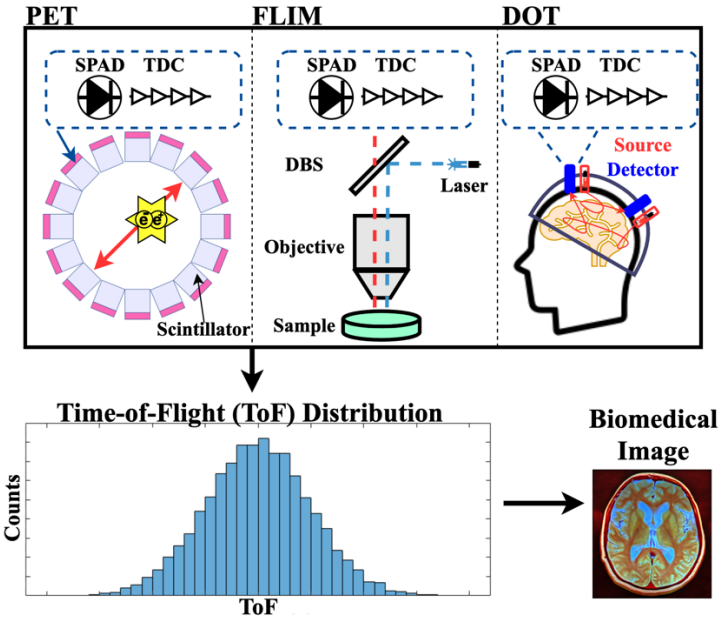

Figure 1-1: Illustration of some applications of dSiPMs and SPAD imagers in biomedical imaging (© 2021 IEEE).

In contrast to traditional PET imaging, ToF PET uses the collected timing information to center the distribution on the true location where the annihilation event took place, as opposed to being uniformly distributed along the line-of-response [9], [10]. The resolution of the detector has a direct influence on the sensitivity of the system, and as a result, the signal-to-noise ratio of the reconstructed images. This is clearly seen from equations (1-1) and (1-2), where D is the diameter of the PET ring, c is the speed of light, Δt is the measured time difference, and G is the effective sensitivity gain [9]. Here, the minimum detectable timing difference is limited by the timing resolution of the photodetector sensor. The timing resolution is largely determined by the scintillator and SPAD timing jitter [11]. With improvements to scintillator crystals and SPAD performance, finer resolution TDCs are needed to allow for further improvements in PET imaging systems.

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

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

For a complete PET detector, it is important to note that the timing, position, and energy information are all required from the system. An estimation of energy can be given by counting the total number of photons that are received during a gamma event. This can be accomplished by using counters integrated with SPAD arrays that track photon counts and are read out with the timing and position data. Usually, only the timing of the first photons of a gamma event is required for ToF PET, and the SPAD-based sensors can use fewer TDCs but benefit greatly from improvements in the resolution. It is worth noting that the first SPAD triggers are not always from photons generated from the true gamma event since SPADs can experience dark counts. Due to the false triggering from dark counts, the first-photon approximation inevitably leads to some timing errors and reduces the triggering efficiency [12].

FLIM is a method of determining the fluorescent lifetime of a sample that can be used to construct biological images (as shown in Figure 1-1). An experiment is performed by exciting the sample with a series of synchronized laser pulses and measuring the time it takes to detect a fluorescent photon back at the detector. After the excitation source is removed, the fluorescence intensity of the sample will decrease over time at a rate corresponding to its fluorescent lifetime, which is determined by the properties of the sample material [13]. Over the course of an experiment, many measurements of the fluorescent photons are made on the sample to determine the fluorescent intensity against the decay time [14]. This dataset can then be fit to an exponential function, where the time constant gives an estimate of the fluorescent lifetime [15]. Major advantages of FLIM are that living samples and their environments can be measured in real-time, and the fluorescence lifetime is largely independent of the concentration of the fluorophore within the sample or the intensity of the incident laser [15].

While there may be advantages to very high-resolution TDCs in FLIM, within much of the literature, 50 – 100 ps resolution was determined to be adequate for typical fluorescence lifetimes [16]–[18]. A key design consideration for FLIM is maintaining high throughput and avoiding pile-up. Pile-up is a phenomenon that occurs when more than one photon arrives in the same timing circuit during a measurement interval, as shown in Figure 1-2. As pile-up occurs, only the first events within a measurement cycle are timestamped, and later events are missed. This causes a skew in the distribution of photon arrivals towards shorter times and results in the system underestimating that lifetime [19]. For this reason, multi-channel systems that employ many TDCs that work concurrently can help to reduce the effect of pile-up, as events from each channel can be combined to detect multiple photons within an excitation period. An additional source of nonideality is the detector's counting loss which results from the dead time of the SPAD and timing electronics. This effect is largely dominated by the timing electronics. Therefore, compact TDC structures with high throughput are preferred for FLIM as they can help in systems aiming to prevent pile-up. Additionally, low power consumption and an easily repeatable structure that can minimize the timing skew are essential as the TDC may be replicated many times across the chip.

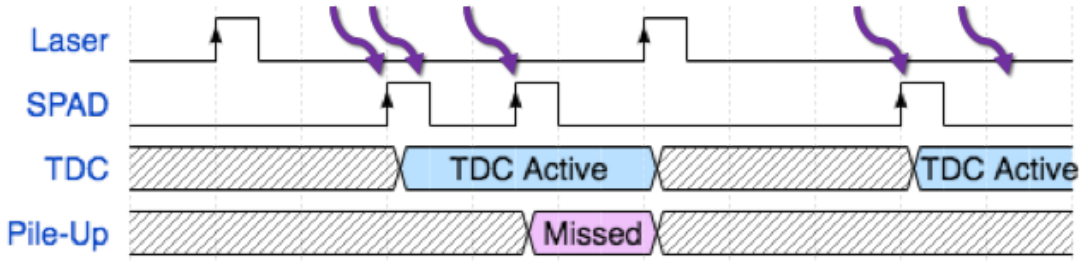

Figure 1-2: Illustration of the pile-up effect. When more than one photon is detected by the same timing unit during a single measurement interval, only the first photon timestamp is recorded (© 2021 IEEE).

Figure 1-1 also shows the application of DOT, where a pulsed laser source is used to shine red or near-infrared light (i.e., NIROT) through a target media. Scattering occurs within the target, and the re-emitted photons can be detected by photodetectors configured in either a transmittance or reflectance geometry. The delay between incident photons and photon absorptions are timestamped by a SPAD-based sensor in order to create a histogram

of the instrument response function (IRF) and the distribution time-of-flight (DToF) [6]. The histograms can then be interpreted by image reconstruction algorithms to produce a biomedical image. Specifically in NIROT, oxygenation values of tissue can be determined by recording photon reflection and scattering within the 650 nm and 850 nm wavelengths [20]. Like the FLIM application, NIROT benefits largely from miniaturized TDCs. These detectors must also be highly sensitive, as weak photon events need to be captured for the highest accuracy. Additionally, a large number of TDCs is desired in order to obtain a high spatial resolution, high-throughput, and produce images at a high framerate in order to avoid movement artifacts from the subject, or tissue variations over the duration of a measurement [20].

Table 1-1: Typical requirements for PET, FLIM, and DOT/NIROT imaging systems.

Parameter	**PET**	**FLIM**	**DOT/NIROT**
Wavelength (nm)	420	250 – 750**	> 600
Timing Resolution (ps)	< 400*	50 - 100	< 50
Timing Range (ns)	~10	50 - 5000	< 100
Time-Gate Compatible	No	Yes	Yes
# of SPADs Per TDC	> 100	< 10	< 10
Direct Patient Contact	No	No	Yes

*Resolution is worsened due to the jitter of the scintillator crystal [21].
**Wavelength range for typical fluorophores obtained from Ref. [22].
Note: Other parameters are estimated from the current published results in Table 2-2.

While PET, FLIM and DOT/NIROT all benefit from improvements to SPAD-based sensor technology, their specific requirements have some differences which are shown in Table 1-1. For PET, typical LYSO scintillators exhibit a peak emission at a wavelength of 420 nm, and therefore the SPADs should aim to have a peak PDP at this wavelength to ensure the highest sensitivity to low-light events. The scintillator is also the largest source of jitter in current PET systems, which can be up to 400 ps. Therefore, the timing requirements for the SPADs and TDCs are reduced, as the scintillator will dominate the jitter performance. Due to the sparseness of events over the detector array in PET, these sensors have commonly been designed to share a TDC by more than 100 SPADs. This allows them to achieve a higher PDE while maintaining SPAD and TDC timing performance that is better than the scintillator jitter.

In FLIM, the wavelength that needs to be detected by the SPAD-based sensor is largely dependent on the fluorophore that is being measured, and may range as far as 250 – 750 nm [22]. Typical single-SPAD designs often have a sharp peak in their PDP response against different wavelengths. If it is desirable to make a design suitable to a variety of fluorophores, multi-junction designs could potentially be used to achieve a wider optical bandwidth for the PDP due to the different depths of the junctions [23]. In addition to the variation in fluorescent wavelengths, the variation in fluorescent lifetimes means that the timing performance in FLIM is heavily dependent on the fluorophore as well. To meet the requirements of a large range of samples, SPAD-based sensors for FLIM have been commonly designed to have resolutions in the range of 50-100 ps, with timing ranges varying from 50 ns – 5 μs.

In DOT/NIROT, the SPADs need to be designed to absorb light of a longer wavelength (> 600 nm), and as such, deeper junctions such as the p-well/deep n-well or deep n-well/p-substrate should be used. In current designs, ~50 ps timing resolution has been commonly achieved for the DOT application, but with < 10 ps timing performance being desirable [6]. Due to the stricter timing requirements as compared to PET, both FLIM and DOT commonly share less than 10 SPADs to a given TDC which can help in optimizing the single-photon timing resolution. In DOT, the large number of TDCs also helps in obtaining images at a higher speed, which can reduce motion artifacts in the measurement since DOT requires direct patient contact. Lastly, both FLIM and DOT operate using a synchronous pulsed laser as excitation sources, allowing them to support time-gated operation which can improve the signal-to-noise ratio of the sensors' outputs. In DOT, this laser will typically operate from tens to hundreds of megahertz, and therefore the TDC dynamic range is generally < 100 ns (i.e., designed to match the period of the laser).

1.2. Components of Single-Photon Detectors

1.2.1. Single-Photon Avalanche Diodes

A. Operation Principle

SPADs are reverse biased p-n junctions that operate beyond their breakdown voltages in what is known as Geiger mode [24]–[30]. In Geiger mode, the large reverse bias of the SPAD makes the electric field across the depletion region so strong that the injection of even a single free carrier can initiate a self-sustaining avalanche as a result of charge multiplication from impact ionization [31]–[36]. From a high-level perspective, the SPAD can effectively act as a switch that probabilistically conducts a large reverse current upon the detection of individual photons, resulting in a digital output pulse. Due to the existence of the quench circuit, this large current is quickly reduced, and the SPAD bias is restored through the reset circuit to its initial operating conditions for subsequent photon detections. The digital operation mechanisms of SPADs allow for their outputs to be sent to TDCs, where the timestamps of individual photon detections are recorded. An illustration of the SPAD operation is shown in Figure 1-3.

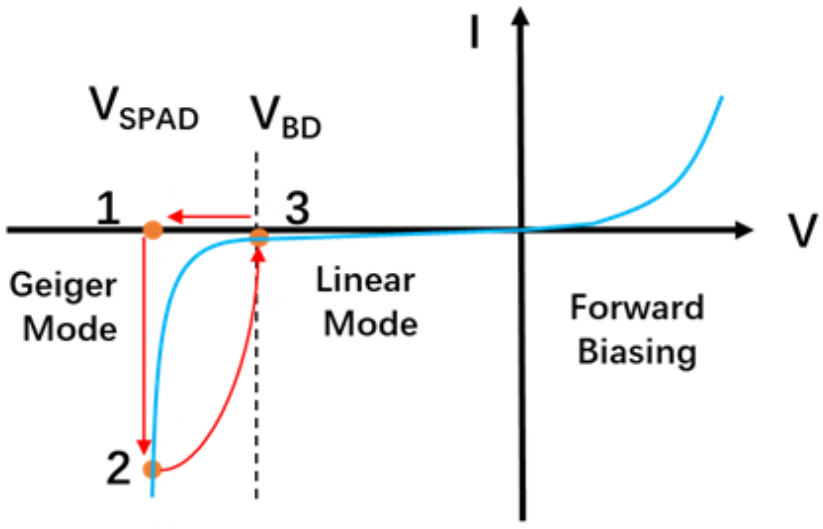

Figure 1-3: Illustration of the SPAD IV characteristics and stages of operation [9].

To achieve their digital operation, SPADs must operate in combination with a quenching and reset circuit. Quench and reset circuits are commonly divided into two categories shown in Figure 1-4: passive quench and reset (PQR), and active quench and reset (AQR). In the PQR configuration, the SPAD is charged beyond breakdown through a large resistor and initially conducts only a negligible reverse saturation current. Upon detection of a photon, the SPAD will turn on and conduct a large reverse current. This current will cause a voltage drop across the quench resistor, reducing the SPAD bias below breakdown to stop the avalanche. The SPAD is then recharged through the same quench resistor to prepare for the next photon detection. Although it has a simple structure and can achieve the highest fill-factors, the PQR configuration suffers from a long dead time due to the large RC time constant during the recharge phase.

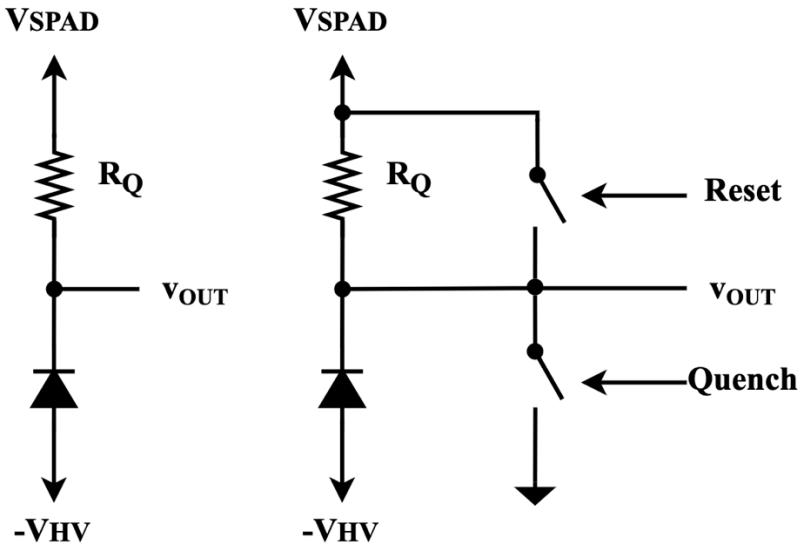

Figure 1-4: Simplified illustration of the schematics for passive quench and active quench SPAD pixels.

This led to the development of the AQR configuration. Here, the SPAD is initially biased beyond breakdown and similarly will conduct a large reverse current through the quench resistor upon detection of a photon. However, after detection of an avalanche, the SPAD is quickly quenched through a parallel quenching switch (i.e., commonly a MOSFET) such that it can be biased below breakdown very quickly. Quenching the avalanche as fast as possible is desirable as it reduces the number of charge carriers that

can become trapped and later released to generate afterpulses (the afterpulsing phenomena will be described in the next subsection). Additionally, after a delay from the quench phase, the SPAD can be reset very quickly through a reset switch in parallel with the quench resistor, bypassing the large RC time constant such that the AQR SPAD can operate at a higher speed.

B. Performance Metrics

Breakdown Voltage: A fundamental performance characteristic of the SPAD is the breakdown voltage. Due to increased phonon scattering at higher temperatures, the breakdown voltage is increased because it is more difficult for charge carriers to reach the avalanche energy threshold [9], [37]. As all the performance characteristics of the SPAD are dependent on the excess voltage, it is important to measure the temperature dependence of the breakdown voltage to ensure consistency when taking temperature-dependent measurements. Note that since the electric field through the depletion region is not constant, and the ionization rate is a strong function of the electric field, the breakdown voltage is most often determined experimentally [38].

Dark Count Rate: Since SPADs have such a high gain that even a single free carrier injected into the depletion region can trigger an avalanche, they are also very susceptible to false avalanches known as dark counts. The number of avalanches per second when the SPAD is not exposed to any light is known as the dark count rate (DCR), and is given in the units of Hz. These dark counts can occur because of multiple carrier generation processes illustrated in Figure 1-5. Due to thermal excitation, minority carriers in the bulk region may move into the depletion region by diffusion and trigger an avalanche. However, in the bulk regions the recombination rates for minority carriers are high, and thus, diffusion of carriers is not a dominating noise mechanism in SPADs [27]. Due to impurities associated with fabricating a SPAD in the complementary metal-oxide-semiconductor (CMOS) process, there will be a significant number of forbidden energy levels in the bandgap that may trap and later release charges to increase the overall DCR. Processes that have significant effects on the total DCR include trap-assisted thermal generation, trap-

assisted tunneling generation (TAT), and band-to-band tunneling (BTBT). It should be noted that direct band-to-band thermal generation is dominated by trap-assisted thermal generation since carriers stuck in the traps have a smaller energy barrier to overcome to trigger an avalanche. Recently, the random telegraph signal phenomenon was observed in SPADs and used to estimate the total defect size, which showed a positive correlation with the total DCR [39].

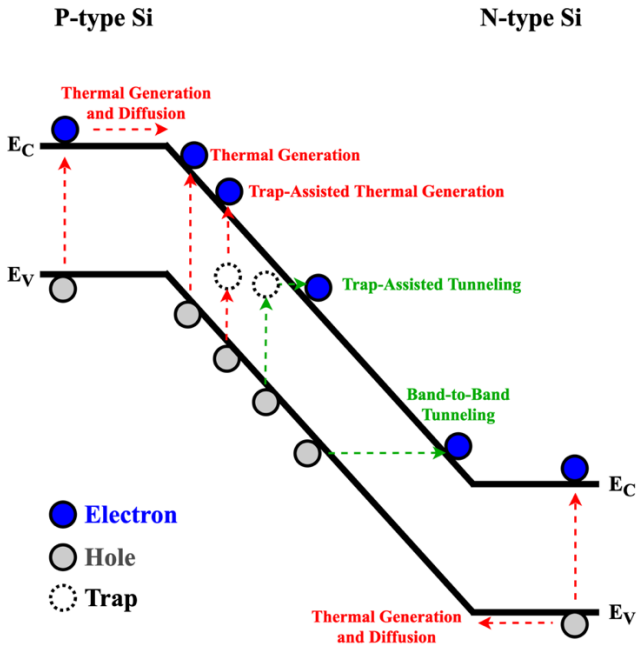

Figure 1-5: Illustration of the main sources of dark counts in SPADs.

Afterpulsing Probability: Another phenomenon that contributes to the overall DCR of the SPAD is the afterpulsing probability (AP). During an avalanche, free carriers will populate the energy traps caused by material defects and impurities. The energy traps have finite lifetimes and statistically can release a trapped carrier at any moment. While recovering from a previous avalanche, if the SPAD is re-biased above the breakdown voltage to prepare for photon detection before all the traps were emptied of charge carriers, then a carrier can be released and trigger an avalanche known as an afterpulse. Afterpulses can also be initiated by charges stored on the parasitic capacitances of the SPAD nodes during a previous avalanche. Therefore, it is important to minimize the capacitance of the front-end circuitry of the SPAD. The total DCR of the SPAD can be given as:

$$DCR = \frac{DCR_0}{1-AP}, \quad (1\text{-}3)$$

where DCR_0 is the DCR without the effect of afterpulsing and is found in practice by waiting for the traps to depopulate after a given avalanche in what is known as the hold-off time [40]. With a large enough hold-off time, it can be nearly ensured that a secondary avalanche will not occur, and thus the DCR can be measured without afterpulsing. However, it should be noted that when using a SPAD for photon detection in a real application, the hold-off time cannot be set arbitrarily large because it will lower the count rate of the device [31]. If the count rate is too low, then some valid pulses may be missed, and the device's performance is decreased. Therefore, for a specific application, there may be an optimal balance between the count rate and AP.

Photon Detection Efficiency: The photon detection efficiency (PDE) of a SPAD is calculated as the product of the geometric fill-factor (FF) and the photon detection probability (PDP). It is defined as the ratio of the number of detected photons to the number of incident photons. There are many "obstacles" that must be overcome by a photon for it to be detected by the SPAD. The photon must pass through the thick passivation layer, which is placed on the top of the chip at the end of fabrication to protect the device from contaminants, as well as through several layers of dielectric that are not optimized to minimize reflections in standard CMOS processes. Additionally, it is not enough for a photon to make it through all these layers because it also needs to be absorbed in the active area of the device. Furthermore, the more in-pixel electronics that are integrated with a SPAD, the lower the fill factor. A low fill factor is a major concern when SPAD arrays are integrated with in-pixel TDCs. In the example shown in Figure 1-6, the SPAD pixel achieves a fill factor of 25%. Therefore, the PDE of the SPAD cannot exceed 25%, even if every photon that reaches the active area generates an avalanche.

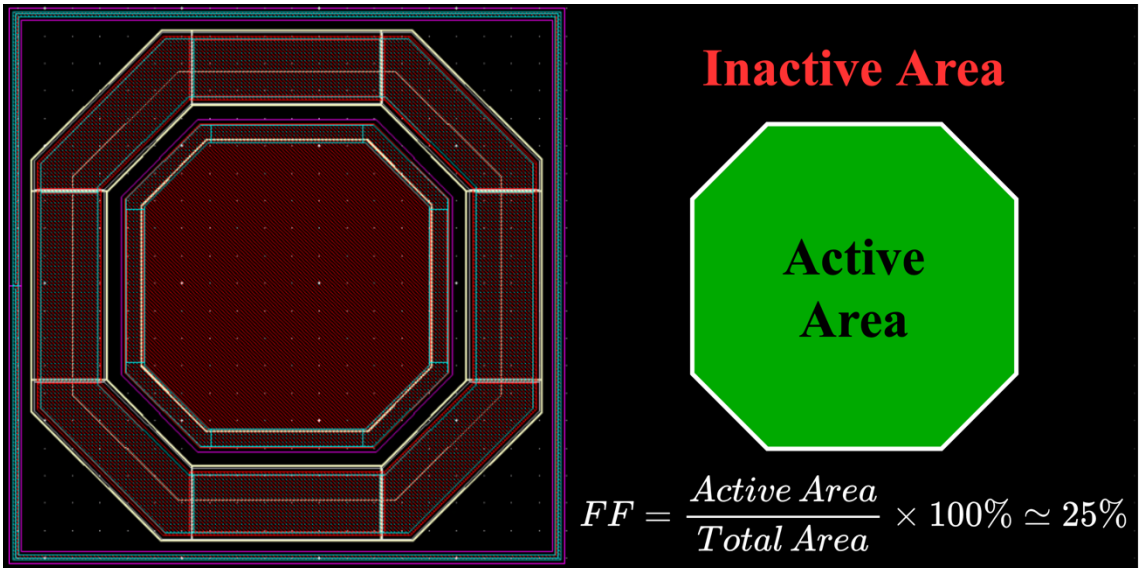

Figure 1-6: Illustration of the SPAD fill-factor, which is a limiting factor on the pixel's achievable PDE.

Timing Jitter: There is a statistical variation in the delay between the photon absorption in the SPAD and the time of the output pulse. This variation is known as the timing jitter. The timing jitter represents the uncertainty in the actual photon absorption time and can be characterized by the full-width at half-maximum (FWHM) or standard deviation of the distribution of the delays of output pulses. A characteristic example of the temporal distribution behaviour of SPADs is shown in Figure 1-7.

The temporal distribution of SPAD pulses contains both a Gaussian and exponential component [34]. The fast Gaussian peak comes from the pulses generated by avalanches caused by photons absorbed in the depletion region, where the width of the peak depends on the avalanche build-up time. The exponential tail component of the distribution is a result of the diffusion of minority carriers from the neutral regions into the depletion region, where they eventually trigger avalanches. As SPADs are often implemented in less advanced processes (i.e., 180 nm or 350 nm CMOS), the SPAD is often the dominating source of jitter when integrated with a TDC. Optimizing SPAD designs in more advanced processes presents a potential solution. However, challenges arise in terms of higher DCR due to tunneling and reduced PDE from thinner depletion regions.

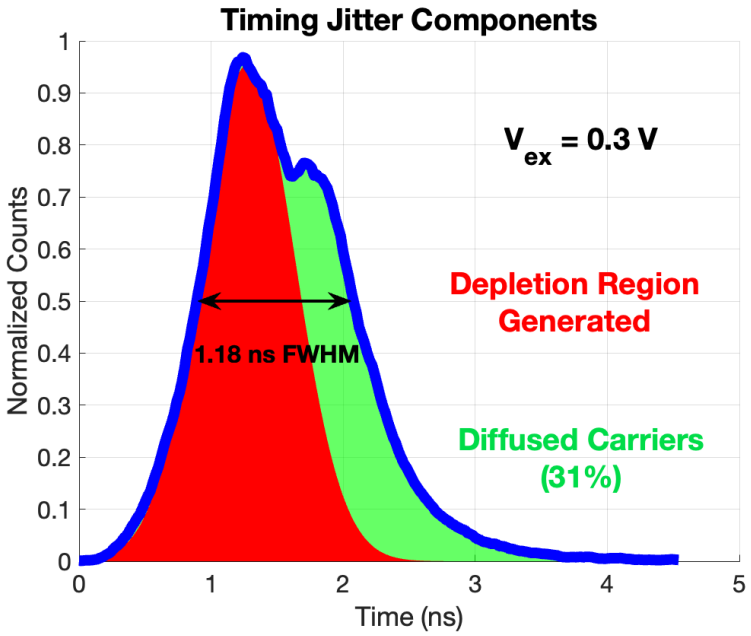

Figure 1-7: Representation of the typical timing jitter distribution for SPADs, consisting of: a Gaussian component from photons absorbed in the active area; and an exponential tail from avalanches generated by carriers diffusing from outside the active area. Note: the same data is presented in Figure 5-12.

Dead Time: As the output of a SPAD is a digital pulse, the absorption of a single photon and the absorption of many photons are indistinguishable. As such, when a given photon triggers an avalanche, not until the avalanche is quenched and the SPAD is re-biased above breakdown can any incident photons be detected. The delay between the arrivals of the photons that trigger an initial pulse to the time in which the SPAD can detect another incident photon is known as the dead-time. The dead-time is the period associated with the maximum frequency of operation of the device, known as the counting rate. As discussed previously, the dead-time of the SPAD can be increased purposefully with the intention of allowing trapped carriers to depopulate from the forbidden energy levels after an avalanche. This can reduce the AP, and by extension, the DCR of the device. The trade-off between the AP and the counting rate is optimized by careful design of the SPAD quenching circuits [41].

1.2.2. Time-to-Digital Converters

A. Operation Principle

TDCs are responsible for determining the length of a time interval between two pulses. As such, the inputs of the TDC are two digital pulses, where the two rising edges denote the start and end of the time interval to be measured, respectively. These signals are generally referred to as *start* and *stop*. In practical SPAD-based sensor applications, one of either the *start* or *stop* signals is a global reference clock, while the other is the output of a SPAD or a SPAD array that was compressed to a single output. Since the output of the SPAD can occur at any point in time, the time intervals between the rising edges form a continuous range of values. The output of the TDC is a series of bits that digitize this continuous range to form the output code, which is analogous to the operation of an analog-to-digital converter (ADC) for analog voltage discretization.

Due to the similarity of their operating principle, early TDCs employed the use of ADCs after first converting the time interval to an equivalent voltage using a time-to-amplitude converter (TAC). A simple implementation of a TAC could be formed according to Figure 1-8. In this implementation, the time interval to be measured is first converted to a pulse width. This input pulse width turns on the NMOS (i.e., M_{IN}) and allows the constant current source to linearly charge the capacitor for a time equivalent to the input pulse width. This results in a unique voltage level corresponding to each time interval within the range. At the end of the conversion, the voltage across the capacitor is discretized using an ADC and is then discharged quickly through a reset transistor. While the TACs can achieve stronger linearity performance and picosecond resolution [42]–[49], digital methods of time interval measurement are often preferred for array designs due to their ease of implementation and scaling with standard digital CMOS, greater insensitivities to process, voltage and temperature (PVT) variations, lower power consumption, and the reduced effect of noise [16].

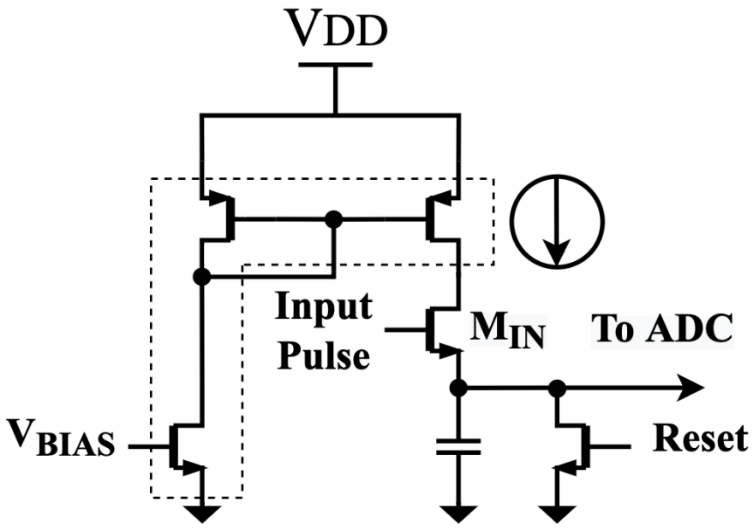

Figure 1-8: Simplified schematic for a basic time-to-amplitude converter (© 2021 IEEE).

B. Performance Metrics

Resolution: The resolution of a TDC is represented by the least significant bit (LSB) of the output word and is the minimum time difference that the TDC can differentiate. If a TDC could have an arbitrarily high resolution, then the quantization characteristics (see Figure 1-9) would be a straight line with a constant slope. However, due to the finite number of bits in the output word, a continuous range of input time intervals are represented by the same output word, giving the TDC quantization characteristics a staircase shape where the average width of the steps gives the resolution. It should be noted that the resolution is defined based on the minimum distinguishable time difference given a large number of measurements. This contrasts with the precision which represents the jitter in a single measurement. In recent works, high-performance TDCs achieved resolutions below 10 ps, with some published TDC results having a resolution in the sub-picosecond range [50]–[54].

Dynamic Range: The dynamic range (DR) is the maximum input time interval that can be converted to an accurate output word. The DR should be considered early in the design process, as it is largely dependent on the converter topology. The DR requirements also vary strongly from application to application. As an example, the required range for FLIM applications may vary significantly from several nanoseconds to within the millisecond

range depending on if the sample is a dye versus a quantum dot or lanthanide [16]. Typically for larger ranges, a system clock can be counted, and the TDC range only needs to interpolate within a system clock cycle. State-of-the-art TDCs are capable of achieving dynamic ranges spanning into the hundreds of nanoseconds or even the microsecond range.

Precision: When repeatedly measuring a constant input time interval, the TDC should ideally give the same output. In the real case, the TDC output will form a distribution around the mean value due to various sources of timing jitter from the system, which is often dominated by the jitter of the SPAD. The standard deviation of this distribution gives the single-shot precision of the TDC. Other sources of uncertainty that worsen the precision of the TDC include the jitter of the *start* and *stop* signals, the jitter of the reference clock, the quantization error, and additional jitters coming from other signals within the TDC [55]. Concerning the TDC itself, the jitter performance can be improved through a trade-off with power consumption and size, as wider transistors can be used to reduce jitter. The precision of an ideal TDC should be in the range of picoseconds and ideally less than the resolution for optimal single-shot performance.

Nonlinearity: In an ideal TDC, the step widths are equal along the entirety of the quantization characteristics. In the real case, the variations in the step widths contribute to the nonlinear performance of the TDC that can be determined by a method such as a delay sweep or the statistical code density test. Common sources of nonlinearity within the TDC include delay mismatches, layout mismatches, and PVT variations. The nonlinearity of the TDC manifests itself in two forms given by the differential nonlinearity (DNL) and the integral nonlinearity (INL). The DNL is the difference between the actual and ideal step widths for each individual step on the quantization characteristics. It represents the error achieved on a conversion pertaining to specific output words. INL is given as the integration of the DNL along the quantization characteristics and may result in missing codes if it is too large. Additionally, a TDC with a high nonlinearity may not exhibit the expected monotonic response and incorrectly categorize the relative lengths of time intervals. The DNL and INL are generally given in units of LSB in order to normalize to the resolution of the TDC, with normal values for the INL being between zero to a few

LSBs. Additionally, the DNL and INL can either be given as an RMS or maximum value across all possible output words.

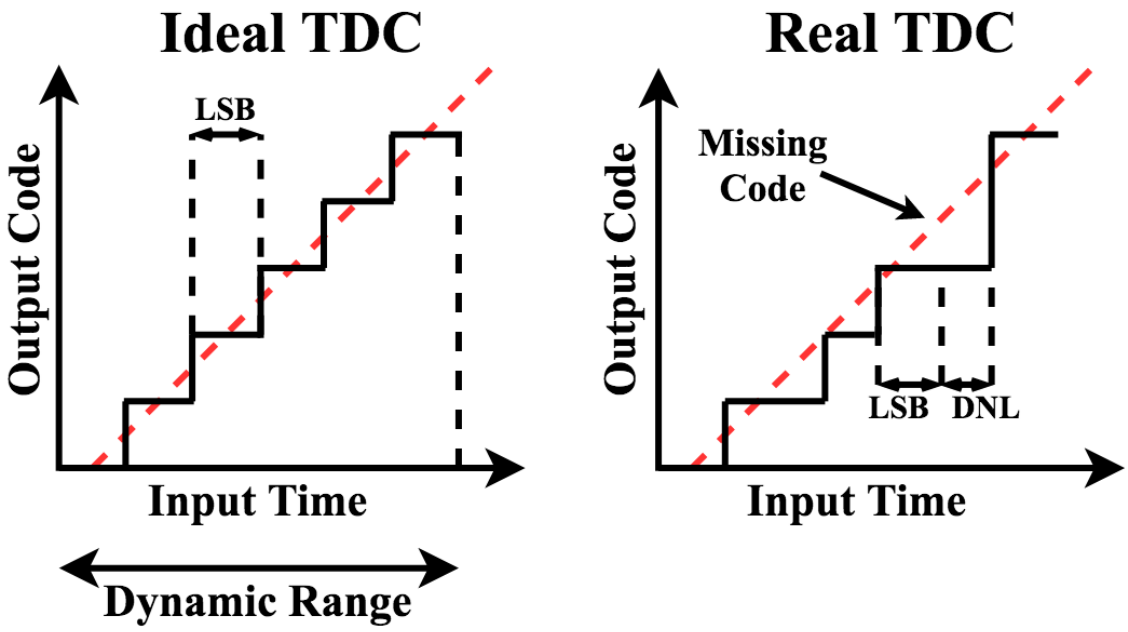

Figure 1-9: Comparison of the ideal and actual TDC quantization characteristics (© 2021 IEEE).

Sampling Rate: The dead time of a TDC is the amount of time that it takes for the TDC to complete a conversion and the inverse gives the maximum sampling rate (or counting rate). The dead time and sampling rate are responsible for determining the maximum operating frequency of the device. In general, having a high sampling rate is important for the SPAD-based sensor applications; otherwise, information from many valid SPAD pulses may be missed while a conversion takes place. Since in most implementations, TDCs are shared between many SPADs, the TDC must operate at a frequency high enough to respond to pulses from several SPADs. The sampling rate normally falls in the range of tens to hundreds of megahertz.

Power and Area: While the power consumption and area requirements do not provide direct information on the performance quality of a given TDC, they must be strongly considered early in the design process since they will affect the overall system performance and cost. In the SPAD-based sensor implementations that integrate a large number of TDCs, each TDC should be kept very small in order to allow the SPAD to have a large

active area for photodetection. This has led to most designs sharing a TDC between many SPADs in order to minimize the negative impact of the TDC size on the fill-factor and, by consequence, the PDE of the SPAD. The power considerations are very important as well since a practical system may have thousands of SPADs. If it is desired to use a 1:1 coupling of SPADs to TDCs, then the total power consumption may increase very quickly, limiting the viability of a particular design. State-of-the-art TDCs report power consumptions in the range of milliwatts, and the area is commonly in the range of hundredths of a square millimeter.

Figures-of-Merit: In the literature, a figure-of-merit (*FOM*) suitable for the SPAD-based sensor application is not yet generally adopted. The most common FOM for TDCs was modified from a common ADC *FOM* shown in equations (1-4) and (1-5), where F_S is the sampling rate, and N_{linear} is the effective number of linear bits calculated using b, the number of bits, and the *INL*.

$$FOM_1 = \frac{Power}{2^{N_{linear}} \times F_s} \tag{1-4}$$

$$N_{linear} = b - log_2(INL + 1) \tag{1-5}$$

While this *FOM* provides some information on the TDC performance, such as the sampling rate, power consumption, number of bits, and *INL*, it neglects the area considerations, which are very important for maximizing the PDE in a dSiPM or SPAD imager. This may be misleading since a TDC could have a very strong *FOM*, but if the area is large, it may be unsuitable for integration with SPADs. In [56], a new *FOM* (i.e., equation (1-6)) was proposed that the authors believed is more suitable to the dSiPM application since it directly considers the area and timing precision given by σ.

$$FOM = \frac{Power \times Area \times \sigma}{F_s} \tag{1-6}$$

While equation (1-6) is more effective in identifying whether a given TDC is valid for a dSiPM application and includes the effect of the precision trade-off with power consumption as a result of delay line jitter, many of the published TDC results do not report the precision, limiting the use of this *FOM* for designers seeking guidance from published results. For that reason, this work presents a new *FOM* in the comparison tables that is

adapted from the aforementioned *FOM*s, in order to try to obtain a *FOM* more suitable to the SPAD-based sensor application while using values that are more consistently reported in the TDC results. The new *FOM* is shown in equation (1-7).

$$FOM_2 = \frac{Power \times Area \times LSB}{2^b \times F_s} \tag{1-7}$$

In this research, the traditional TDC *FOM* will be used alongside the proposed FOM. Note that the lower the *FOM*, the better is the design considering the specific parameters used in the *FOM* expression. Also, the *INL* of TDCs in the recent literature is roughly half the time presented as the maximum value and half the time as an RMS value across the dynamic range. For this reason, FOM_1 cannot always be compared directly between papers. Here, the convention is adopted that the maximum value of the *INL* is used when calculating FOM_1 since that was most commonly reported. The *INL* was not included in the proposed *FOM* despite it being a very important performance metric in an attempt to create a FOM that can be calculated for the majority of TDC results. This is additionally why the resolution was used in place of the precision from the *FOM* presented in equation (1-6). Due to the lack of the *INL* term in FOM_2, it should not be taken as a direct performance indicator but as a general interpretation of the performance that may be missing or hiding some of the finer details that can be obtained by reading closely into the specific TDCs linearity performance. Similarly, the precision and its trade-off with size and power consumption as a result of delay line jitter should be considered for the highest single-shot performance.

1.3. Research Contributions

The focus of this research was on the design of high-performance TDCs and SPAD structures within standard CMOS processes. This work targets the optimization of these sensors to performance levels where low-cost time-resolved SPAD-based sensors can be used in high-performance medical imaging systems for applications such as PET, FLIM, and DOT. The main contributions of this work are the following:

- **An extensive literature review was conducted on high-performance TDCs, and their results when integrated with SPAD arrays to form dSiPMs, and SPAD**

imagers. Through the process of this review, fundamental concepts of TDC methods were explored, and the results are presented for the most advanced state-of-the-art TDCs in the literature. Furthermore, system-level considerations encountered when integrating SPADs together with TDCs are described, and several research challenges are identified for future improvements.

- **Design and measurement results of a prototype TDC using feedback time-amplification in the TSMC 65 nm standard CMOS process.** The proposed TDC aims to achieve high resolution by implementing a multi-stage structure along with a feedback time amplifier. The measurement results of the initial prototype verify the functionality of the design in achieving resolutions below 5 ps, although future iterations should aim to improve the nonlinearity performance.
- **Design of initial measurement results of a multi-time-gated SPAD array with integrated TDCs.** This design aimed to maximize the fill-factor of the array by using the same multi-purpose delay line to provide shifted gate windows to an array of SPADs, generate the SPAD gating signals for each pixel, and perform coarse time-to-digital conversion. The p+/n-well SPAD is characterized in a passive quench configuration, and measurement results are provided for the time-gated pixels in the first design iteration.

Publications:

1. R. Scott, W. Jiang, and M. J. Deen, "CMOS Time-to-Digital Converters for Biomedical Imaging Applications," *IEEE Reviews in Biomedical Engineering* (Accepted June 19, 2021).
2. W. Jiang, Y. Chalich, R. Scott, and M. J. Deen, "Time-Gated and Multi-Junction SPADs in Standard 65 nm CMOS Technology," *IEEE Sensors Journal*, pp. 1–1, 2021, doi: 10.1109/JSEN.2021.3063319.

3. I. Faisal, S. Majumder, R. Scott, T. Mondal, D. Cowan, and M. J. Deen, "A Simple, Low-Cost Multi-Sensor-Based Smart Wearable Knee Monitoring System," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 8253–8266, Mar. 2021, doi: 10.1109/JSEN.2020.3044784.

1.4. Thesis Organization

In Chapter 1, several applications of time-resolved single-photon counting measurements were described. As a motivation for the following chapters, the system operation of PET, FLIM, and DOT were briefly described to provide context for the single-photon detector requirements within these applications. Next, we described the operation principle and main performance specifications of the two main building blocks that form time-resolved single-photon detectors: SPADs and TDCs. Lastly, a summary of the major contributions of this research and the organization of the thesis are described.

In Chapter 2, an extensive review of CMOS TDCs is presented. The fundamental circuit building blocks that provide the basis for the more complex methods implemented in recent years will be described. Then, the results of state-of-the-art TDCs are presented, broken down into several architectural categories. Following this, a review is presented on the integration of TDCs with SPAD arrays. While the TDCs themselves in these implementations often employ simpler topologies, we also consider system-level design aspects such as TDC sharing and readout approaches.

The design of a TDC using feedback time-amplification is presented in Chapter 3. While multi-stage TDCs can offer improved resolution and dynamic range, it often comes at the cost of a larger area and power consumption in the majority of designs. We utilized a feedback topology that reuses circuitry to provide equivalent resolution of competitive TDC structures in a smaller layout area. The TDC was fabricated in the TSMC 65 nm process, and the detailed measurement results are presented.

In Chapter 4, the design of multi-time-gated SPAD arrays is described. Shifting the gate window of time-gated SPAD arrays with respect to a synchronous laser pulse was demonstrated previously to generate histograms in time-resolved single-photon

measurements. In comparison with using a single time-gate window for an entire SPAD array, the proposed multi-time-gated approach aims to reduce measurement time by applying several shifted gate windows to different SPADs within the array. In this way, all bins of the histogram can theoretically be measured simultaneously. The proposed designs were implemented in the TSMC 65 nm process, and an example of the post-layout simulation results are shown. Lastly, we proposed an example of how the 2D design can be easily scaled to share a fine interpolating TDC for improved timing resolution in a larger design while minimizing the impact on the fill factor.

The results of the aforementioned multi-time-gated SPAD design fabricated in the TSMC 65 nm process are presented in Chapter 5. As a precursor, the same p+/n-well SPAD results are presented for a passive quench configuration. The SPADs are characterized in terms of their breakdown voltage, DCR, time-gate window, PDP, TDC resolution and nonlinearity, and the timing jitter.

In Chapter 6, a summary of the work that was performed, and the achieved results are given. Based on our work, we outlined several challenges that were identified for future research.

Chapter 2

Review of CMOS Time-to-Digital Converters

2.1. Fundamental Concepts

A straightforward TDC implementation that is the basis for most methods is the use of tapped delay lines (DLs) formed from the series connection of buffers or inverters as in [57]–[60]. In the basic DL approach from Figure 2-1, the rising edge of the *start* signal propagates through the DL, and the output of each stage is connected to the input of a D flip-flop (DFF) or arbiter (the term "sampler array" will be used in the images to refer to an array of arbiters or DFFs). When the *stop* signal arrives, the state of the DFFs is stored, resulting in an output code that represents how far along the DL the *start* signal propagated before the arrival of the *stop* signal. The result of the conversion is easily determined as it is simply the product of the delay of a single stage by the number of elements the *start* signal propagated through during the conversion.

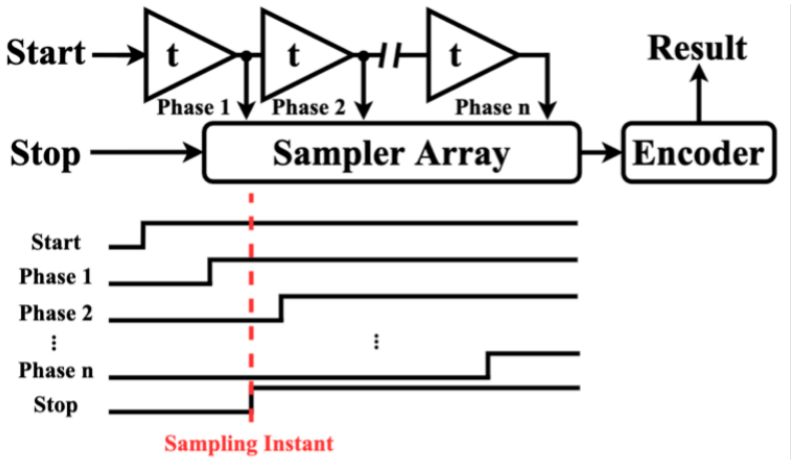

Figure 2-1:Illustration of a basic delay line TDC (© 2021 IEEE).

*Part of this work was accepted for publication in IEEE Reviews in Biomedical Engineering as: R. Scott, W. Jiang, M. J. Deen, CMOS Time-to-Digital Converters for Biomedical Imaging Applications.

An important consideration for delay line TDC design is that the delay at each stage is susceptible to PVT variations that are normally compensated for by locking the delay along the line to a reference clock period with a delay-locked loop (DLL) [61]–[65]. This is a general method that may be used in any TDC architecture that employs the use of delay lines. While the delay line based TDCs offer a very simple structure, their resolutions are limited to the logic gate delay in a given process, which led to the development of the Vernier and pulse shrinking TDCs (Figure 2-2 and Figure 2-3, respectively).

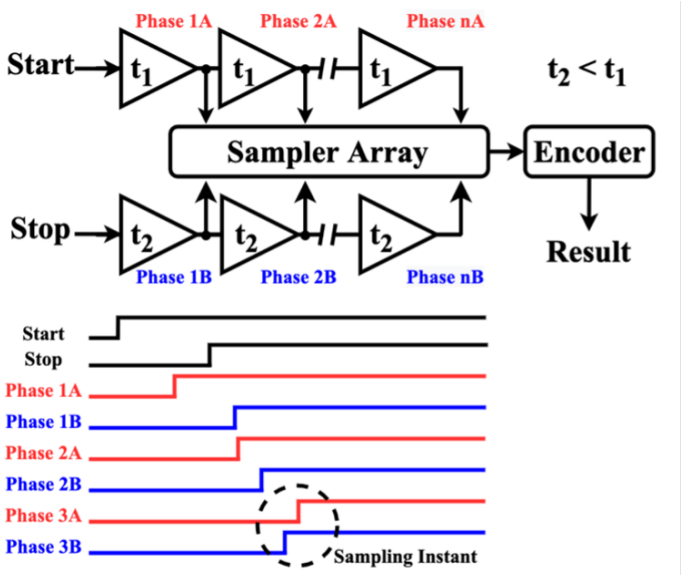

Figure 2-2: Illustration of a Vernier delay line TDC (© 2021 IEEE).

In an attempt to achieve superior timing resolution, TDCs were implemented utilizing Vernier delay lines (VDLs) [66]–[69]. The basic concept of the VDL is that the *stop* signal is no longer directly clocking the DFFs upon arrival. Instead, the *stop* signal propagates through a second DL that is designed to have slightly less delay than the *start* line. As such, the rising edge of the *stop* signal will propagate faster than that of the *start* signal. After each stage, the *stop* signal will clock the corresponding DFFs, and the gap between the *start* and *stop* signals will decrease by the difference in delays in the slow and fast delay elements. This is equivalent to one LSB. The end of the conversion will occur when the first zero is stored in a DFF, which occurs when the *stop* signal surpasses the *start* signal. Since the resolution is defined as the difference in delays between elements in the slow and

fast DLs, the VDL can achieve sub-gate delay resolution by designing the delay elements in the lines to have a very small delay difference. However, some trade-offs arise in the design. For example, although the resolution of the VDL is superior to the basic DL TDC, the DR and latency are worse. This is due to the fact that the *stop* signal must have enough time to surpass the *start* signal before it reaches the end of the line.

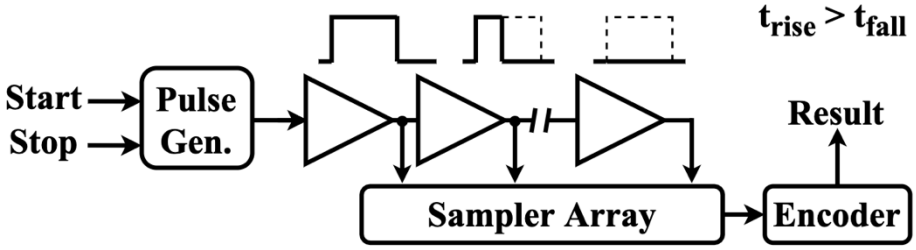

Figure 2-3: Illustration of a pulse shrinking delay line TDC. (© 2021 IEEE)

Sub-gate delay resolution can also be achieved by a topology known as pulse shrinking (PS) TDCs [70]–[74]. In the PS TDCs, the input time interval is applied as a pulse width on a single line rather than as the rising edges of separate *start* and *stop* signals. By designing the delay elements to have a shorter falling transition than rising transition using asymmetric transistor sizing or current starving, the pulse width is shrunk at each stage by the difference of the rising and falling transitions. If the difference between rise and fall transitions is made small enough, the PS line can achieve sub-gate delay resolution while only using a single DL, theoretically providing advantages in size and power consumption over the VDL method. To form the PS TDC, the data inputs of the DFFs are generally held high, and the output of consecutive PS buffers are applied as the clock inputs. At a certain point down the line, the pulse will be diminished due to the shrinking, and the DFF will not be clocked, resulting in a 0 being stored, indicating the end of the conversion. A major limitation of the PS TDCs is that as the pulse width becomes small, the shrinking rate becomes non-uniform and worsens the linearity performance [75], [76]. As the pulse is shrinking while it moves down the line, the non-uniform shrinking rate issue for narrow pulse widths may be encountered if the pulse width is not large enough in comparison to the stage pulse shrinkage [77].

The aforementioned TDC methodologies provide the basis for the more sophisticated methods seen in the literature today. In the next subsection, we will discuss the recent advancements made and the results achieved by published state-of-the-art TDCs.

2.2. Results of State-of-the-Art TDCs

The previous section gave an overview of fundamental principles used for TDCs, such as basic DLs, VDLs, and PS techniques. Now, modern state-of-the-art TDCs have expanded these approaches in order to achieve higher performance. A general improvement that was made is to fold the DLs into ring oscillators (ROs) so that a large range can be obtained within a smaller silicon area. In this configuration, the DL becomes a RO [78]–[85], the VDL becomes the Vernier ring oscillator (VRO) [86]–[95], and the pulse shrinking DL becomes the pulse shrinking ring (PSR) [50], [51], [75], [76], [96]–[99]. Aside from these approaches, multipath ring oscillators [100]–[103], ΔΣ TDCs [104]–[111], and time amplification based approaches [52], [112], [113] were proposed in recent years. A summary of the results in recent TDCs is presented in Table 2-1.

2.2.1. Vernier TDCs

Perhaps the most common topology for standalone TDCs published in recent years is based on the VRO, shown in Figure 2-4. The VRO is a TDC method adapted from the VDL, where the DLs are now formed into ROs with different frequencies. In this configuration, the per-stage delay difference between the slow and fast ROs determines the resolution. The operation of the VRO TDCs is generally such that the arrival of the *start* and *stop* signals enable the oscillations of the slow and fast ROs, respectively. As the *start* signal arrives first, the rising edge of the slow RO begins propagating through the ring, and the number of iterations is tracked by a loop counter, thereby allowing the VRO TDC to have a large dynamic range with a smaller number of delay stages. The arrival of the *stop* signal will enable the propagation in the fast RO, and arbiters or DFFs are used to compare the oscillator phases in order to find the position in which the fast RO rising edge passes

Table 2-1: Results of State-of-the-Art TDCs (© 2021 IEEE).

Year Ref.	Type	Tech.	LSB	DR	INL	Rate	Power	Area	FOM1	FOM2
Unit	-	nm	ps	ns	LSB	MHz	mW	mm²	pJ/conv. step	pJ x mm² x ps/conv. step
2009 [101]	MP-GRO[a]	130	6/1	12	-	100	2.2~21 @50MHz	0.0405	-	0.00831
2012 [114]	V-GRO[a]	90	5.8/3.2	40	-	25~100	4.5 @25MHz	0.027	-	0.00190
2013 [103]	MP-GRO[a]	65	5.035/4.22	1	-	200	1.73~2.20	0.02	-	0.0113
2013 [55]	Multiple Interp. (VRO)[b]	350 HV	10	160	0.98 rms	3	15/80	0.3	-	0.458
2013 [115]	Two-step (GDL)	65	3.75	0.476	2.3 max	200	3.6	0.02	0.464	0.0105
2014 [112]	Pipe.	65	1.12	0.578	1.7 max	250	15.4	0.14	0.325	0.0189
2015 [53]	Multiple Interp. (SRO)	350	0.61	327000	7.4 max	0.8	80	0.64	12.8	0.596
2016 [116]	2DV-GRO	65	10.6/2.2[a]	20	-	50	2.3	0.068	-	0.000840
2016 [117]	V-GRO	130	7.3	9	1.2 rms	2.4	1.2 @1MHz	0.03	-	0.257
2016 [76]	PSR	180	1.8	0.92	8.7 max	4.4	3.4	0.07	14.6	0.190
2016 [111]	3rd Order ΔΣ	110 1P6M	4.7	39.06	-	12.8	0.4	0.11	-	0.00197
2017 [56]	V-GRO	65	15	3.44	0.39rms/ 0.83 rms	5	0.160 @1MHz	0.0013	-	0.00272
2017 [52]	Feedback V-GRO[c]	65	0.98/6.01	5.76	2.2 max	10/250	3.0/17.5	0.02	0.117	0.000718
2018 [118]	2D-VDL	45 SOI	1.25	0.319	0.34 max	80	0.3	0.04	0.0196	0.000732
2018 [81]	RNS-RO	45	9.4	1.96	1.8 max	500	27.3	0.08	0.481	0.172
2019 [75]	Two Step (PSR)[d]	180	2.0	130	4.2 max	3.3	18.0	0.08	0.433	0.0133
2019 [113]	Two Step (DL)[d]	180	5.3	1.3	2.8 max	30	1.1	0.05	0.544	0.0380
2019 [110]	2nd Order ΔΣ	65	5	0.7	-	50	3.5	0.09	-	0.223

[a] Raw resolution/effective resolution with noise shaping; [b] Power given for single TDC channel/with other circuitry such as 3 DLLs that would be shared by a TDC array; [c] Results given for feedback mode/feedforward mode; [d] Simulation results. Abbreviations: (G)DL: (gated) delay line; (G)RO: (gated) ring oscillator; MP: multipath; PSR: pulse shrinking ring; RNS: remainder number system; SRO: switched ring oscillator; V: Vernier; 2DV: 2-dimensional Vernier.

that of the slow RO, denoting the end of the conversion. The phase in which the conversion ends and the loop counter value are then combined to give the final result.

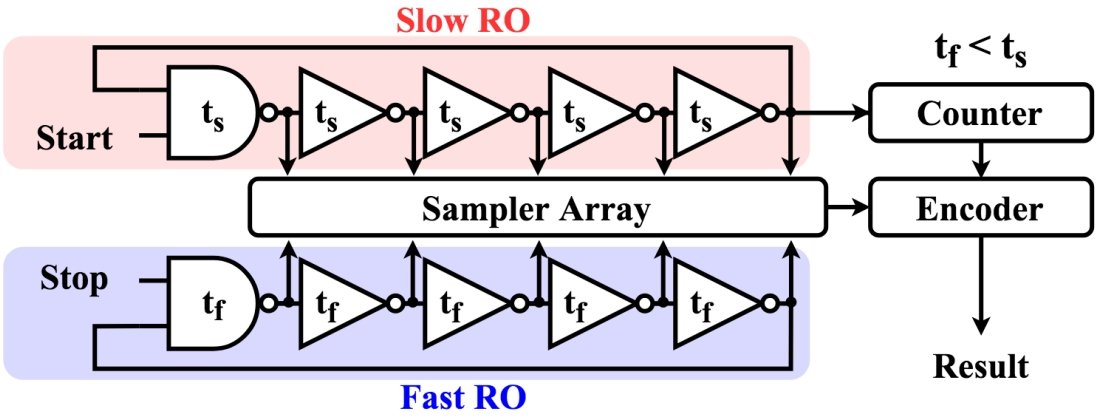

Figure 2-4: Diagram of a Vernier ring oscillator (VRO) TDC (© 2021 IEEE).

A Vernier gated ring oscillator (V-GRO) TDC in 90 nm CMOS was designed in [114], where the slow and fast oscillators were gated (i.e., enabled and disabled) by transistors in the delay elements. This has the effect of integrating the quantization error of a conversion to improve the resolution through first-order noise shaping. Since the phase of the ROs is held at the end of the conversion, each conversion will start from a different phase. Therefore, a multiphase counter was therefore used to obtain the conversion result by counting the transitions of each phase, adding some complexity and additional area compared to the more commonly used single-phase counter. This TDC achieved a 3.2 ps effective resolution within a small area of 0.02 mm^2. At a sampling rate of 25 MHz, the power consumption in the Vernier mode was moderate, being reported as 3.6 mW.

Two-dimensional Vernier (2DV) TDCs, depicted in Figure 2-5, had superior performance compared to the V-GRO TDCs, with some of the best FOM_2 scores provided in Table 2-1. The sampling rate in the 2DV TDCs is increased by allowing comparisons between all possible phase pairs of the slow and fast ROs. In a standard Vernier TDC, the end of the conversion occurs when the *start* signal is surpassed by the *stop* signal at the same phase. However, since the 2DV TDC allows comparisons between all phases, the TDC conversion can be ended when the *stop* signal passes the *start* signal at other phases,

reducing the dead time. A two-dimensional Vernier gated ring oscillator (2DV-GRO) TDC was developed in [116] using 65 nm CMOS that achieved a sampling rate of 50 MHz by reducing the latency time to less than a sixth of the equivalent V-GRO TDC. Digital calibration was used to set the slow and fast oscillator periods using a capacitor bank, and an effective resolution of 2.2 ps was obtained through first-order noise shaping. Additionally, the power consumption and area were reported as 2.3 mW and 0.068 mm^2, respectively.

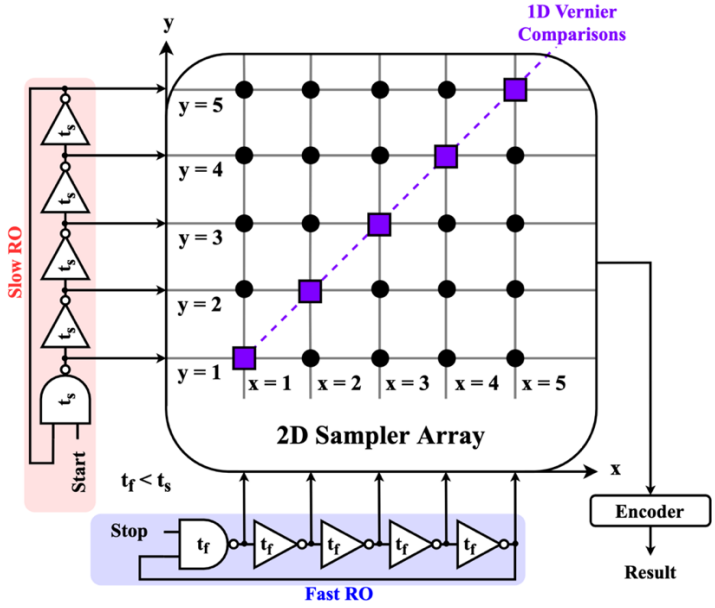

Figure 2-5: Diagram of a 2D Vernier ring oscillator (2D-VRO) TDC. The 1D array of samplers from the traditional VDL falls along the main diagonal. The 2D array reduces size and latency by allowing comparison between all phase differences in the 2D plane (© 2021 IEEE).

A simple method of improving the linearity of TDCs was explored in the single-stage VRO TDC of [55]. In this TDC, the *start* and *stop* signals are used to enable and disable a 100 MHz reference clock counter, respectively. By designing the TDC such that both *start* and *stop* are asynchronous to the reference clock, the conversion result is then given by the counter value as well as two clock interpolations that are the time differences between the *start* and *stop* signal rising edges to the following rising edges of the reference clock. This is sometimes referred to as the Nutt method [119], [120]. In this single-stage VRO example, the time between the asynchronous rising edges of the *start* and *stop* signals and the

subsequent rising edge of the reference clock are random values that are subtracted from each other when computing the final conversion. This has the effect of averaging the bin widths, and thus improving the linearity in what is known as the sliding-scale method [55]. The linearity is also improved since the fine interpolation is achieved by a single-stage VRO where minimal delay elements are used. This reduces delay mismatch by using only one phase comparison between the slow and fast ROs.

2.2.2. Pulse Shrinking Ring TDCs

PSR TDCs should theoretically have size and power advantages over the VROs since only a single RO is required (as shown in Figure 2-6); however, issues arise in terms of linearity. In [76], the issue of the non-uniform pulse shrinking rate was addressed in an attempt to achieve better linearity. In this TDC, rather than injecting the pulse width into a PSR and ending the conversion when the pulse disappears, the measurement interval is added to a 50% duty cycle signal using a pulse injection scheme. The conversion is ended when the duty cycle subsequently falls below 50%. This end of conversion event is detected by using DFFs connected to opposite sides of the PSR. When designed such that the 50% duty cycle pulse width is within the uniform shrinking range, this topology will ensure that the pulse never encounters the non-uniform shrinking rate issue due to a narrow pulse width. However, while this method offered a novel topology, a large maximum INL of 8.7 LSB was reported due to the package inductance of the fabricated chip.

This work was expanded upon by the same research group in [75]. In this later work, the area and power consumption performance were sacrificed in order to design a two-step TDC using a ring oscillator coarse counting stage that increased the dynamic range from 0.92 ns to 130 ns. This TDC also reported a relatively high maximum INL of 4.2 LSB (simulation result). These results would indicate that the PSR TDCs are capable of achieving high resolution around 2 ps, but suffer from high nonlinearity in these works.

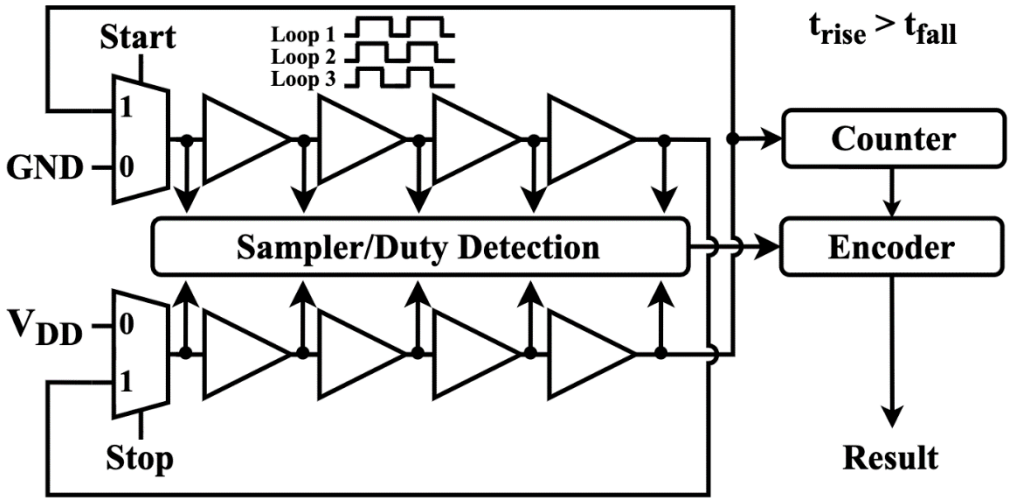

Figure 2-6: Diagram of a pulse shrinking ring (PSR) TDC (© 2021 IEEE).

2.2.3. Multipath Ring Oscillator TDCs

An alternative method of achieving sub-gate delay resolution is using multipath gated ring oscillator (MP-GRO) TDCs, as shown in Figure 2-7. Based upon the ideas presented in [57] and [58], MP-GRO TDCs used inverters with multiple inputs tapped to different stages of the oscillator to decrease the effective delay per stage. This, in turn, increased the oscillator frequency and resolution. As the performance improved when the number of stages was a prime number, the MP-GRO introduced in [101] employed a 47-stage multipath oscillator, resulting in an improvement by a factor of 5 in the raw resolution compared to the standard RO approach. Additionally, NMOS and PMOS transistors are used to perform gating, which carries over the quantization error to the next conversion, resulting in first-order noise shaping of the quantization noise. This improved the effective resolution from 6 ps to 1 ps at a 50 MHz sampling frequency.

In the design of a multipath oscillator, the inputs of each stage are connected to earlier stages compared to a typical RO, so the state transitions occur earlier, and multiple transitions may occur at the same time. For this reason, the MP-GRO TDCs have more complicated sampling circuits, which requires the oscillator to be segmented into multiple sections where it is known that only one transition is occurring at a time. The conversion is then computed separately for each section and combined at the end to give the final result.

It should be noted that the MP-GRO TDCs of [101] and [103] achieved high sampling rates of 100 MHz and 200 MHz, respectively. This is because the conversion result is available very shortly after the arrival of the *stop* signal (i.e., flash operation) as opposed to the Vernier TDCs and pulse shrinking methods.

While the MP-GRO TDCs were able to achieve a high sampling rate and a very fine resolution, due in part to the noise shaping used in these examples, their disadvantages often include higher power consumption resulting from the lengthened state transition times and the complex sampling circuitry.

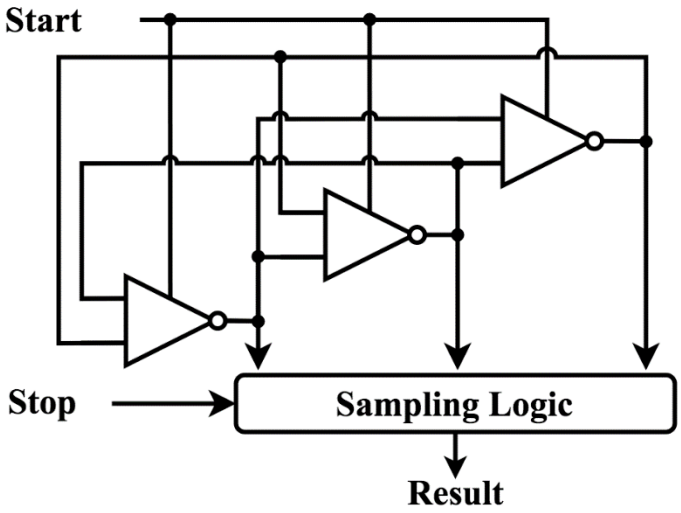

Figure 2-7:Diagram of a multipath gated ring oscillator (MP-GRO) TDC (© 2021 IEEE).

2.2.4. ΔΣ TDCs

Higher-order noise shaping was also shown to be an effective means of achieving high resolution, as demonstrated by ΔΣ TDCs. Analogous to ΔΣ ADCs, ΔΣ TDCs perform data conversion by using ΔΣ modulators to achieve quantization error noise shaping, thus increasing the effective resolution. In the general structure shown in Figure 2-8, a low-resolution TDC is used to coarsely quantize the input. The output code is then passed through a digital-to-time converter (DTC), which converts the code to a pulse width that is subtracted from the previously quantized value. The remainder can then be processed with the same low-resolution TDC to achieve a finer resolution. The main difficulties that were

faced in the development of ΔΣ TDCs are the implementation of circuits to perform time-domain operations such as addition, subtraction, and integration.

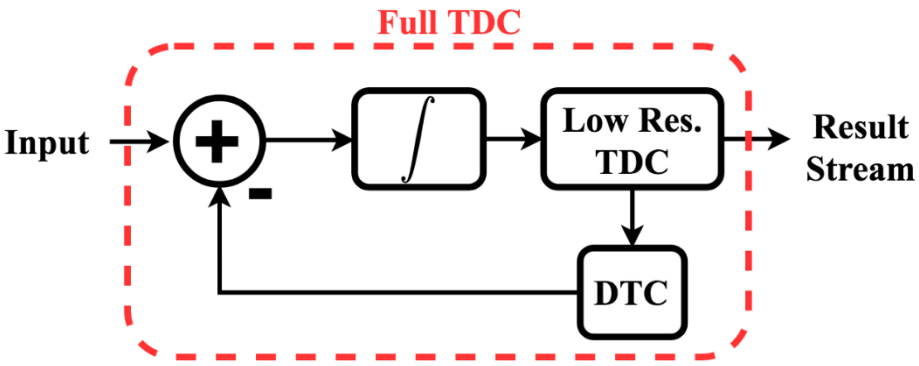

Figure 2-8: Diagram of a ΔΣ TDC (© 2021 IEEE).

In [111], a 3rd order ΔΣ TDC was developed using half-delay time integrators formed by two *AND* gates, a charge pump, and a voltage integrator. The charge pump was responsible for converting time information to a voltage prior to integration. This TDC achieved an effective resolution of 4.7 ps in a 110 nm technology at a sampling rate of 12.8 MHz with a moderately large dynamic range of 39.06 ns. The power consumption of this TDC was also very low, being reported as 0.4 mW, due to the simplicity of the half-delay time integrator. Another ΔΣ TDC developed in [110] was able to obtain an effective resolution of 5 ps at a sampling frequency of 50 MHz. While this TDC achieved an improved sampling rate for a comparable resolution and area requirement, the power consumption was increased to 3.5 mW due to the gated delay lines (GDLs) chosen to perform the time-domain arithmetic operations.

2.2.5. Time Amplification TDCs

The GDLs used for implementing time integration in the ΔΣ TDC from [110] were also used to implement time amplification for the TDCs designed in [115] and [112]. The pulse-train time amplifier (TA) proposed in [115] and [112] achieved linear and programmable time amplification without a calibration circuit. Using DLs and an *OR* gate to generate non-overlapping replicas of an input pulse width, the time interval is amplified by passing the replicas as the enable signals to a GDL. These replicas enable the input signal (tied to V_{DD})

to propagate through a number of stages in the GDL equal to the total pulse width of all the replicas. The new pulse width stored in the GDL is then equal to the original pulse width multiplied by the number of replicated pulses, which can be programmable.

A two-step TDC in 65 nm CMOS was developed in [115] that used this pulse train TA in between its coarse and fine DL-based TDC stages. This allowed the design to achieve a higher resolution and dynamic range than the basic DL TDC with the same number of delay elements. The coarse and fine stages were designed identically such that the increased resolution of the fine stage comes as a result of the time amplification. Due to the flash conversion nature of DL-based TDCs, this design was able to achieve a high sampling rate of 200 MHz, at which the power consumption was 3.6 mW. The proposed TA occupied only 0.0024 mm^2 on chip and the total TDC area was 0.02 mm^2, being one of the most compact designs in Table 2-1. A resolution of 3.75 ps was achieved with a dynamic range of 0.476 ns, which could be improved by using a RO counter method in the coarse stage as opposed to a DL TDC.

In [112], a 9-bit TDC was developed in 65 nm CMOS that achieved a resolution of 1.12 ps by using GDLs to implement a pipelined TDC. The proposed TDC consisted of 3 pipelined stages with intermediary 4x pulse-train TAs, resulting in a total gain of 64. Each pipelined stage was responsible for quantizing their input signal using a DL, and subsequently amplifying the residual and passing it to the next stage. The pipelined operation allowed for multiple conversions to occur at a given time as each stage operates independently and synchronously. This allowed the TDC to achieve a high sampling rate of 250 MHz; however, the power consumption and area requirements were degraded to 15.4 mW and 0.14 mm^2 respectively, from the TDC of [115].

While the pulse train TA has shown promising results, time amplification was most commonly achieved in TDCs by exploiting the metastability window of SR latches. The first proposed coarse-fine TDC utilized an SR latch to achieve a resolution of 1.25 ps [122]. This can often result in inaccurate gains with narrow input ranges if methods of calibration or linearization are not considered in the design. However, due to its compact nature, the SR latch TA can be implemented in a smaller silicon area and with lower power

consumption. Additionally, while the previous time amplification TDCs utilized a feedforward approach as in Figure 2-9, a feedback time amplification topology was demonstrated in the literature as in Figure 2-10.

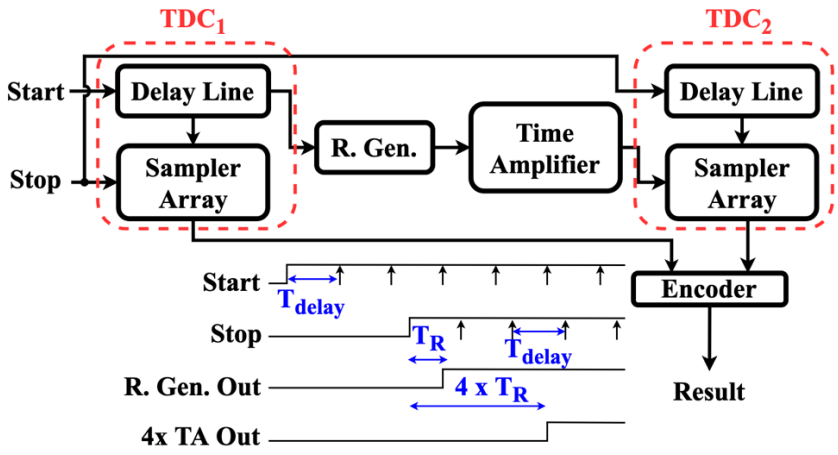

Figure 2-9: Diagram of feedforward time amplification (TA) TDC (© 2021 IEEE).

In fact, the best performance for FOM_2 was demonstrated by the TDC from [52] that utilized a single-stage V-GRO with a feedback path containing an SR latch-based TA. This topology would theoretically allow for repeated conversions on the residual of the measurement result; however, the proposed design implemented only a single feedback iteration. During operation, the *start* and *stop* signals undergo a coarse conversion using a counter on the slow RO. Afterwards, the conversion residual was generated, amplified, and applied as the input to the same V-GRO, where it then underwent another coarse conversion. On the last feedback cycle, a fine conversion was also performed using the single-stage VRO method. Due to the feedback nature, this TDCs throughput was limited to 10 MHz. This sampling rate was a moderate performance compared to most of the results of Table 2-1. However, if a higher throughput is desired, then the resolution can be sacrificed to 6.01 ps in order to operate the TDC in a feedforward mode where the sampling rate is given as 250 MHz, being one of the highest of Table 2-1. It should be noted that this TDC implemented offline calibration for digitally controlling the oscillators as well as for correcting the TA, which may be impractical or undesirable for certain applications such

as ToF PET which requires thousands of channels of TDCs in a complete system. This TDC was able to achieve the second highest resolution of Table 2-1 at 0.98 ps over a wide dynamic range of 5.76 ns. The area was only 0.02 mm^2 and the power consumption was moderate at 3.0 mW. The TDC also achieved an acceptable maximum INL of 2.2 LSB, which was likely improved due to the single-stage structure of the Vernier TDC. Overall, this TDC was able to achieve competitive performance in all the metrics with no major weaknesses, which is indicated by it being the highest-ranked TDC in terms of FOM_2.

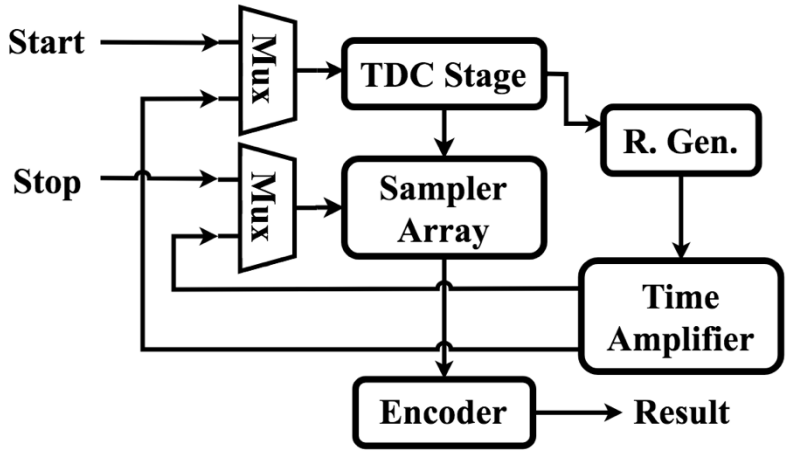

Figure 2-10: Diagram of a feedback time amplification (TA) TDC. The feedback TA TDC places the TA in the feedback path and therefore requires only a single TDC stage (© 2021 IEEE).

2.3. TDCs for Biomedical Imaging Applications

In recent years, the integration of TDCs with arrays of highly sensitive photodetectors known as SPADs has led to incredibly powerful sensors such as SPAD imagers and dSiPMs (summarized in Table 2-2) that are capable of timestamping the detection of individual photons. SPAD-based sensors have found numerous applications in biomedical imaging, such as: Raman spectroscopy [123], functional near-infrared spectroscopy (fNIRS) [124], fluorescence correlation spectroscopy (FCS) [125], PET [2], FLIM (widefield [126], confocal [127], and Förster resonance energy transfer (FLIM-FRET) [128]), and near-infrared diffuse optical tomography (NIROT) [20]. In this section, we focus on the

Table 2-2: Summary of TDC integrated with SPADs to form dSiPMs and SPAD imagers (© 2021 IEEE).

Year Ref.	SPAD				TDC				SYSTEM				
	Cell Pitch	FF	Peak PDP	Median DCR	Type	LSB	DR	DNL/INL	Tech.	# SPADs	# TDCs	TDC Sharing	App.
Unit	μm²	%	%	Hz (@RT)	-	ps	ns	LSB	nm	-	-	-	-
2011 [129]	50x50	1	27.5 @ 460nm Vex=1.40V	50 Vex=0.73V	GRO	55	55	0.3/2	130^{CIS}	20480 (160x128)	20480	Integrated 1:1 with SPAD	FLIM
2012 [19]	21.5x21.5	10	-	5.5k Vex=3.0V	GRO	56.5	3700	-	130	1024 (32x32)	16	8 interleaved TDC pairs shared through OR-tree	FLIM
2012 [18]	50x50	2	25 @ 500nm Vex=1V	100 Vex=1V	Interp. (DL)	119	100	0.4/1.2	130^{CIS}	1024 (32x32)	1024	Integrated 1:1 with SPAD	FLIM
2014 [3]	48x48	0.77	30 @ 425nm Vex=1.5V	544 Vex=2.5V	Interp. (DL)	62.5	64	<4/<8	130	4096 (64x64)	4096	Integrated 1:1 with SPAD	FLIM
2014 [130]	30x50	21.2	30 @ 420-430nm Vex=4V	~70k Vex=4V	GRO	51.8	3390	1.97/2.39 (LUT)	350^{HV}	416 (16x26)	48	Column parallel TDCs (3 per column)	PET
2014 [2]	-	42.6	45 @ ~410nm Vex=1.5V	13.7k Vex=1.5V	GRO	64.56	261.59	0.28/3.9	130^{CIS}	92160 ([24x30]x[8x16])	256	2 interleaved TDCs per pixel shared through OR-tree	PET
2015 [4]	23.78x23.78	43.7	-	1.4k -	GRO	40	>2600	-	130^{CIS}	2048 (256x8)	256	Shared by columns through OR-tree and select logic	FLIM Raman
2015 [20]	11.75x11.75	23.3	12.2 @ 800nm Vex=1.5V	35k Vex=1.5V	RO 2-spd.	49.7	200	0.44/0.47	130^{3D}	800 (2x400)	100	Shared by 2x4 groups of SPADs through WTA circuit	NIROT
2015 [42]	8x8	19.6	-	-	TAC	6.66	50	-	130^{CIS}	65536 (256x256)	65536	Integrated within SPAD cell	FLIM
2015 [131]	30x50	~39	18.6 @ 420nm Vex=3.5V	37k @ 20°C Vex=3.5V	GRO	48.5	6360	0.75/4	350^{HV}	67392 ([16x26]x[9x18])	432	Column parallel TDCs (48 per mini-dSiPM column)	PET
2018 [132]	19x5	-	-	-	RO	40	1000	0.12/<1	40	32768 ([128x64]x[2x2])	512	Column parallel TDCs (1 TDC per semi-column)	PET
2019 [133]	~16x16	32.1	~31 @ 450 nm Vex=5V	500 Vex=2.8V	GRO	80	81.8	0.2/2.4	150	3840 ([3x10]x[16x8])	128	Shared by 3x10 groups of SPADs through OR-tree	Particle Therapy
2019 [134]	18.4x9.2	13	34 @ 560 nm Vex=1V	25 Vex=1.5V	GRO	33	135	$0.9/5.64_{p\text{-}p}$	40	24576 (192x128)	24576	Integrated 1:1 with SPAD	FLIM
2020 [135]	62.3x202.4 (2 SPADs)	37	~54 @ 400 nm Vex=3.3V	424 Vex=3.3V	VDL	78	~10	0.039/0.58	350	1728 ([12x36]x[2x2])	1	Shared through OR-tree	NIRS

applications of PET, FLIM, and DOT.

It should be noted that SPAD imagers generally refer to SPAD arrays where photon counts and timing information are commonly available at the pixel level, while dSiPMs often refer to sensors in which the outputs of an array of SPADs are summed to give a totally digital output on a single channel (analogous to an analog silicon photomultiplier). As a convention for simplicity and clarity in this thesis, we will often refer to dSiPMs and SPAD imagers using the umbrella term "SPAD-based sensors" in order to avoid any misclassification while still providing a comprehensive overview of the hardware being developed in this field. Additionally, while it is not required for a SPAD-based sensor to contain a timing circuit, this thesis will only discuss SPAD-based sensors that integrate at least one TDC. In this section, TDCs will be discussed in how they relate to the system-level design and performance of SPAD-based sensors. Examples of the most common TDC topologies integrated with SPADs are described. An overview of methods for sharing TDCs between arrays of SPADs is then discussed, along with the impacts this has on the detector's performance. Lastly, the readout methods currently being employed are considered, as bandwidth requirements become a key concern in the development of large SPAD-based sensors with high throughput.

2.3.1. TDC Topologies

As shown in Table 2-2, since the early integration of TDCs with SPADs, a common approach to time interval measurement was delay line interpolation of a reference clock using a DLL (shown in Figure 2-11). These TDCs are constructed by using a coarse counter that records the number of periods of a reference clock during the measurement interval. By locking a delay line to the reference clock, multiphase clocks can be generated that are equally spaced throughout the reference period. Sampling the multiphase clock can then provide a finer measurement of the time interval, where a larger number of delay elements in the DLL (i.e., more phases) gives a finer resolution.

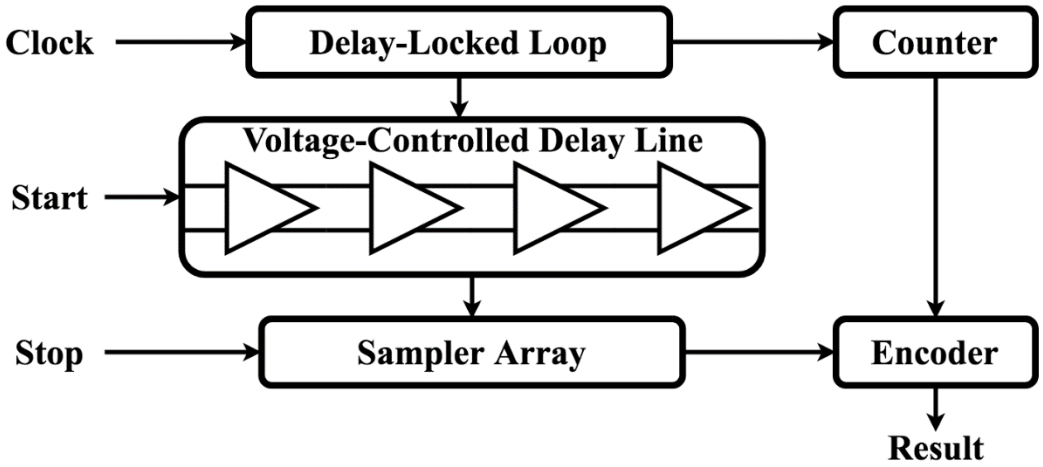

Figure 2-11: Diagram of a differential delay-locked loop (DLL) interpolation TDC (© 2021 IEEE).

In [17], a 32-phase clock was generated from a DLL and used alongside a 5-bit coarse counter to obtain a 350 ps resolution, with a dynamic range of 358 ns. Further levels of interpolation were used in other works to achieve improvements in the resolution. In [136], a three-level interpolating TDC was used to give a 97.66 ps resolution with a 100 ns dynamic range. Using the multiphase clock generated by a master DLL, a 40 MHz reference clock was divided into 16 evenly spaced phases. By performing the first level of interpolation on a global level and then performing the fine interpolation within each TDC, the size of the TDC circuitry was greatly reduced due to the sharing of the coarse DLL by the fine interpolators. Locally, consecutive phases of the multiphase clock were then used to divide the frequency once again with another DLL, using a separate 32 stage delay line.

In recent years, RO-based TDCs were the most common TDC architecture for SPAD-based sensors (shown in Figure 2-12). In contrast to many of the state-of-the-art TDCs discussed in the previous section, RO-based TDCs integrated with SPADs generally opt for the use of differential delay elements due to their ability to reject common-mode noise. In this approach, a RO can be formed from differential buffer elements, with reversed polarity feedback connections from the first to last stages, in order to ensure stable oscillations [137].

At the system level, there are likely to be many TDC circuits, so gating was generally adopted to lower the power consumption as in [129]. Here, additional logic in the RO starts

the TDC and freezes the state upon detection of the *stop* signal. This work also used this logic to reset the state of the TDC between conversions. In other works, gating was used as a means of obtaining noise shaping. In [4], the phase of the TDC was held at the end of a conversion instead of being reset. This allows the next conversion to start from where the previous measurement stopped in order to improve the linearity through noise shaping by effectively integrating the error across consecutive measurements. Improvements in the nonlinearity performance were also demonstrated through adopting the sliding scale property as in [132]. By allowing the ROs to run asynchronously to the system reference clock, the phase in which a measurement begins is no longer fixed, which allowed this work to minimize the impact of global and local transistor variations, as well as any oscillator mismatch, for a reported 6.25 times improvement in the linearity.

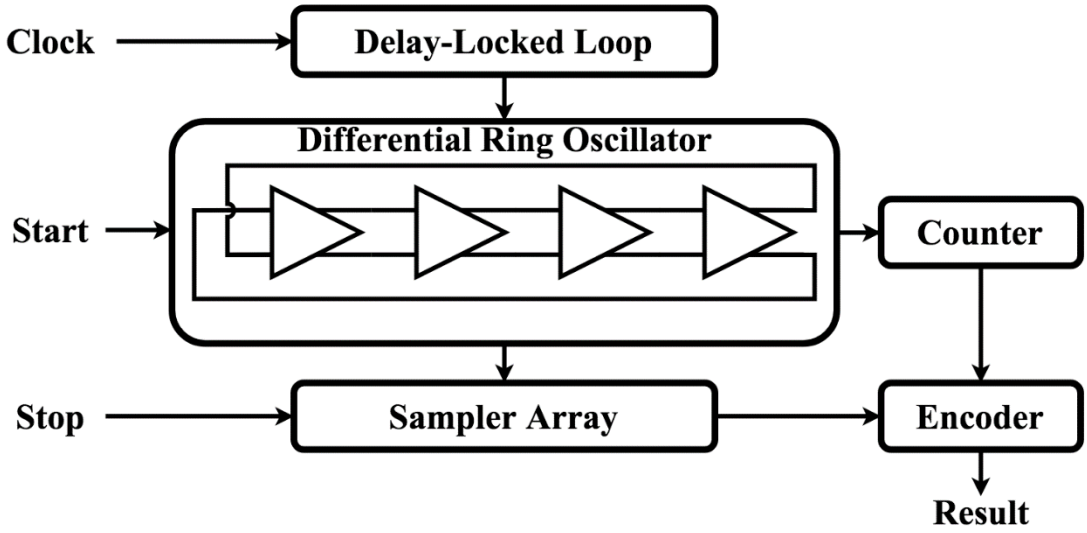

Figure 2-12: Diagram of a differential ring oscillator TDC (© 2021 IEEE).

While basic ring oscillators were most commonly reported in SPAD-based sensors, in the fully integrated SiPM Blumino, a MP-GRO (see Section 2.2.3) was used to form the TDC [138]. Due to their simple structure, the MP-GRO TDCs can be integrated in a small area and are viable candidates for higher resolution TDCs in SPAD-based sensors. An additional variation on the RO-based TDC was presented in [20], where a dual-speed RO was used in order to lower the power consumption. As traditional RO TDCs require high-frequency operation for fine resolution, this work aimed to limit the RO frequency for the

majority of the conversion without sacrificing the fine LSB. At the start of a conversion, the oscillator is enabled in a slow mode at 246 MHz, which determines the 5 MSBs of the result using a coarse counter. Following the next falling edge of the clock signal, the RO is switched into a fast mode at 2.52 GHz, where it is tracked by a fast counter. The state of the slow and fast counters, as well as the RO phase at the end of the conversion, give the measurement result. This topology resulted in a very low power consumption of 15 μW at a sampling rate of 500 kHz.

Although most of the designs integrating TDCs with a large number of SPADs have used simpler topologies, the design of [135] implemented a VDL TDC. A variation on the sliding scale method was used, where the TDC signals were injected on different sections of the delay line, which had the effect of averaging the bin widths across this range to improve the linearity by ~3.5 times. Two DLLs were used to lock the slow and fast delay lines to global references, and a resolution of 78 ps was obtained within a 10 ns range. In future, the VRO topology could be used to obtain this range in a smaller area, allowing for higher resolution. Gating of the oscillator could additionally be used in place of the sliding scale method as a means of improving the linearity through noise shaping.

While not as common, TACs were used in [42]. Due to the low transistor count of the TAC circuitry, this work was able to achieve a high fill factor of 19.6%, even considering that it used a TAC integrated into the front-end of each SPAD. Additionally, this work showed a fine resolution of 6.66 ps. The main advantage of this approach is that the TAC circuitry can be directly integrated with the SPAD biasing and quench/reset circuitry on the pixel level. However, while this design achieved a high resolution and high fill factor, it does not take into account that the complete system was not fully integrated into the chip and required external ADCs to complete the time interval measurement. In the works of [139]–[142], a SPAD array was routed to a set of parallel TACs, that recently reported precision of less than 10 ps FWHM, high-throughput, and a strong linearity performance with a peak-to-peak DNL less than 1.5% of the LSB.

2.3.2. TDC Sharing Schemes

Although the work of [42] was able to achieve a relatively high fill factor of 19.6% with TACs integrated into the SPAD front-end circuitry, in Table 2-2 it is shown that in general, TDCs being integrated 1:1 with SPADs provide the lowest fill factors. The sensor designed in [129] consisted of a 160 x 128 array of SPADs, integrated directly with RO-based TDCs in a 130 nm CMOS process. This resulted in a low fill factor of 1%. A higher fill factor of 2% was demonstrated with 1 TDC per SPAD in [18]. However, the TDC resolution was also worsened by a factor of approximately half, indicating a trade-off between fill factor and timing resolution when integrating a TDC with each SPAD.

Recently, a 192 x 128 SPAD array was fabricated in a 40 nm process, including 33 ps per-pixel TDCs [134]. The advanced technology node not only helped in achieving improved resolution due to the faster logic transitions in the TDC, but also in achieving a high fill factor of 13%, considering per-pixel TDCs. Using microlenses, the fill-factor was further improved to 42%, resulting in an effective PDE of ~14%. Additionally, the SPADs in this design demonstrated a very low median dark count rate of 25 Hz at a 1.5 V excess bias.

It should be noted that while designs using TDCs integrated directly with individual SPADs could potentially provide the highest single-photon timing resolution, in the majority of the published works, TDC sharing schemes are one of the most common ways to improve chip fill factor when a 1:1 integration is not viable. A summary of TDC sharing schemes is illustrated in Figure 2-13.

In an effort to improve the SPAD-based sensor performance to have a higher PDE, the spatial resolution of the SPAD array can be compressed in order to provide a higher fill factor by sharing TDCs between a group of SPADs. This approach is particularly effective when the sensor is expected to operate in a low photon density mode, such as with PET imaging. A common method was to use row or column parallel TDCs. In [136], a 128 × 128 SPAD array was partitioned into rows, where groups of 4 adjacent SPADs shared a single TDC. This work achieved a fill factor of 6% and a peak PDE of ~2%, showing 4 times peak PDE improvement over a 1:1 SPAD to TDC interface. This work also used a

row selection transistor within the SPAD pixel in order to enable one row at a time for acquisition. Therefore, the number of TDCs is reduced, as each row can share the same set of 32 TDCs since they will never be activated at the same time. A key disadvantage of this approach is that events can be missed if only a single row is activated at a time.

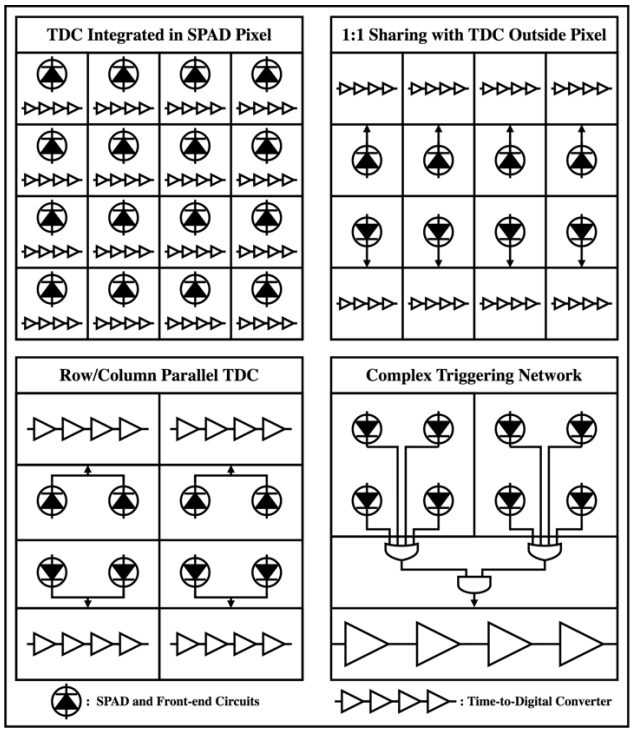

Figure 2-13: Illustration of various methods of sharing TDCs between an array of SPADs (© 2021 IEEE).

Using variations on this methodology, considerable improvements were made in the following years. In [130], a 16 x 26 array of SPADs shared 48 TDCs. With 3 TDCs per column, a peak PDE of over 6% and a fill factor of 21.2% was achieved. Later, a high fill factor using this approach was designed at 43.7% by the 256 x 8 SPAD array developed in [4]. Here, 8 SPADs in a column shared a single TDC through an *OR* tree. It should be noted that only a 256 x 4 subset of these SPADs are enabled at a given time, based upon whether the red or blue mode of the sensor is selected.

Although the row or column wise sharing of TDCs was a popular approach, even larger groups of SPADs were shared using more complex TDC triggering networks. In [2], arrays of 24 × 30 SPADs effectively shared a pair of two interleaved TDCs. Each 24 × 30 group

of SPADs was split into 12 × 15 SPAD arrays, and 3 adjacent SPAD outputs were passed to *OR* gates. This method sacrificed spatial resolution in order to achieve a high fill factor, since in the PET imaging application, the photon surface density is low. After the *OR* gates, each of their outputs was passed to a monostable. Here, the SPAD pulses are compressed in time in order to minimize the chance of an overlap between the outputs of separate SPADs. Thus, the photodetection dead time is reduced by a method known as temporal compression, which can also help to reduce pile-up effects for applications such as FLIM and NIROT. The monostable outputs can then be passed to a final *OR* tree and routed to the appropriate TDC. Using a multiplexer that allows for one interleaved TDC to be active while the other writes its result, the TDC dead time is avoided through ping-pong operation. This general scheme of 720 SPADs sharing a pair of 2 interleaved TDCs is replicated in an 8 x 16 array, resulting in a high fill factor of 42.6% across the detector (35.7% at the chip-level including I/O, power ring, and bonding pads), and a peak PDE of ~16% at a wavelength of 410 nm, and an excess bias of 1.5 V. This work indicated that the spatial and temporal compression approaches could provide a design alternative capable of achieving high fill factors.

An approach that could perhaps achieve the highest fill factors is through the use of 3D IC technologies. By designing the photodetection circuitry and the digital/mixed-signal circuits, including the TDCs to be in separate tiers, it is possible to obtain a sensor with a high fill factor, even when sharing very few SPADs to a given TDC. In [20], an array of 2 x 400 SPADs was implemented in the top tier, while time interval measurement was performed in a bottom tier of a 130 nm 3D CMOS process. Specifically, in the top tier the SPADs are grouped into 2 x 4 arrays, connected by through-silicon vias (TSVs) to processing blocks in the bottom tier. While in this work, conservative design choices were made in the SPAD design, a relatively high fill factor of 23.3% was still achieved. It was estimated by the authors that with optimal SPAD design from experienced designers, the fill factor could be made higher than 70%. The growing trend of 3D-stacked imagers (e.g. [143]–[145]) employing a large number of TDCs is expected to continue in the coming

years, and as such, optimal readout of the large amount of data generated by these TDCs should be considered.

2.3.3. Data Readout

In general, there were two main readout methods used in SPAD-based sensors for biomedical imaging. Frame-based sensor operation and readout were commonly used for applications such as FLIM, NIROT, and Raman spectroscopy, where the photon detections are correlated with a reference clock signal. One example was presented in [18]. Since a TDC was integrated 1:1 with each SPAD, a large amount of data needs to be transferred in order to generate high-speed images over the entire detector array. To achieve this, the 32 x 32 detector array was split into two 16 x 32 sections. A 1 μs data acquisition period was used, where the data stored from the previous frame would be serialized column-by-column and read out. Simultaneously, new data could be collected to allow for continuous data collection. The data was read out in a rolling shutter approach, which has the advantage of not requiring any additional address information, as the order in which data is passed is predetermined.

The second method of readout is based on an event-driven mechanism and is highly effective when photon detections are sparse [146]. While the frame-based approach can be used for obtaining high-speed time-correlated images across an entire pixel array, this approach does not allow for any reduction of readout data in low photon density situations. The benefits of event-driven readout are even greater when photon events are uncorrelated with the reference clock signal as in PET imaging [132]. In [130], a threshold was used to decrease the bandwidth requirements by using the fact that photon events must exceed a specified energy threshold in order to be considered valid gamma events in PET imaging. For output data to be generated, the number of photon events triggering the TDCs must exceed a specified threshold within a predetermined timeframe. Otherwise, the system will be reset since the event will be deemed invalid. A dual threshold approach was used in [2], which was able to discriminate gamma events from background noise. A bin clock was used, during whose period the photon events are counted. Using an initial threshold only slightly above the noise floor, the first photons of gamma events can be detected even if

they occur near the end of a bin clock period. Detecting the first photons of these gamma events can lead to a more accurate ToF estimation, resulting in improved images. The second threshold can then be made much larger, such that only true gamma events are recorded, and background noise is rejected.

A detailed mathematical derivation comparing frame-based and event-driven readout methods was performed in [147]. Using analytical expressions and numerical simulations, it was shown that in a frame-based readout architecture, a larger number of events could be recorded, but resource utilization is low. The event loss in the frame-based approach arises from only the first event in a pixel being recorded during the synchronous measurement interval, and a long clock period being needed to read out a large array. Conversely, the event-based readout scheme may save area and power by sharing readout resources more efficiently throughout the array when photon detections are sparse. In this approach, the sharing of resources is often the primary event loss mechanism. Also, in the event-based readout, the throughput will saturate more quickly at higher data rates and are therefore most useful when it is expected that data will be sparse over the detector array; but this is not always the case.

Based on these considerations, a hybrid of the frame-based and event-driven readout methods was implemented in [148]. This was termed as a router-based readout. Here, a set of timing lines are shared by all pixels within the array and connected to external data converters capable of operating at the same frequency as the laser. When an event occurs, the timing information is delayed, and selection logic that communicates with all pixels will determine which of the shared lines to connect to the timing data. The selection process is allowed to last longer than a clock cycle, as a pipelining approach is utilized to enable high throughput. In addition to traditional frame-based event loss, this approach also has loss in each pixel during the time in which the selection logic of that pixel is active. If more events occur than selection lines that are available, loss will manifest as well. This work could achieve higher throughput than even the frame-based approach when photon detections were less than 1% of the laser frequency. At higher rates, the throughput approaches that of the event-based approach, with the bandwidth being dominated by the

required address bits.

2.4. Conclusions

TDCs were a heavily studied research topic in recent years. While the common TDC methods apply different principles, they all attempt to balance the set of trade-offs illustrated in Figure 2-14. By choosing a metric to optimize from the left, potential trade-offs are shown within that row. For example, if the sampling rate (F_S) is to be optimized, the LSB will be worsened if the range is fixed. Optimizing F_S also inherently limits the dynamic range as you cannot sample faster than flash conversion. Higher sampling rates are also noted to give higher power consumptions because of high-speed clocks or parallelized conversions that utilize more hardware, which additionally increases the size. By observing these trade-offs, a designer should identify the requirements of the targeted applications and aim to balance the key performance requirements. Additionally, we expect that new TDC topologies will continue to be developed to reduce the impact of some of these trade-offs, similar to how recent pipelined TDCs have allowed high-throughput TDCs to be designed while maintaining fine resolution.

		Affected Metric					
		LSB	**DR**	**Fs**	**INL**	**Power**	**Area**
Optimized Metric	**LSB**		↓ for fixed area	↓ for fixed DR	↑	↑ for fixed DR	↑ for fixed DR
	DR	↓ for fixed area		↓	-	↑ for fixed LSB	↑ for fixed LSB
	Fs	↓ for fixed DR	↓		-	↑	↑ (pipelining)
	INL	↓	-	-		↑	↑
	Power	↓ for fixed DR	↓ for fixed LSB	↓	↑ (no DLL)		-
	Area	↓ for fixed DR	↓ for fixed LSB	↓ (no pipelining)	↑ (no DLL)	-	

Figure 2-14: Graphical summary of the general TDC trade-offs (© 2021 IEEE).

Due to their capabilities to be integrated with high-performance photodetectors known as SPADs, TDCs can be used to timestamp individual photon detections in dSiPMs and SPAD imagers. SPAD-based sensors integrated in CMOS technology provide the capability to readily commercialize high-performance and low-cost biomedical imaging sensors for applications such a PET, FLIM, and DOT. In this chapter, we reviewed the

fundamental principles and performance metrics of CMOS TDCs. In addition, we performed a detailed analysis of state-of-the-art CMOS TDCs, and the integration of TDCs with SPAD arrays to form dSiPMs and SPAD imagers.

Chapter 3
Time-to-Digital Converter Using Feedback Time-Amplification

3.1. Operating Principle

As seen from the previous chapter, state-of-the-art TDCs can achieve resolutions in the order of picoseconds. However, TDCs integrated with SPAD arrays often use simpler topologies such as the delay line interpolation of a DLL or basic ring oscillators since higher resolution TDC structures were shown to occupy larger silicon areas. Therefore, in this work, we aimed to design a TDC capable of achieving fine resolution while minimizing the potential negative impacts on the fill factor if integrated with a SPAD. To achieve this goal, a multi-stage TDC structure was used in a feedback configuration where the first TDC stage and the time amplifier were reused through a multiplexing scheme to improve the area efficiency. A block diagram of the proposed TDC is depicted in Figure 3-1, and a simplified diagram of the timing path is provided in Figure 3-2.

The time difference between the *start* and *stop* signals are converted to a pulse width on a single line and routed through a 2:1 multiplexer (MUX) as the enable signal to the gated-delay line (GDL) of the upper TDC stage. While the enable pulse is high, propagation of a rising edge is enabled in the GDL, and the phase of the GDL is held after the falling edge of the pulse. The control logic is designed to sample the state of the first delay line, and the sampler result can be encoded to give the most significant bits (MSBs) of the output code. The sampler state is also used to generate the remainder by once again enabling the GDL and generating another pulse width equal to the time it takes for the rising edge to reach the second-next delay element. The second-next delay element is used to generate the

remainder instead of the next delay element in order to ensure the pulse width is wide enough to be accurately generated by the remainder generation logic, which uses an SR-latch for pulse generation.

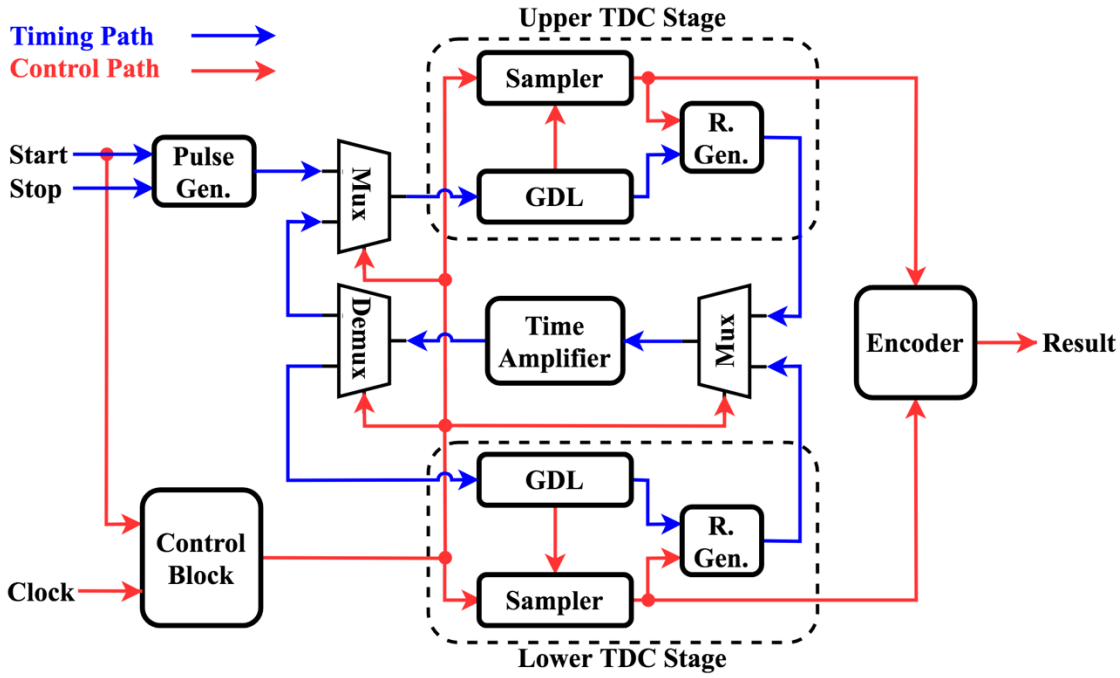

Figure 3-1: Block diagram of the proposed feedback time amplification TDC.

The remainder that is generated from the upper TDC is then routed through a 2:1 MUX by the control logic and passed to the pulse-train time amplifier (TA). The pulse-train TA takes a pulse width as input and creates 8 non-overlapping replica pulses with the same width. These replicated pulses are passed through a 1:2 demultiplexer (DEMUX) as the enable signals to the lower TDC stage. This results in the time residual being amplified by 8 times, as the GDL in the next TDC stage accumulates the width of all 8 pulses. The state of the lower TDC's GDL is sampled by the control logic, and the result is encoded to generate the medium-resolution bits of the output code, and to generate the next remainder.

The control block then switches the select inputs on all the MUXs and DEMUXs. This allows the remainder of the lower TDC stage to be routed to the TA, and then undergo the third and final conversion in the upper TDC to give the least significant bits (LSBs) of the output code. This MUX/DEMUX scheme effectively shares the TA, which is one of the

largest sections of the circuit. Additionally, it reduces the need for a third TDC stage that would otherwise be required in this topology for the three levels of conversion that are achieved. While this topology would theoretically allow for repeated conversions on remainders from each stage, here we designed for only three levels of conversion, as the nonlinearity performance will degrade for a larger number of remainders that are time amplified. The optimal level of conversions for a given application will depend on the required resolution and the tolerable level of nonlinearity.

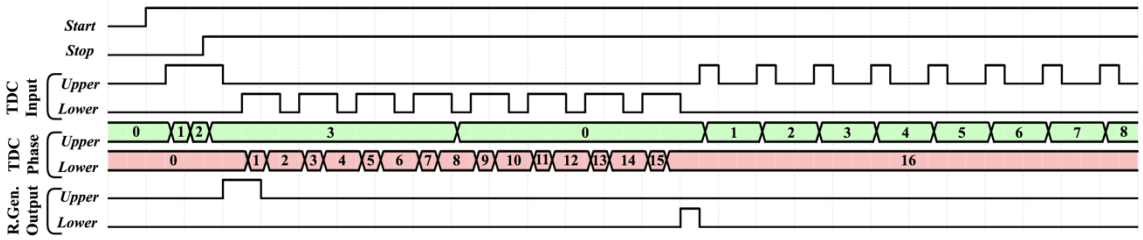

Figure 3-2: Timing diagram of the proposed feedback time amplification TDC.

3.2. Circuit Design

In the previous section, the general operation of the TDC was discussed using the block diagram in Figure 3-1. Here, we will discuss the hardware implementation of the main blocks in detail, highlighting additional design considerations that needed to be made at the circuit level.

3.2.1. Control Block

An illustration of the TDC control block is shown in Figure 3-3. The function of this block is to allow for synchronous operation of the TDC with respect to a reference clock. The control block uses an edge detection circuit to generate pulses on the rising and falling edges of a 200 MHz reference clock. These edges increment a 4-bit ripple counter, whose states are passed to a combinational logic network to generate the signals SET_1, SET_2, RST_1, RST_2, RST_{master}, and SEL.

The *SET* signals are applied as the inputs to the GDLs, and transition high shortly before the input pulse width is applied to the GDL enable input. The rising edge of each *SET* signal will also clock the sampler array in the opposite TDC stage. A delayed version of each *SET* signal, generated by passing the signal through a small number of buffers, is used to initiate the remainder generation logic. The delay is used to ensure that the remainder is not generated until the state of the sampler array is stabilized (i.e., after a buffer time to avoid metastability of the DFFs).

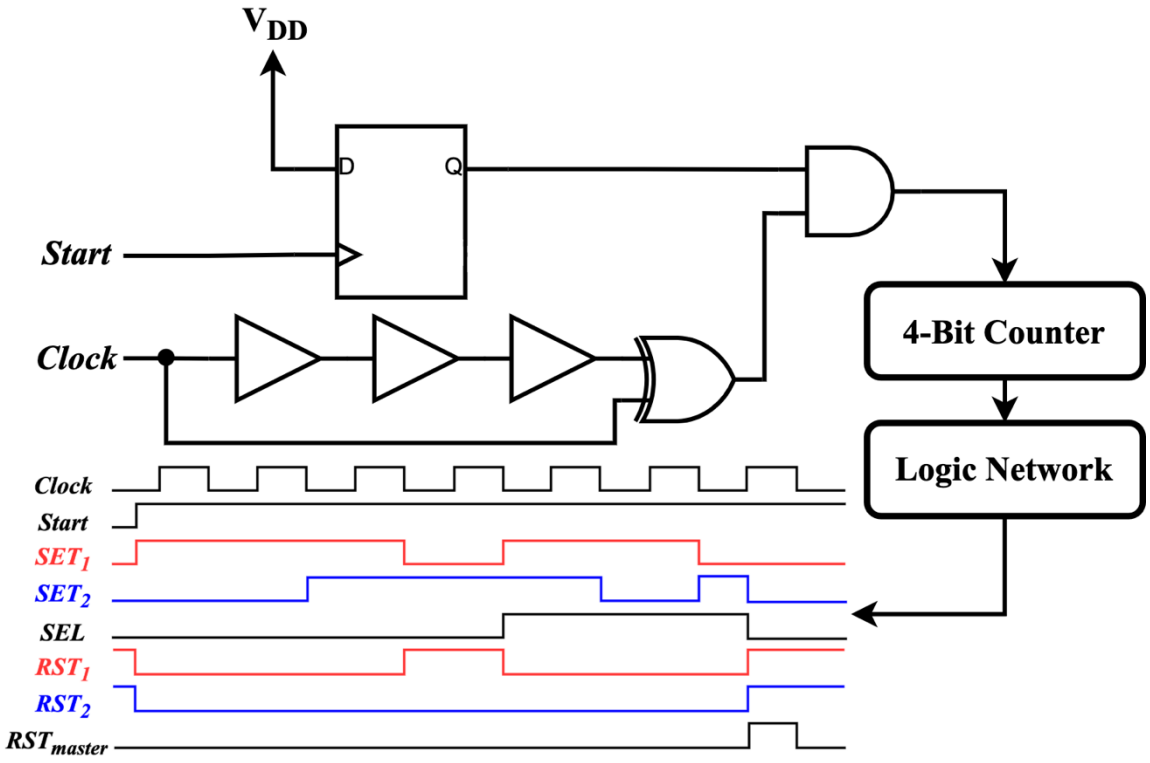

Figure 3-3: Simplified schematic of the control block.

The RST_1 and RST_2 signals are used to ensure the state of the GDLs are cleared before the start of a conversion. Since the first TDC stage will quantize the initial input pulse and the second remainder, it is reset during the time in which the second TDC stage completes its conversion. Since in this implementation the second TDC stage only completes one conversion during the measurement cycle, the RST_2 signal is not required. However, it was included in the control block in case this work was expanded to employ a larger number of feedback cycles. The TDC also employs a signal RST_{master}, whose output comes from a

monostable, that generates a reset pulse that is passed to all flip-flops within the TDC after a measurement is completed. The RST_{master} pulse can also be generated from a momentary pushbutton switch on the PCB to ensure the TDC is in a known state at power-up.

Lastly, the control block generates a signal called *SEL*. This signal is used for controlling the select inputs of all MUXs and DEMUXs in the TDC. This allows the TDC to be configured in a feedback loop, reducing the number of delay lines and sampling circuits required compared to the feedforward TDC approach. This also means that only a single TA is required, which reduces the area requirements.

3.2.2. Digitally-Controlled Gated Delay Line

Since the pulse train TA output must be routed to a GDL, the delay cells in this design are formed from gated buffers, as shown in Figure 3-4. The buffer input is set to V_{DD} during the conversion by the *SET* signal in the control logic, and the time intervals are applied as pulses on the delay cell enable inputs. The buffer delay is set using an 8-bit digital delay bias circuit as in [149], which is illustrated in Figure 3-5. Digital control signals which are connected to external pins of the chip are routed to the gates of 8 binary sized PMOS transistors. Each 8-bit digital code will result in a unique current being sourced to the NMOS current mirror, generating the analog control voltages for the delay line V_{cn} and V_{cp}. The achievable delay range of this circuit was kept large to ensure the desired delay is achievable in all process corners and that the delay line can be finely tuned such that the resolution is well controlled. The delay bias circuit is shared between both TDC stages and should be replaced by a delay-locked loop to reduce jitter in a final application. Both GDLs were designed for a ~5 ns delay across 16 delay elements, resulting in ~312.5 ps delay per stage. However, for the purposes of testing, two additional delay elements were placed at the end of each GDL to ensure that any nonlinearity in the remainder generation does not result in amplified remainders that exceed the dynamic range of either TDC stage.

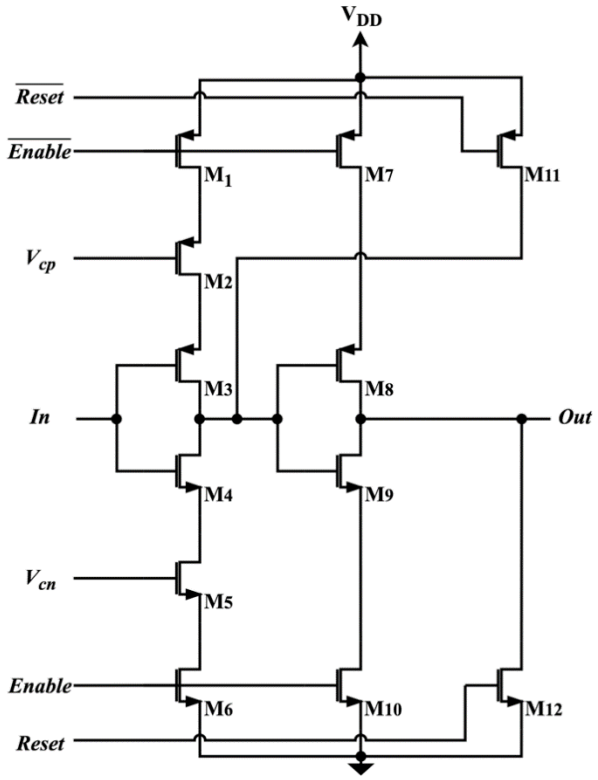

Figure 3-4: Schematic of the current-starved gated delay cell that is replicated to form the gated delay lines.

A replica of the TDC delay lines was placed on the chip to perform initial calibration and determine the correct 8-bit bias code that would give the desired delay. As each stage of the GDL within the TDC is capacitively loaded by an arbiter from the sampler array and a buffer to the remainder generation logic; dummy cells of the arbiter and buffer were included in the delay line replica to ensure the same capacitive loading for the calibration procedure.

In the delay cell is shown in Figure 3-4, M_3, M_4, M_8, and M_9 are two series-connected inverters that construct a buffer. The transistors M_2 and M_5 limit the first inverter's switching current, such that the propagation delay through the buffer is dependent on the voltages V_{cn} and V_{cp} that are generated by the digital delay bias circuit in Figure 3-5. Gating is achieved by means of transistors M_1, M_6, M_7, and M_{10}. These transistors are turned on while the input pulse width to each stage is active, and otherwise isolate the inverters from V_{DD} and GND in order to hold the delay line state. After each conversion result is

determined, the delay line can be reset through the transistors M_{11} and M_{12} by a reset pulse from the control logic. As this implementation aimed to achieve small size, the minimum length is used for all transistors in the delay cell, and the width is kept just above the minimum for the NMOS. The PMOS width was sized for symmetric rise and fall times (~2.5 times the NMOS width for these buffers).

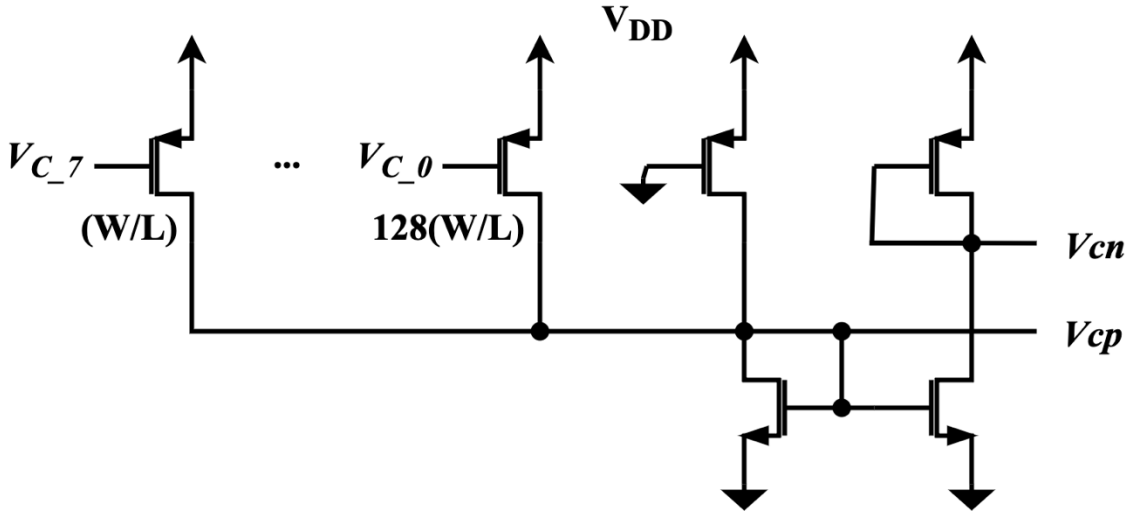

Figure 3-5: The digitally-controlled delay line (DCDL) biasing circuit.

3.2.3. Input Pulse Generator

A simplified depiction of the input pulse generation stage is shown in Figure 3-6. In order to convert the rising edges of the TDC *start* and *stop* signals to a pulse width, a *NOR* gated SR-latch was used (note that Figure 3-6 does not show the full gated latch). The gating was required in order to guarantee that the latch is in a known state prior to the conversion, due to the indeterminate state of the conventional NOR SR-latch when both inputs are low. Additionally, the finite rise and fall times of the latch limit the ability of this circuit to convert narrow input time intervals into reliable pulse widths. Therefore, to avoid the narrow pulse width issue, a 1 phase offset (i.e., the delay of 1 element of the GDL) is added to the input time before being sent to the latch. This offset is added by using replicas of the GDL elements at the input stage, with dummy cells that replicate the capacitive load seen by the delay element within the GDL to ensure consistent delay. Since the offset is equal to 1 phase of the GDL, it is easily removed by post-processing to give the final result.

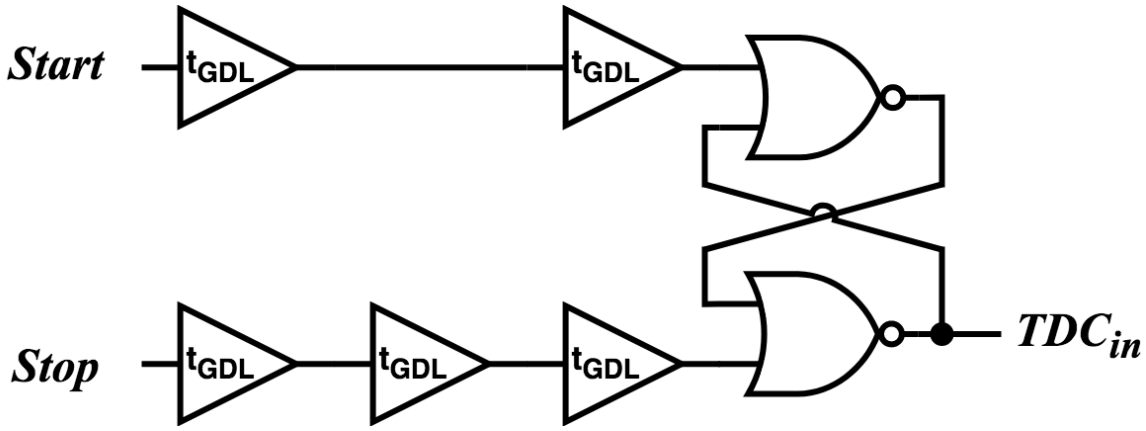

Figure 3-6: Schematic of the input pulse generator.

3.2.4. Sampler Array

In the TDC, an array of samplers is used to store the state of the GDLs for each stage of the conversion. The samplers are constructed as shown in Figure 3-7 and are connected to the output of each stage of the GDLs. Since the layout of DFFs have asymmetric propagation delays for the data and clock signals, symmetric arbiters can more accurately determine the order of arrival of two signals. A *NAND* SR-latch based arbiter is used in this implementation [150].

Initially, the output of both *NAND* gates are high, and thus the inverter outputs are low. The rising edge of either *Phase*[*i*] or *SET* causes the output of the respective *NAND* gate to go low, and then the output of the connected inverter goes high. Since the *NAND* gate's output sets the source voltage of the PMOS in the opposite inverter to GND, the second inverter's output is blocked from transitioning high. The *SET* signal passes through a delay in the control block, such that it clocks the DFF and stores the state shortly after the arbiter result has settled from its metastable state. True single-phase clock (TSPC) DFFs are used in this circuit due to their low transistor count and resulting compact size.

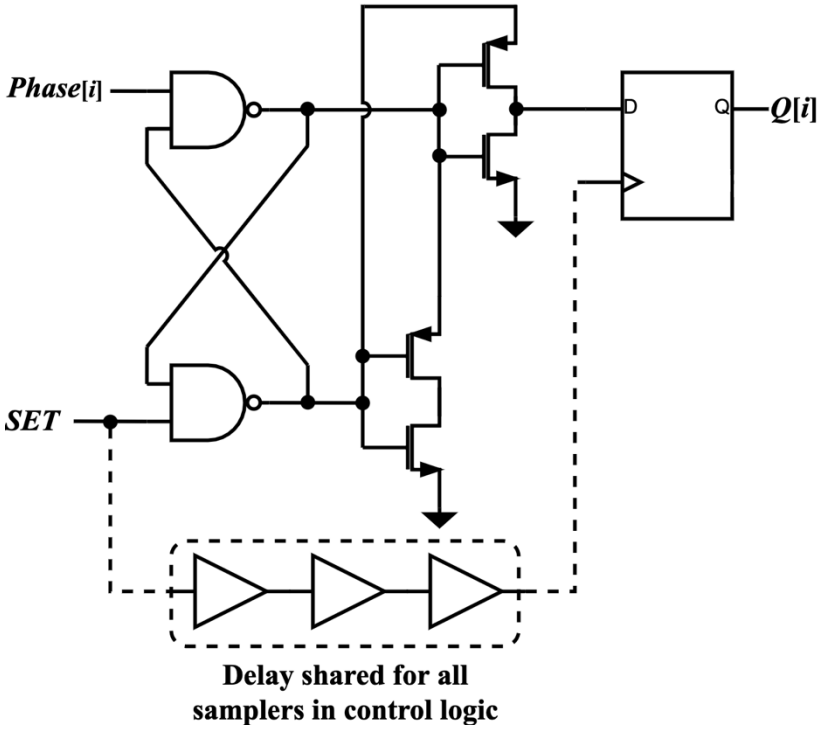

Figure 3-7: Schematic of the sampler, consisting of an arbiter and DFF, that is replicated to sample the state of each gated delay line.

3.2.5. Remainder Generation Logic

After a time interval was applied to the GDL and the result was sampled, the remainder needs to be generated so that it can be time amplified and quantized in the next TDC stage. The complete remainder generation logic is shown in Figure 3-8 and consists of 16 remainder generation cells (i.e., one for each phase of the delay line), whose outputs are routed to a 16-input *OR* gate. The remainder generation is initiated by the delayed *SET* signal of the previous stage, which will pass through a duplicate of the remainder generation logic in order to have symmetric propagation delays in each path so that no offset is added to the remainder, which contributes to the nonlinearity of the TDC. This causes a rising edge on the output of the SR-latch. At the same time, neighbouring results from the samplers of the previous stage are each passed to *XNOR* gates, whose active low signals will enable the inverter that passes the second next rising edge from the delay line. The

signal from the delay line is buffered to improve its drive strength and, due to the *OR* gate, will generate a single rising edge on the R input of the SR-latch. The result is an output pulse whose width corresponds to the remainder of the previous delay line.

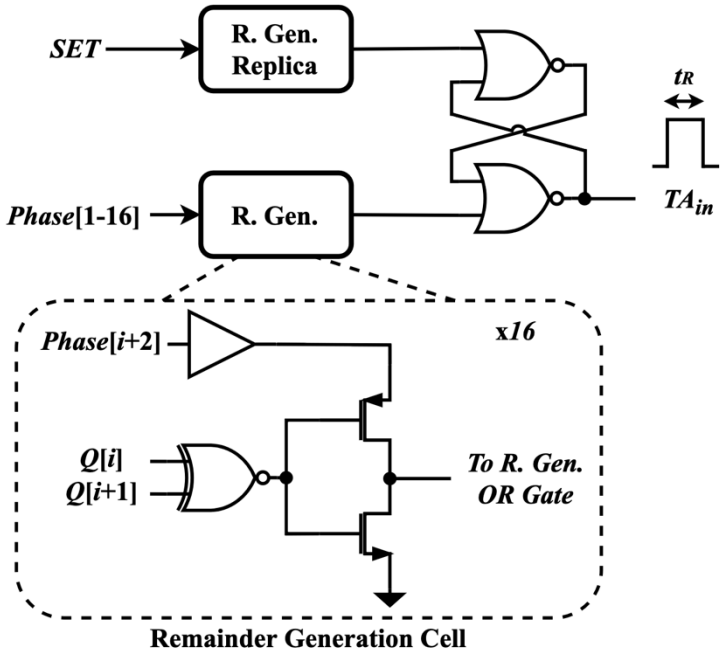

Figure 3-8: Schematic of the remainder generation logic.

3.2.6. Time Amplifier

In this work, a pulse train TA is used as it was previously demonstrated to provide linear time amplification over a wide input range without calibration [115]. The pulse train TA also has the benefit of being based on standard digital logic cells, providing ease of implementation. The principle of the pulse train TA is that by replicating a pulse and applying the replicas as the enable inputs to a GDL, the width of all replicas is accumulated. The result is time amplification of the input pulse width by the number of replicas.

The structure of the pulse train TA is shown in Figure 3-9. To generate the replicas of the input pulse, a series of delays are tapped and applied as inputs to an 8-input *OR* gate. In order to ensure that the replicas are nonoverlapping, the delay elements are designed to have a longer delay (t_{TA}) than the maximum input pulse width (t_R). To achieve this across

all process corners in the post-layout simulation, each delay cell in the pulse train TA is composed of the series connection of 32 standard buffer cells. Since the goal is to achieve a time amplification factor of 8, 7 delay blocks are needed, resulting in a total of 224 buffers. A benefit of the pulse train TA is that mismatch or jitter in t_{TA} do not affect the linearity or jitter of the time amplification since it only changes the spacing between the replicas, not their pulse width.

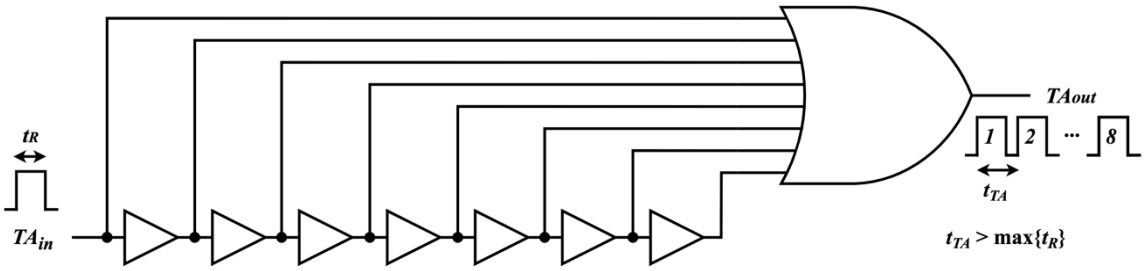

Figure 3-9: Schematic of the pulse-train time amplifier.

3.3. Measurement Results

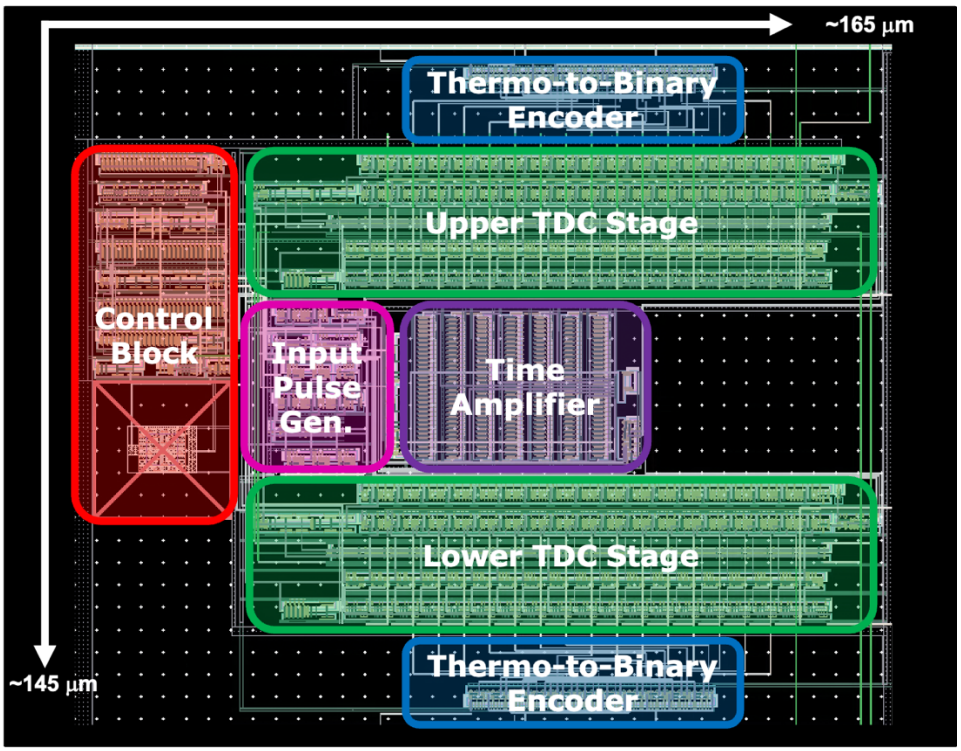

Figure 3-10: Annotated layout of the complete TDC in the TSMC 65 nm CMOS process.

In this subsection, the results of the proposed TDC are presented for a design that was fabricated in the TSMC 65 nm process. The total square area of the TDC is 0.024 mm^2; however, the core circuit area occupies only ~0.016 mm^2. This would indicate that this structure is competitive in size with the most compact topologies of high-resolution TDCs. It should also be noted that the TDC was designed entirely of custom cells, but it could be significantly reduced in size by using the optimized standard cell libraries provided by TSMC, which were unavailable at the time of the design.

3.3.1. Measurement Setup

A custom PCB was designed for the measurement of the TDC. The clock, *start*, and *stop* signals are input to the board using SMA connectors and 50 Ω termination resistors to avoid reflections from an impedance mismatch. 6 dB RF attenuators are used to reduce the 2 V minimum output of the delay generator to the 1 V required by the TDC chip. To supply the input code for the DCDL, an 8-bit DIP switch was included on the board. At the output of the TDC, buffer ICs are used to improve the drive strength of the TDC outputs to ensure they can drive the large capacitive load of the scope probes.

To measure the performance of the TDC, the setup illustrated in Figure 3-11 was used. A 1 V bias for the TDC was provided by an Agilent E3646A DC power supply. The clock was generated using a PLL from a Xilinx Spartan VI FPGA evaluation board, and the TDC *start* and *stop* inputs came from a Berkeley Nucleonics Model 745 Digital Delay Generator. Lastly, a Lecroy Waverunner 625Zi mixed-signal oscilloscope was used for collecting and saving the data for further analysis. For all measurements, the delay generator and oscilloscope were programmed using MATLAB in order to automate the measurement process due to the significant time required to fully assess the performance of the 10-bit TDC.

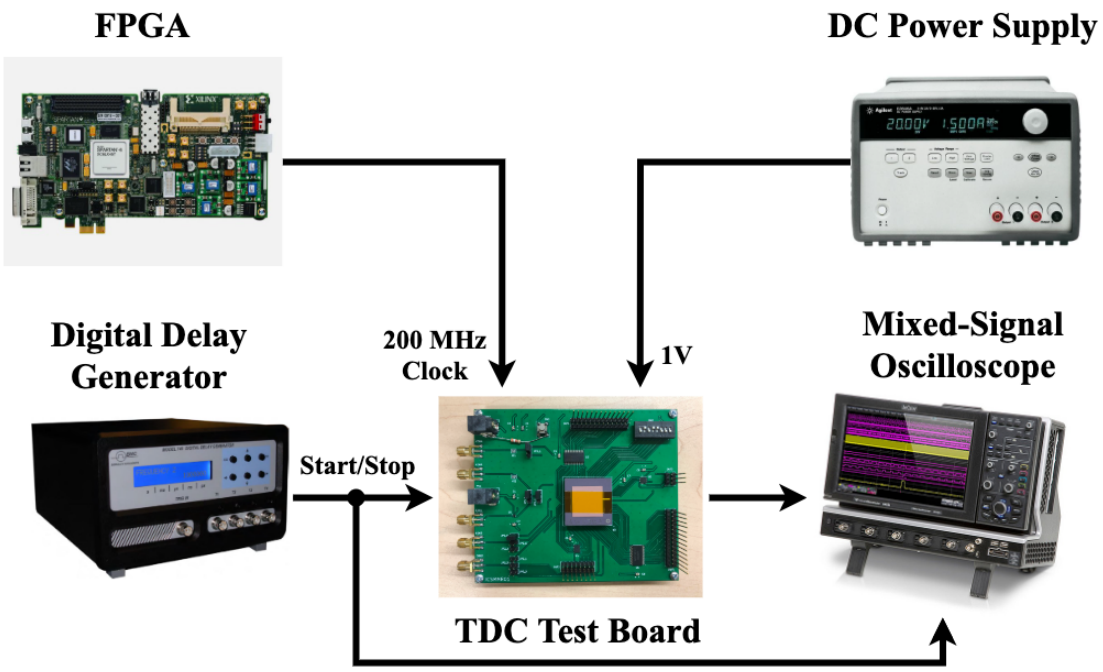

Figure 3-11: Block diagram of the measurement setup used for the TDC characterization.

3.3.2. Results and Discussion

The first measurement that was conducted was to assess the delay response of the DCDL. As the delay line takes a digital code as input from an 8-bit DIP switch, not every code was measured, and only a coarse measurement was performed spanning from the minimum to maximum codes. It can be seen in Figure 3-12 that the delay line can achieve delays ranging from ~2.75 ns to 26 ns. While at low input codes, the delay increases very quickly, within the range of delays below 5 ns that the TDC was designed to operate in, the resolution of the DCDL is ~13 ps/code. This allows for the fine-tuning of the TDC resolution within the measurement range.

While the DCDL is capable of achieving delays with fine resolution within the desired range, it should be noted that the jitter of the delay line is still quite high since it is not locked by a DLL. Figure 3-13 shows the distribution of delay measurements for several input codes. The average jitter of the delay line was measured to be 62.3 ps. This is higher than the target resolution of 4 to 5 ps but could be reduced in the future if a DLL was used.

As such, the TDC can still achieve a high resolution but suffers from high jitter, which will be seen from the precision measurement results.

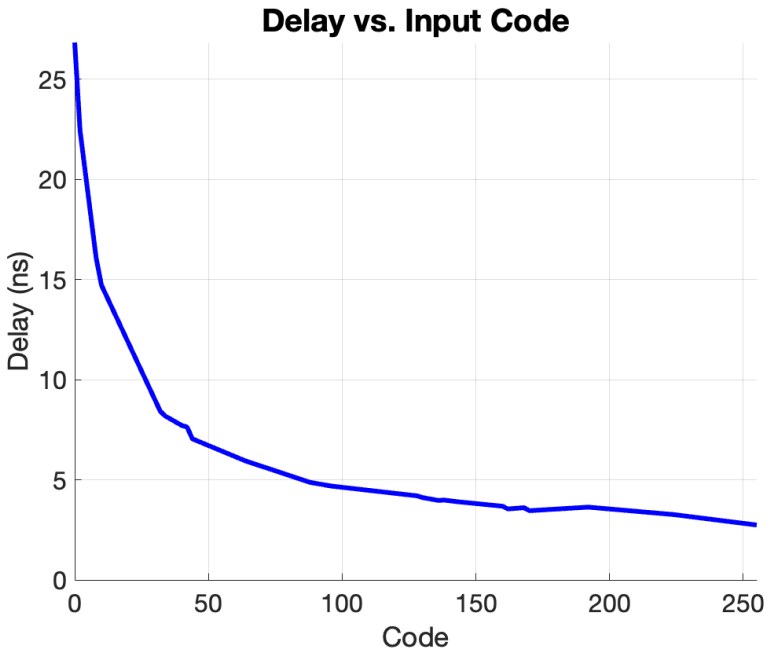

Figure 3-12: Delay vs. bias code for the DCDL.

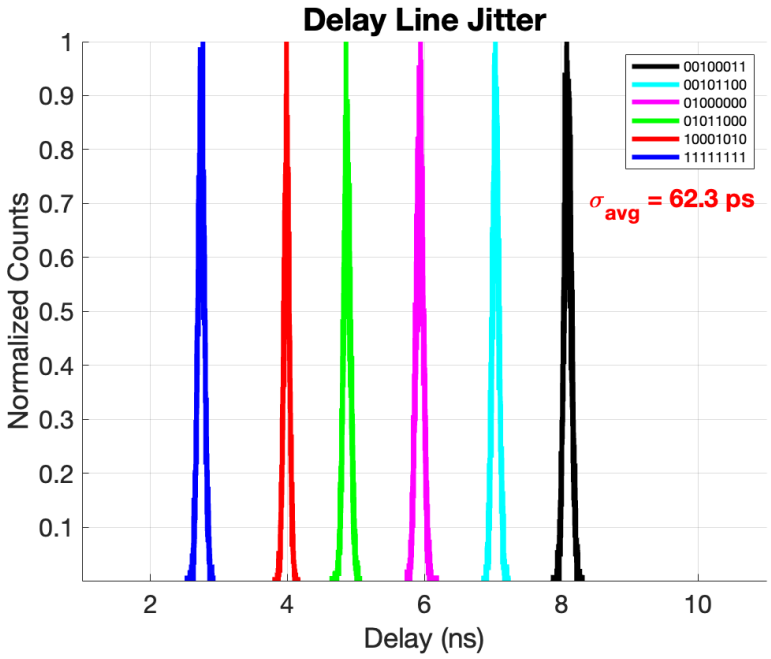

Figure 3-13: Results of the DCDL jitter for several input codes.

The most fundamental measurement for the TDC's performance is the statistical code density test, which determines: the resolution, dynamic range, and nonlinearity performance. The statistical code density test applies a large number of uniformly distributed inputs to the TDC. In this measurement, ~500000 measurements were taken, which gave a 97% confidence level and a 10% tolerance level [151]. Since the input time intervals to the TDC are uniformly distributed across the entire dynamic range, each output code should ideally occur with equal probability. Due to nonlinearity within the TDC, certain codes will occur more frequently during the measurement. Using equations (3-1) and (3-2), the step widths for each bin of the TDC can be computed and summed to give the cumulative distribution function (CDF). Here, τ_n is the i^{th} step-width, DR is the dynamic range of the measurement, N_i is the number of counts observed for that code, N_{total} is the total number of counts, and $H(n)$ is the CDF. From the CDF, we can plot the TDC's full quantization characteristics where the average step width gives the resolution, and the difference between the first and last codes gives the dynamic range. The difference between each step width of the TDC response and the resolution gives the DNL, and its integral (i.e., the deviation of the TDC response from the best-fit line) gives the INL.

$$\tau_i = \frac{DR \times N_i}{N_{total}} \tag{3-1}$$

$$H(n) = \frac{\tau_i}{2} + \sum_{i=1}^{n-1} \tau_i \tag{3-2}$$

The result of the statistical code density test is shown in Figure 3-14. The TDC achieves a resolution of 4.14 ps over a dynamic range of ~3.2 ns. Due to increased nonlinearity at the start and end of the response, the ends were slightly truncated before fitting the result to a best-fit line. The outcome is that the number of TDC codes is effectively reduced from 10-bits to ~9.6 bits, degrading the dynamic range. Even with the correction at the endpoints, the TDC deviated significantly from the expected result due to high nonlinearity throughout the response. Figure 3-15 qualitatively shows the TDC's nonlinearity. The DNL was determined to be 0.671 LSB_{rms}, while the INL was determined to be 15.3 LSB_{rms}.

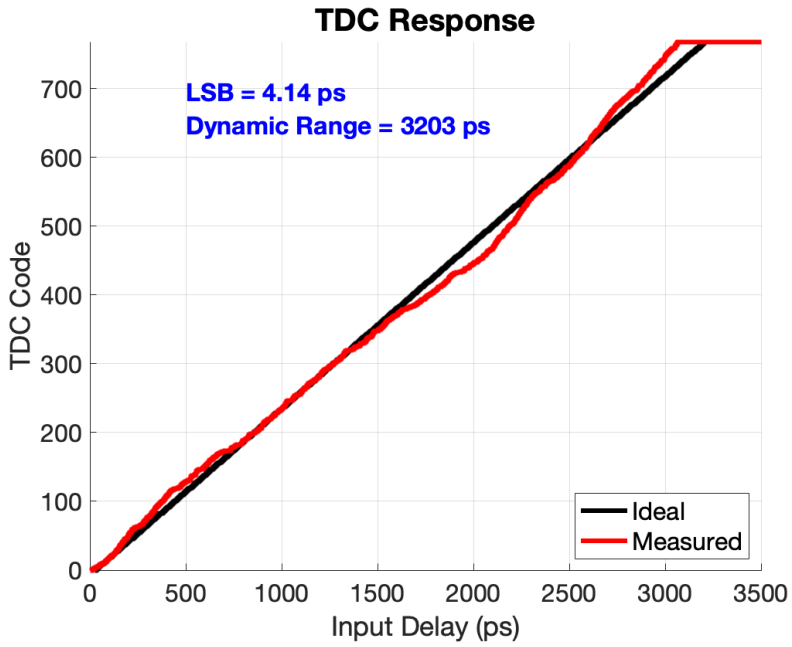

Figure 3-14: Quantization characteristics of the 10-bit TDC, determined using a statistical code density test.

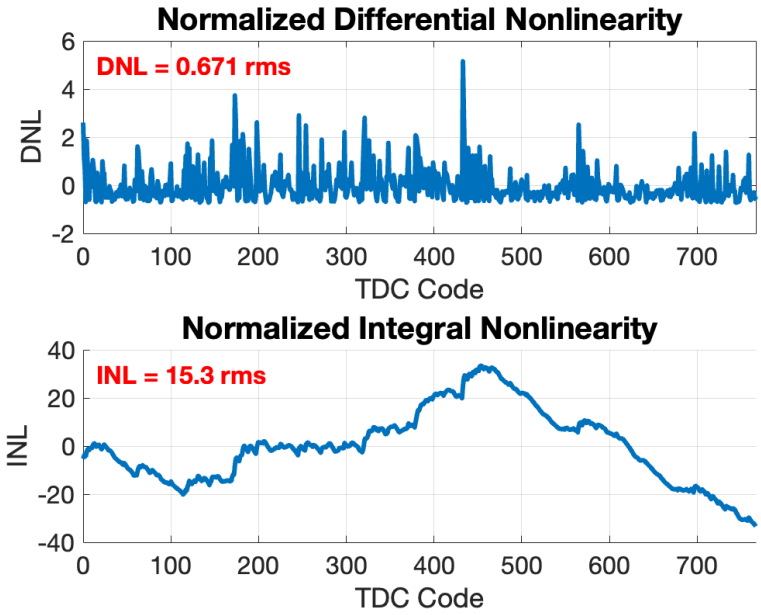

Figure 3-15: Nonlinearity performance of the 10-bit TDC.

Since the TDC computes the 10-bit output by combining the result of 3 separate measurements (i.e., 4-bits in the first conversion, and 3-bits in each of the second and third

conversions), the TDC response was also considered in the 4-bit and 7-bit cases in order to assess the functionality of the first two stages of operation. In the 4-bit case of Figure 3-16, it can clearly be seen how the TDC closely follows the expected response. The resolution (LSB) is reduced to 255 ps in this case. However, the nonlinearity near the start and end of the response is reduced, and the dynamic range is improved to ~4.1 ns as the ends of the response no longer need to be truncated due to excessive nonlinearity as in the 10-bit case. Specifically, the DNL and INL are improved to 0.231 LSB_{rms} and 0.245 LSB_{rms}, respectively.

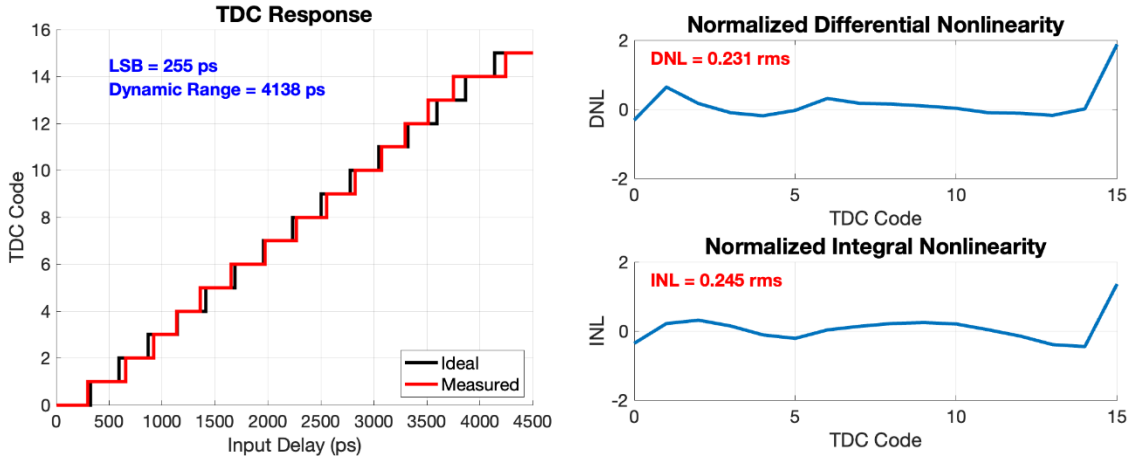

Figure 3-16: Response of the 4-bit TDC obtained by truncating the 6 LSBs. Effectively, this uses 1 of the 3 available conversion results.

In the 7-bit case of Figure 3-17, the TDC still achieves adequate results in terms of linearity, with a DNL of 0.352 LSB_{rms} and an INL of 1.29 LSB_{rms}. This indicates that the remainder generation logic of the TDC is functioning correctly, as it is capable of generating the first stage remainder for quantization in the second stage. Here, the TDC achieves a resolution of 38.8 ps with a dynamic range of 3.8 ns that is only slightly reduced by truncating the ends of the response to effectively achieve a ~6.6-bit response. Based on the results in the 4-bit, 7-bit, and 10-bit cases, it can be concluded that each component of the TDC proposed in the block diagram of Figure 3-1 is functioning properly but fails to maintain adequate nonlinearity performance when integrated together to form the complete 10-bit TDC. Potential areas for improving the linearity in a subsequent iteration will be outlined in the conclusions section of this chapter.

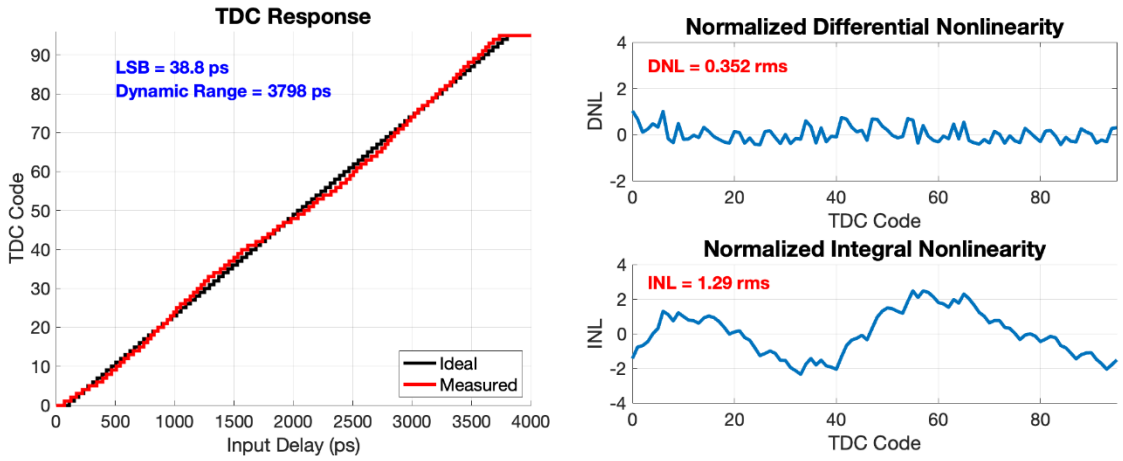

Figure 3-17: Response of the 7-bit TDC obtained by truncating the 3 LSBs. Effectively, this uses 2 of the 3 available conversion results.

The final measurement that was performed on the TDC was the single-shot precision test. Using the same setup illustrated in Figure 3-11, the distribution of TDC output codes for constant input times was generated for several inputs across the dynamic range, and 2500 measurements were taken in each setting. The standard deviation of the distribution is then plotted for each delay setting in the 4-bit, 7-bit, and 10-bit cases to give Figure 3-18. This result further confirms that the final 3-bits of the TDC are not reliable and suffer from large measurement errors. While changing the TDC from 4-bit to 7-bit operation improves the precision by a factor of 2, the 10-bit case shows no significant improvement over the 7-bit case. In addition to the delay line jitter that was illustrated in Figure 3-13, the TDC precision may be degraded by leakage of the gated delay line stages during the time in which they are disabled. Since the TDC *start* and *stop* signals arrive asynchronously, the first stage of the TDC conversion will have to hold the charge associated with the coarse measurement result in the gated delay line for varying amounts of time. The difference in hold time of this charge in the presence of leakage current manifests as variation in the remainder that is passed to the next stage and degrades the precision. As the second remainder is generated synchronously, this issue is only present in the initial stage of the operation when the conversion begins asynchronously.

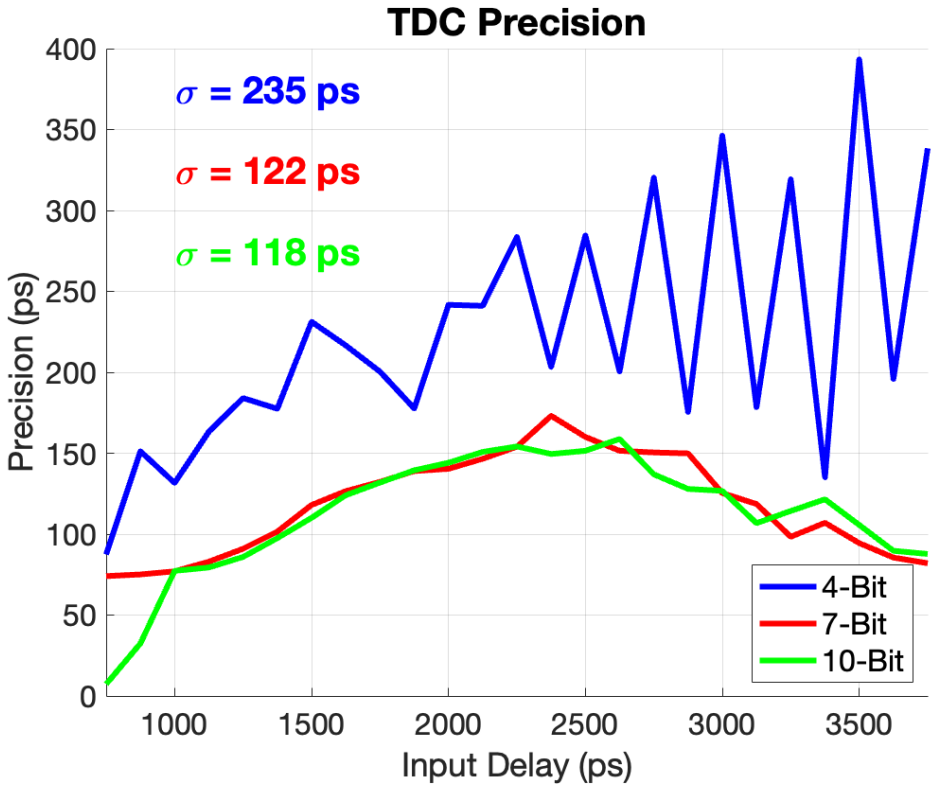

Figure 3-18: Comparison of the TDC precision in the 4-bit, 7-bit, and 10-bit cases.

3.4. Conclusions

A summary of the TDC performance compared to similar works is provided in Table 3-1. The proposed TDC achieves a resolution of 4.14 ps over a dynamic range of 3.2 ns. Due to a lack of I/O pins on the prototype chip, the TDC power consumption was estimated from the post-layout simulation as 1.87 mW at a 25 MHz sampling rate. Despite achieving fine resolution in a very compact area of 0.016 mm^2 with relatively low power consumption, the fabricated TDC in the 10-bit case suffers from poor INL and single-shot precision performance of 15.3 LSB_{rms} and 118 ps, respectively. By truncating the result to include only the first and second conversions, the TDC demonstrated a resolution of 38.8 ps over a 3.8 ns dynamic range, with comparable precision to the 10-bit case of 122 ps. Here, the nonlinearity is improved with a measured DNL of 0.352 LSB_{rms} and a measured INL of 1.29 LSB_{rms}.

Table 3-1: Comparison of TDC performance to published works.

Ref.	[115]	[112]	[110]	[52]	[113]	This
Type	Two-step (GDL)	Pipelined (GDL)	ΔΣ (GDL)	Feedback TA (VRO)	Two-step (DL)	Feedback TA (GDL) (7-bit/10-bit)
Tech. (nm)	65	65	65	65	180	65
LSB (ps)	3.75	1.12	5	0.98/6.01	5.3	38.8/4.14
DR (ns)	0.476	0.578	0.7	5.76	1.3	3.8/3.2
INL (LSB)	2.3	1.7	-	2.2	2.8	1.29/15.3
F_S (MHz)	200	250	50	10/250	30	33
Power (mW)	3.6	15.4	3.5	3.0/17.5	1.1	1.87 (25 MHz)
Area (mm^2)	0.02	0.14	0.09	0.02	0.05	0.016

Due to some limitations of the initial design, the following improvements should be made to optimize TDC in terms of its nonlinearity performance and precision:

- The current design aggressively targeted compact size and low power by using transistors very close to the minimum feature sizes. The trade-off of this design choice is that process variations in the fabrication process will have a relatively larger effect. As this TDC structure was implemented in a significantly smaller area than most works of similar resolution, sizing up the transistors in the design to minimize process variations may help in obtaining consistent results across the chips. Additionally, as most of the TDC is built from standard logic cells, if a standard cell library was utilized for a subsequent design, it would likely be possible to implement the design in a smaller area while still increasing the base transistor size.
- The DCDL in the prototype TDC was measured to have a significant average jitter of 62.3 ps in comparison with the target resolution of 4-5 ps. For optimized single-shot performance, the delay lines of both TDC stages would ideally be locked by a DLL to a reference clock. Since a 200 MHz clock is used for the control logic, the same clock could be used in the DLL to achieve the desired 5 ns range. If a DLL is not used, then increasing the aspect ratios of the transistors in the timing-sensitive

paths and minimizing any leakage in the delay elements while they are holding the remainder of previous conversions can help to minimize the jitter.

- Further considerations should be made during the TDC layout to ensure optimal matching of all timing-sensitive signals. This is particularly important for the routing of both TDC stages to the shared time amplifier, and in the remainder generation logic. The accuracy of these circuit blocks is crucially important for optimizing the nonlinearity performance and maintaining consistent step-widths along the quantization characteristics.
- To improve the resolution through time amplification, gated delay lines were used for the TDC stages to accumulate the replicated pulses from the time amplifier. Due to the gating of the delay cells, an effect that should be considered is the charge injection that occurs when the delay line is gated off. The charge that is within the channel of the gating transistors of the delay elements at the instant of switching cannot remain in the channel and will exit through either the drain or source. This excess charge modulates the phase of the delay line, resulting in errors in the amplified remainder from the previous stage. The delay element could be modified to include a switching mechanism that can minimize the effect of charge injection [152]. Another potential solution to the charge injection issue would be to use a time amplifier that does not rely on switching, such as an SR-latch based time amplifier. However, while this alleviates the charge injection error, SR-latch based time amplifiers frequently require calibration and tuning to achieve the desired amplification factor and may suffer from small input ranges.

Chapter 4
Design of Multi-Time-Gated SPAD Arrays

4.1. Operating Principle

Time gated (TG) SPAD pixels can be viewed as a subset of AQR SPADs that operate only within a specified time window (i.e., the time gate). A simplified schematic of a TG SPAD is shown in Figure 4-1. Initially, the quench switch is closed, and the SPAD bias is held below breakdown. Just prior to the start of the gate window, the quench switch is opened, and the precharge switch is momentarily closed to bring the SPAD bias beyond its breakdown voltage. When the precharge switch is opened, the SPAD bias is maintained by the capacitance at the SPAD's cathode. To initiate the start of the gate window, the gate switch is closed, enabling any photon detections or dark noise from the SPAD to be passed to the output. At the end of the gate window, the gate switch is once again opened, and the quench switch will be closed such that the SPAD is below breakdown and unable to generate output pulses. A summary of the results achieved for time-gated SPAD arrays is shown in Table 4-1.

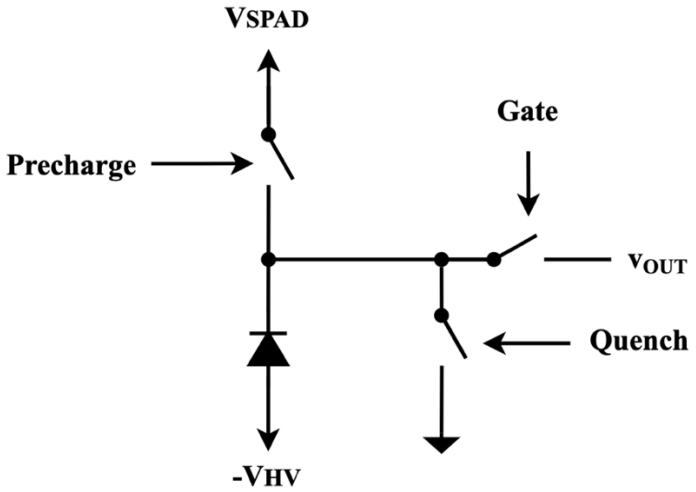

Figure 4-1: Conceptual diagram of a SPAD pixel with time-gated front-end circuitry.

Table 4-1: Summary of the results of time-gated SPAD arrays.

Year Ref.	Tech.	#SPADs	FF	PDP	Median DCR	System Timing Jitter	Gate Window	Gating Frequency/ Framerate
Unit	**nm**	**-**	**%**	**%**	**Hz (@RT)**	**ps (FWHM)**	**ns**	**MHz/fps**
2014 [153]	350	8192 ([64x8]x[16x1])	44.3	21.7 @ 465nm Vex=3V	5.7k Vex=3V	-	0.7 FWHM	950 Hz/-
2015 [154]	350	1024 (2x4x128)	23	~25 @ 532nm Vex=3.3V	19.2k Vex=3.3V	-	1.1-4 (4 gates)	0.1/-
2015 [155]	350	60 (60x1)	52	50 @ 420nm Vex=5V	2.5k Vex=5V	~470	300	-/1.7M
2016 [156]	350	19200 (160x120)	21	-	580 Vex=3V	-	0.75	50/486
2018 [157]	350	4096 (16x256)	12	-	31.25k Vex=?	225	-	-/400k
2018 [158]	180	1024 (32x32)	9.6	-	-	-	10	-/360k
2019 [159]	350	64 (8x8)	12.5	55 @ 450nm Vex=6V	60 Vex=5V	>410	350	-/1M
2019 [160]	350	2000 ([10x5]x[8x5])	32	4 @ 810nm Vex=3.3V	200k Vex=3.3V	~200	<1	-/1.2M
2019 [126]	180	262144 (512x512)	10.5	50 @ 520nm Vex=7V	7.5 Vex=6.5	97.2-139.5 (SPAD)	5.75-100	2.5/97.7k
2020 [135]	350	1728 ([12x36]x[2x2])	37	~54 @ 400nm Vex=3.3V	424 Vex=3.3V	300	4-8	100/- (Single channel)

The time gating of SPADs was commonly used as a method of reducing the DCR and afterpulsing in applications where the photons' arrivals are correlated with a reference clock. As the approximate arrival time of photon absorption is known, the SPAD can be enabled only during a specified window where the anticipated photon events are most likely to occur. During the time in which the SPAD is disabled, it is biased below its breakdown voltage, and impinging photons cannot initiate an avalanche. Also, excess free carriers captured by traps during the previous avalanche have the ability to recombine without initiating afterpulses. While time gating is primarily adopted to reduce the effects of noise, it has also shown the capabilities of obtaining timing information of the photon events by shifting the gate window with respect to a synchronous laser pulse to create histograms of the SPAD counts.

In [153], an array containing 8192 SPADs utilized a combination of on-chip and off-chip delay lines to shift a single gate window of the SPAD array in 250 ps increments over a range of 32 ns. By measuring the SPAD outputs for a large number of cycles in each gate delay setting, a histogram with 250 ps time bins was constructed by subtracting SPAD counts from adjacent gate delay settings. More recent works using a similar approach have demonstrated gate shifting capabilities as fine as 18 ps when generated from an FPGA clock generation block [126]. The main disadvantage of these approaches is that histograms need to be constructed sequentially in each gate delay setting, reducing the frame rate. In [154], four time gates of different widths, but starting at the same instant in time, were applied to a SPAD array using an off-chip delay generator. This approach allows for simultaneous acquisition of the data for different gate windows, which could reduce the time needed to produce a valid histogram. Acquisition speed is a key consideration in DOT, where fast images are required to avoid motion artifacts or tissue variations during the measurement.

A conceptual diagram of the proposed 1D multi-time-gated SPAD array is depicted in Figure 4-2, and the timing diagram is shown in Figure 4-3. The input clock to the circuit is synchronous with a pulsed laser. Delayed rising edges of the clock are tapped from a multi-purpose delay line, which can be used to provide shifted gate windows that are overlapped by approximately half the gate width to adjacent SPADs. Therefore, each SPAD is responsible for constructing a small section of the histogram, which can later be recombined to give the final result. Since the delay line can be integrated using a chain of buffers, which is an identical structure to a delay line TDC, the multi-purpose delay line's state at the time of a SPAD output can be stored in an array of DFFs to quantize the photon arrival time relative to the start of the gate window. The time to construct a full histogram can then be potentially reduced by a factor equal to the number of simultaneous gates applied to the SPAD array, compared to shifting a single gate window across the array. Additionally, as the circuitry for generating the delayed gates is shared, it can be fully integrated with the SPAD array while minimizing the impact on the fill factor. As a proof of concept, both a 1D design containing a 6×1 SPAD array and a 2D design with a 4×4 SPAD array were designed in the TSMC 65 nm process utilizing this structure. The 2D

design, in particular, could serve as a repeatable structure to create a larger array due to the scalability of this structure as a result of using shared circuitry.

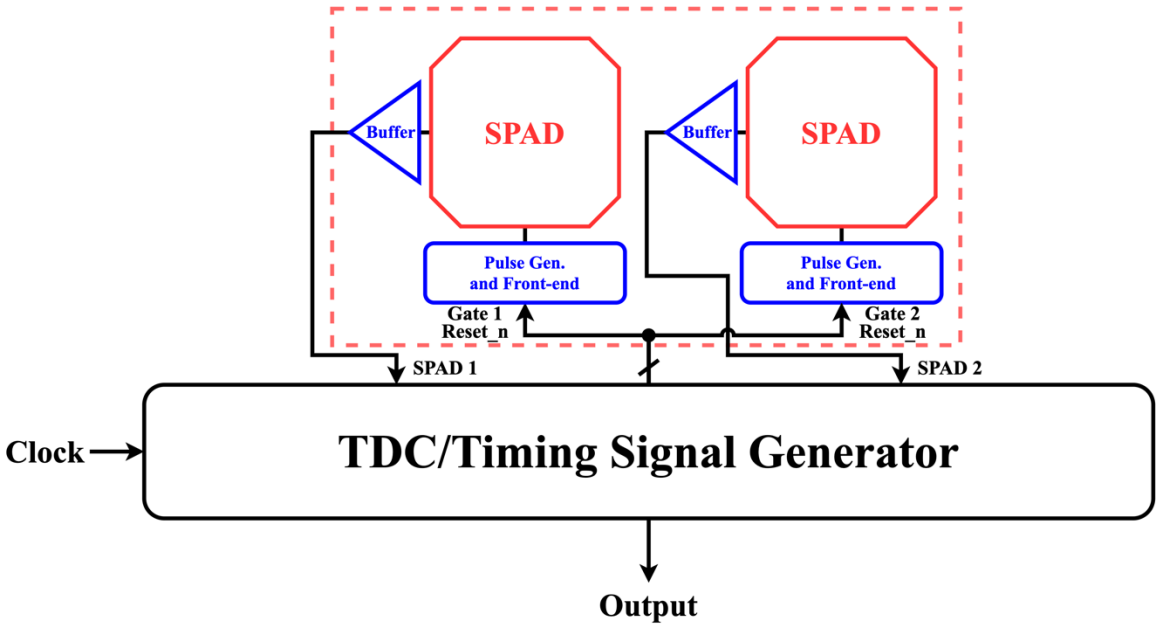

Figure 4-2: Conceptual block diagram of the proposed 1D multi-time-gated SPAD array.

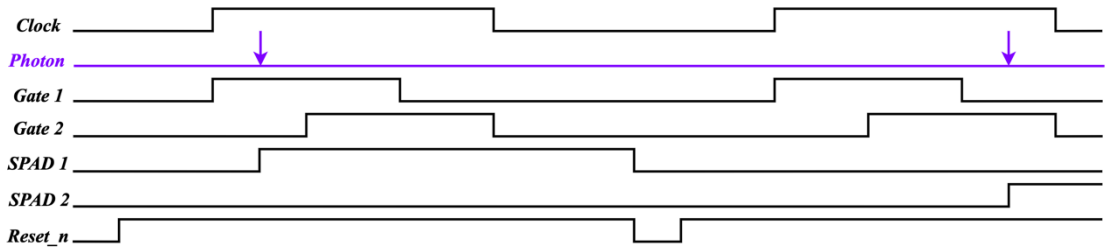

Figure 4-3: Conceptual timing diagram of the proposed 1D multi-time-gated SPAD array.

4.2. SPAD Pixel Design

4.2.1. SPAD Structure

For this design, a p+/n-well SPAD was used in both the 1D and 2D multi-time-gated arrays. The top view in Cadence Virtuoso and an illustration of the cross-sectional view are shown in Figure 4-4. The SPAD was designed conservatively to ensure correct functionality in a single design iteration; and consists of a ~10 µm diameter active area and

achieves a fill-factor of ~24.5%, which could be greatly improved with an optimized design. A silicide blocking layer was placed over the active area, and an upper metal layer was placed over the SPAD, with an opening only above the active area in order to ensure impinging photons are more likely to be absorbed in the active area of the SPAD.

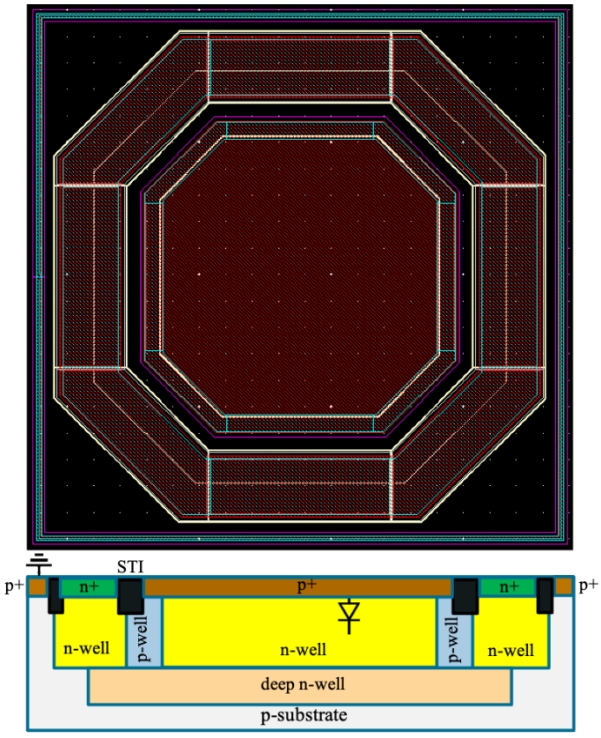

Figure 4-4: Top view of the SPAD layout in Cadence Virtuoso, and the cross-sectional view. The p+ region extends into the lesser doped p-well and pushes the STI away from the high field region. This mitigates premature edge breakdown of the junction.

An octagonal shape was used for the SPAD active area as obtuse corners within the SPAD structure were demonstrated to reduce the PEB compared with using sharp 90-degree corners [9]. An additional method of reducing the PEB is by using a p-well guard ring[161]. As the p+ region extends beyond the n-well and into the p-well, the lower doping of the p-well reduces the electric field strength near the edges. This ensures the strongest field is within the planar junction, which is verified by the electric field distribution TCAD simulation shown in Figure 4-5. The p+ region extending beyond the n-well had the added benefit of pushing the STI away from the active area, mitigating the effect of STI-generated carriers initiating avalanches and contributing to the dark noise of the SPAD [162].

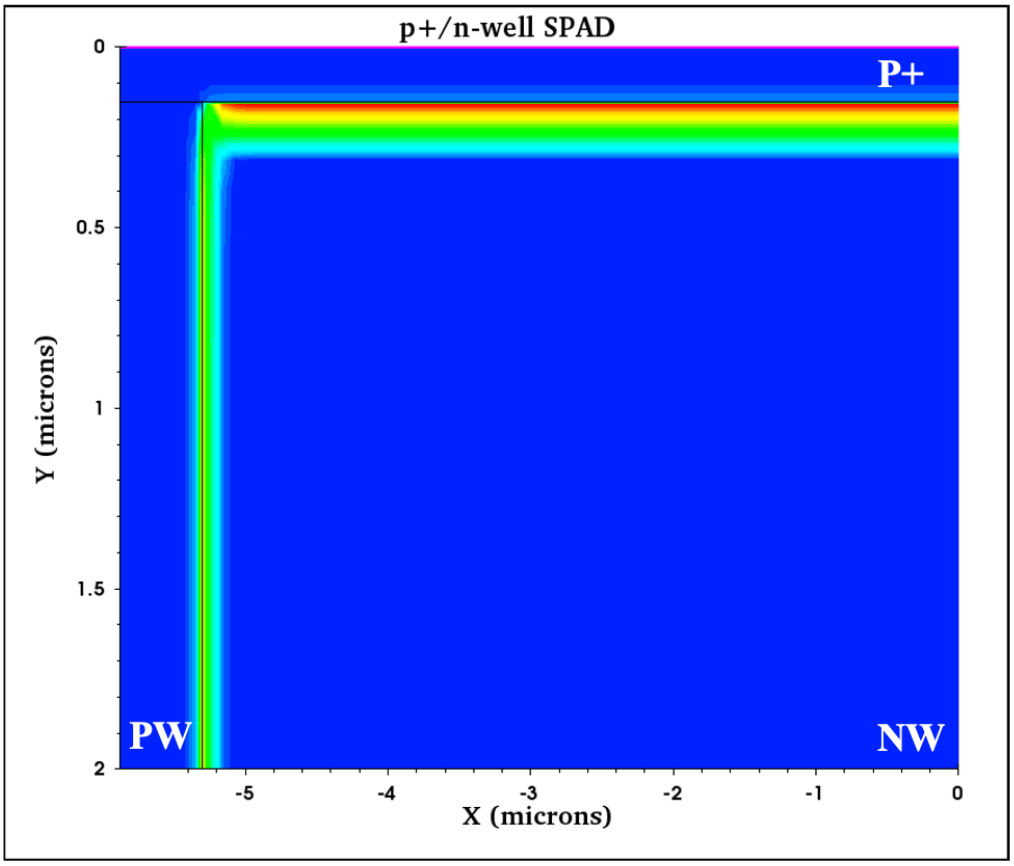

Figure 4-5: TCAD simulation of the proposed SPAD verifying the highest field being within the planar junction.

4.2.2. Front-End Circuitry

In order to obtain the gating operation, the SPAD front-end circuitry was designed as shown in Figure 4-6 with the active-low P1 and P3 pulses and the active-high P2 pulse being generated by the circuit shown in Figure 4-7. This front-end circuit is based on previous designs from our group as in [162]. A negative high voltage is applied to the anode of the SPAD, and in the initial state, P2 is high to ensure the SPAD cathode is discharged to ground, such that the SPAD is biased below its breakdown voltage. A short P1 pulse (i.e., a few hundred picoseconds) then turns on M1 at the same time that P2 turns off M2, which charges the SPAD cathode to V_{SPAD}. The SPAD bias is then above its breakdown voltage. After the end of the P1 pulse, the P3 pulse transitions and enables the output branch of the SPAD front-end. Thus, when a photon allows conduction in the SPAD, the gate of

M4 will be discharged below its switching point, and M4 and M5 will form a low-resistance path to V_{DD} that generates the output pulse.

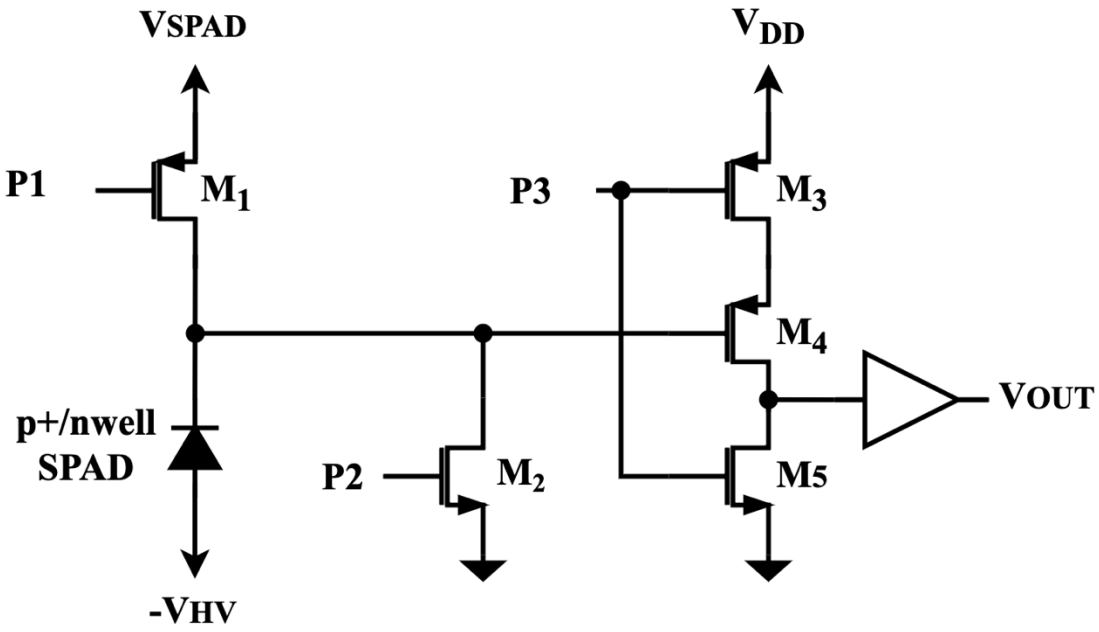

Figure 4-6: Schematic of the SPAD front-end circuit. P1, P2, and P3 are timing signals generated by the pulse generation circuitry within the SPAD pixel that produces the gate window.

In our previous designs, the P1, P2, and P3 pulses were generated using buffers with large capacitive loads in order to create the required delays [162]. In this design we use a tapped delay line to create the delays, such that taps from a single delay line can be used to: generate the gating pulses P1, P2, and P3 for each SPAD pixel in the array; perform shifting of the gate windows for adjacent pixels; and coarse time-to-digital conversion. The delay line is shared by each pixel of the design and will be discussed in detail in the next subsection, while the pulse generation logic is integrated directly with each pixel. Conceptually, the rising edge of the clock signal can be used to denote the start of the P1 and P2 pulses, and the next tap of the delay line can quickly disable the P1 pulse such that the pre-charge of the SPAD occurs very quickly. After a number of buffers required to achieve the desired gate width, another tap of the delay line can then be passed to the in-pixel pulse generation circuit in order to end the gate window.

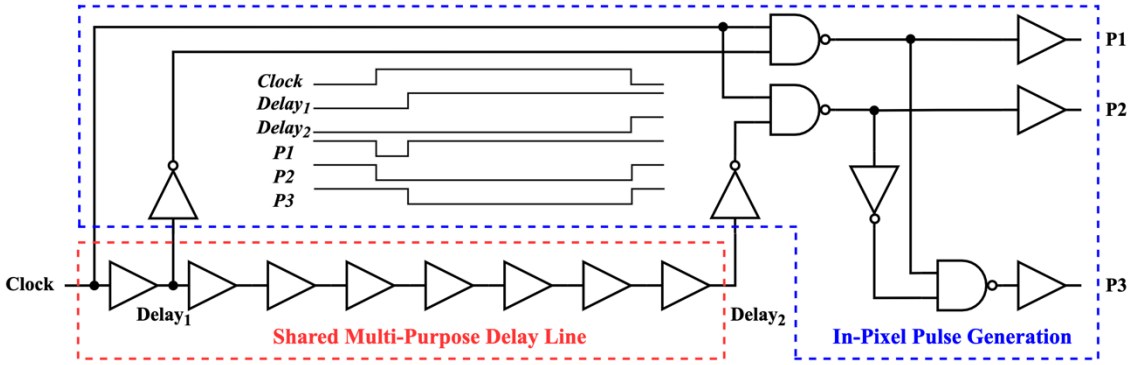

Figure 4-7: Schematic of the SPAD pulse generation circuit. Delayed replicas of the clock are tapped from a shared multi-purpose delay line located outside the SPAD pixel. The in-pixel pulse generator uses combinational logic to generate the gating signals (i.e., P1, P2, and P3).

A time-gated pixel consisting of a p+/n-well SPAD integrated with the time-gated front-end and pulse generation logic circuitry is shown in Figure 4-8. As mentioned previously, the SPAD itself was designed conservatively to ensure correct functionality (i.e., low dark noise with no edge breakdown) in a single design iteration. In an optimized design, the spacing between the SPAD and its front-end/pulse generation circuitry can be reduced for a more compact pixel. In the current design, the SPAD pixel achieves a fill factor of ~18%, compared to the ~24.5% fill factor of the SPAD itself. This fill factor can be improved primarily by optimization of the SPAD, as there is more area to be saved in the SPAD structure than in the optimization of the size of the front-end circuits.

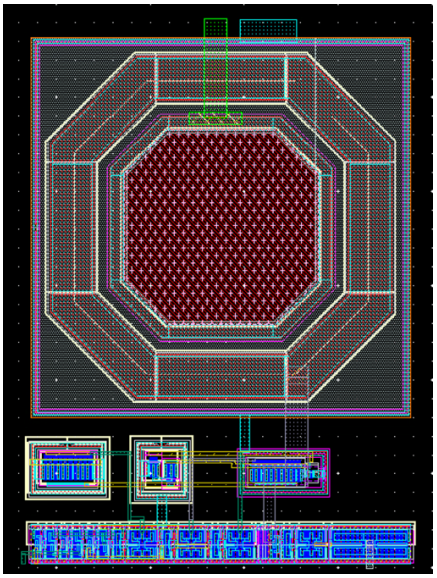

Figure 4-8: Layout of the SPAD with front-end and pulse generation circuits.

4.3. Multi-Purpose Delay Line

As delayed replicas of the input clock are required to: generate the P1, P2, and P3 pulses; shift the gate windows of adjacent SPADs; and perform coarse timestamping of photon arrivals; we implemented a single multi-purpose delay line to simultaneously accomplish all these functions. By sharing the circuitry in this manner, the impact on the fill factor can be minimized. The general structure of the multi-purpose delay line is shown in Figure 4-9. An external clock signal is applied as the input for the first buffer of the delay line, which initiates the gating pulse generation for the first pixel. At the same time, the clock's rising edge propagates through the series of buffers towards the delay taps associated with the next SPAD pixels. In order to overlap the gate windows of SPADs in adjacent columns by half, the delay taps used to generate the P1, P2, and P3 of the next pixel will start from a delay tap occurring halfway through the previous gate window. When a SPAD generates an output pulse, it will clock the DFFs whose data inputs are connected to the delay line taps within its respective gate window.

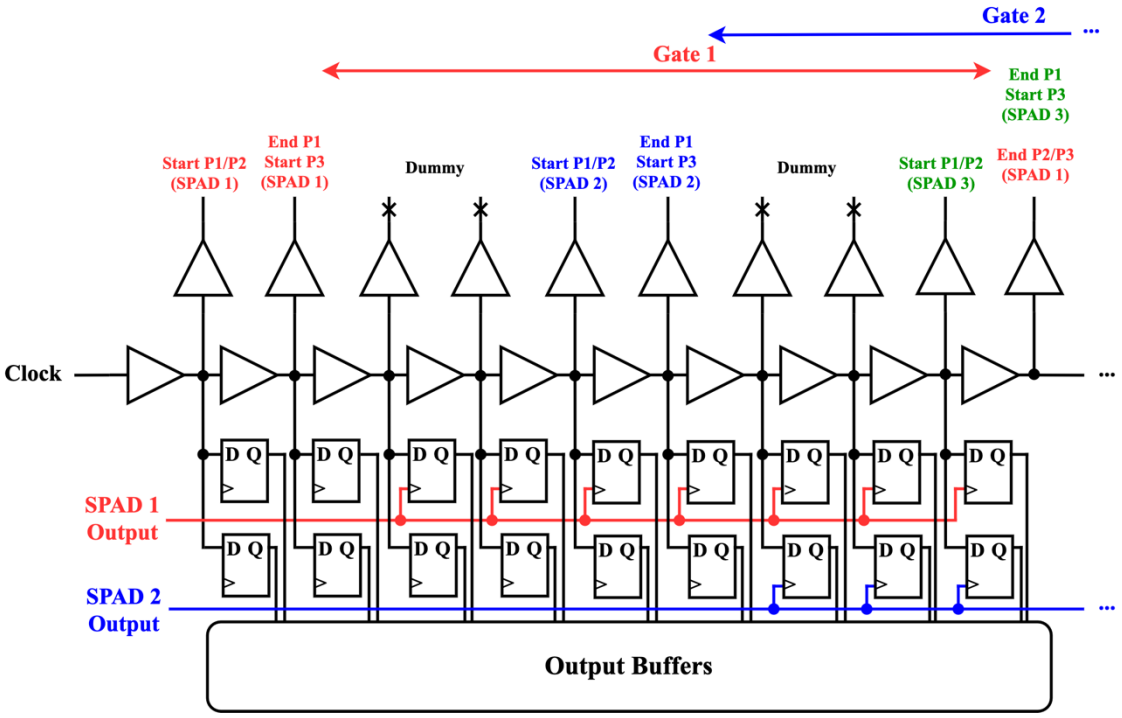

Figure 4-9: Schematic of the shared multi-purpose delay line that drives the multi-time-gating and TDC operation.

A primary consideration when designing the delay line is to maintain equal capacitive loads for each delay buffer in order to reduce delay mismatches that contribute to nonlinearity in the timing performance. For this reason, each tap of the delay line is not directly connected to the pulse generation logic of the corresponding SPAD. Instead, the tapped delay is first passed through an intermediate buffer, such that the load on the actual delay cells is consistent regardless of the connections to the SPAD front-end. A similar approach is used on the DFFs that sample the delay line state for coarse time-to-digital conversion. As the gate windows overlap by half, this means that most delay line taps will need two DFFs clocked by two different SPADs. Dummy DFFs are used to ensure the same capacitive load at each stage, even when they are not clocked by a SPAD. This case mainly happens at the start and end of the multi-time-gated array, where only a single SPAD will cover that portion of the timing range. A robust implementation in the future should also aim to use a DLL to lock the delay line to a reference in order to ensure consistent performance across PVT variations and to minimize the jitter of the delay line.

The final output of the circuit will consist of 4 bits for each SPAD that are buffered to the output through DFFs. The first bit is a status signal that indicates if a SPAD pulse occurred during the given time-gated period. This signal is generated from a DFF with the data input tied high and the associated SPAD serving as the clock input. The remaining three bits for the output of a given SPAD come from the 3-bit output code of the coarse TDC. The DFFs are then reset after the falling edge of the clock when the measurement interval is complete. Since in DOT experiments, the event rate is kept low to avoid pile-up, a single fine-interpolating TDC could be shared in future by a set of SPADs while missing a minimal number of events. This second-stage TDC would increase the resolution by quantizing the remainder of the coarse time-to-digital conversion to generate additional fine time resolution bits.

In the layout of the multi-purpose delay line, matching between the traces in the array was a key consideration. The delay line was designed such that the spacing of the buffers was evenly distributed along the bottom of the array of SPADs. Variations in routing are also a factor that contributes to the timing nonlinearity, as it impacts the capacitive load

seen by each stage. By designing the spacing of buffers to match the pitch of SPAD pixels within the array, the trace lengths can be more closely matched to avoid larger variations in the capacitance that degrade the timing performance. The multi-purpose delay line circuit occupies an area of 0.0017 mm^2 for the 1D design, which can be easily scaled to match the length of a SPAD array of a different size, as in the implemented 2D array where the multi-purpose delay occupies a reduced area of 0.001 mm^2.

4.4. A 2D Multi-Time-Gated SPAD Array

The proposed 1D SPAD array can be easily extended to a 2D array. Here, the design was extended to a 4x4 2D multi-time-gated SPAD array within the same TSMC 65 nm process. A simplified illustration of a 2x2 array is depicted in Figure 4-10 for ease of understanding. In the 2D case, SPADs within each column will operate on the same gate window, and their outputs are ORed to give the output of that column. SPADs within separate columns will operate on shifted gate windows in the same manner as the 1D array.

As previously mentioned, since the spacing of buffers within the multi-purpose delay line was designed to align with the pixel pitch of the SPAD array, the delay line was easily modified to fit the shortened length of the 4x4 array. In the 2D design, the additional feature was added that each SPAD pixel could be disabled by externally configurable mask bits. The mask bits control the select input of MUXs that are placed within each SPAD pixel. The MUXs will either connect the taps from the multi-purpose delay line that initiate the gating pulses to the SPAD front-end, or tie the P1, P2, and P3 inputs of the SPAD front-end circuit to GND such that the SPAD is constantly disabled. This allows for the disabling of “hot-pixels” that have a much higher DCR than the majority of the pixels within the array. For the purposes of testing, it allows us to easily assess the performance of individual SPADs.

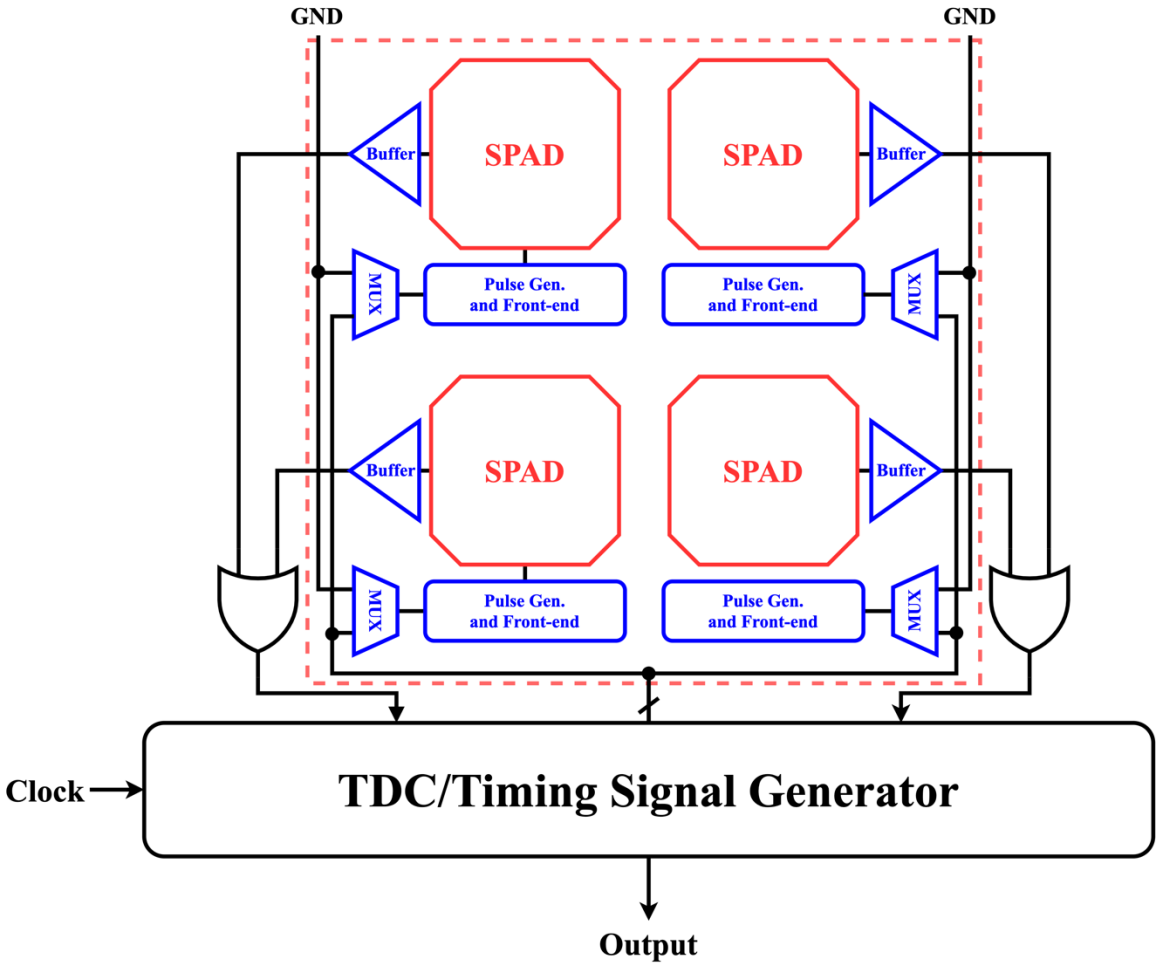

Figure 4-10: Conceptual block diagram of the proposed 2D multi-time-gated SPAD array.

4.5. Layout and Simulation Results

The final layouts of the 1D and 2D multi-time-gated SPAD arrays are shown in Figure 4-11 and Figure 4-12, respectively. For the 1D case, the total area of the design, including the output buffers and readout electronics is ~0.008 mm^2, for a total fill factor of ~7.7%. Note that in the 1D case, the output buffers and readout electronics largely contribute to the inactive area. The 2D array fill-factor is improved, as the 4 SPADs within a column are ORed together, and therefore can share the same channel of the multi-purpose delay line and readout circuitry. The limitation here is that the counts from different SPADs within the same column are indistinguishable from each other due to the spatial compression when

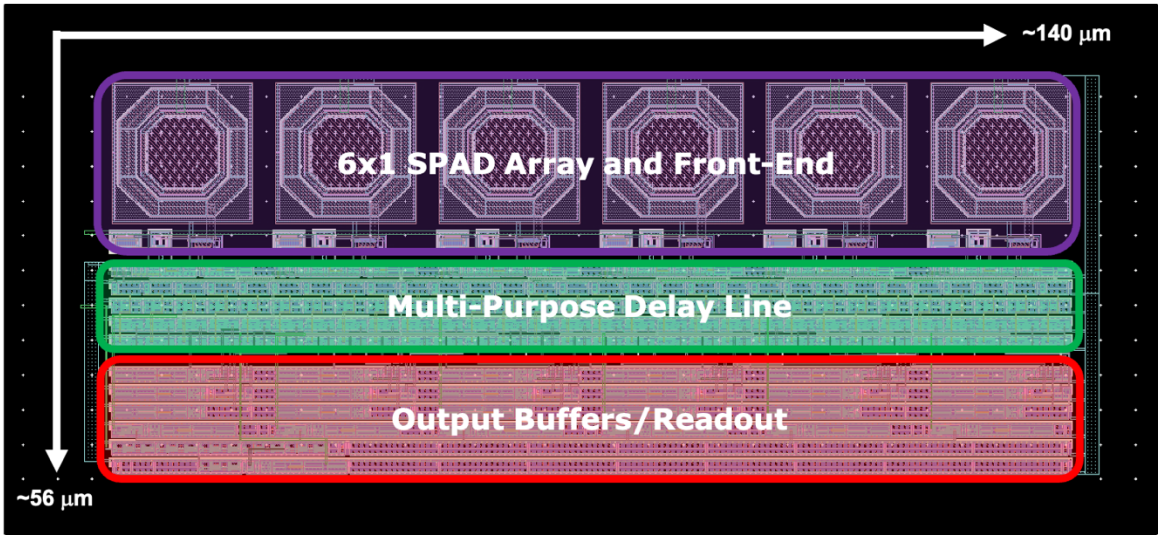

Figure 4-11: Layout of the 1D multi-time-gated SPAD array.

Figure 4-12: Layout of the 2D multi-time-gated SPAD array.

ORing their outputs. The 2D design occupies a total area of 0.017 mm^2 for a fill factor of ~9.6%. The SPAD pixels themselves largely limit the fill factor due to the conservative design choices made to ensure correct functionality on the first fabrication attempt. For an

optimized design, the fill-factor of each SPAD pixel should be optimized, and approaches such as well sharing or guard ring sharing could be used to minimize the pitch between adjacent SPADs [163].

Based on the layouts shown previously, we conducted a post-layout simulation to verify the functionality of both multi-time-gated designs. Here, the post-layout simulation of the 2D design is presented. As shown in Figure 4-13, a 20 MHz clock was applied at the input of the array. A simulation model for the SPAD was used according to [164], and the photon arrival time was increased with respect to the clock by 1.5 ns in each cycle in order to validate the functionality across the sensor's dynamic range.

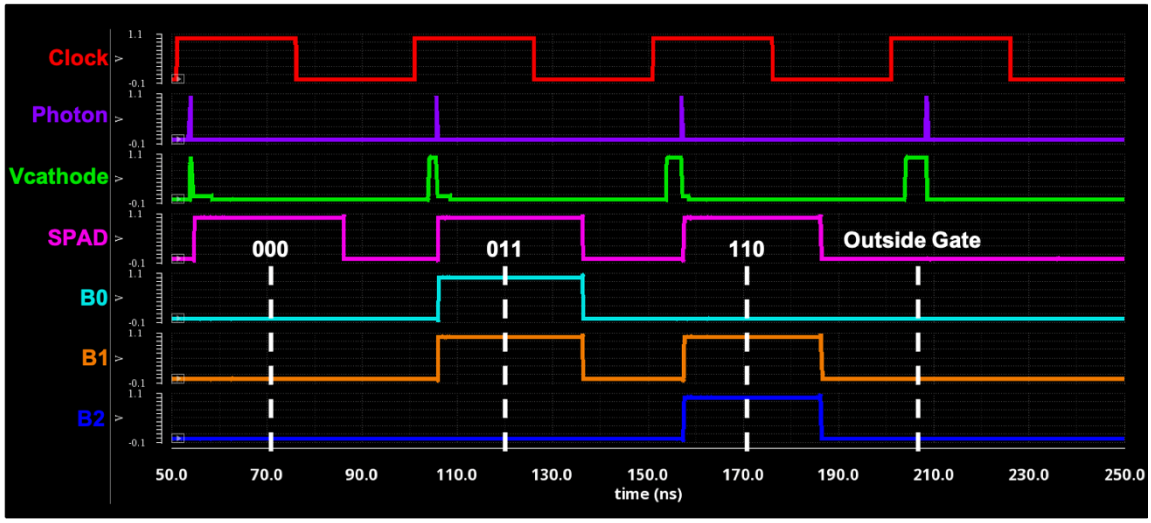

Figure 4-13: Post layout simulation verifying the correct operation of the 2D multi-time-gated SPAD array.

It can be seen that after a delay from the rising edge of the clock, the SPAD pixel's cathode voltage is pre-charged by its front-end circuitry to arm the SPAD for photodetection. In the first clock cycle, the photon arrives at the very start of the gate window, causing the cathode voltage to discharge and the SPAD to generate an output pulse. The resulting TDC code (i.e., read as $B_2B_1B_0$) is 000, as the photon arrives at the very start of the gate window opening, and the clock's rising edge has propagated a minimal distance through the delay line. As the photon arrives 1.5 ns later in each cycle, it can be seen that the TDC code gradually increases to 011 and 110 before the photon arrives outside the gate window in the fourth cycle and fails to generate an output pulse as expected. Note

that the displayed output waveforms are buffered through DFFs that are reset after a delay of the falling edge of the clock. This design choice was made to ease in the testing of the physical chip, as the falling edge of the clock could be used as a trigger for sampling the result during a measurement cycle.

4.6. Conclusions

In this section, the operation and design of multi-time-gated SPAD arrays were discussed. Common downfalls of current time-gated designs are that a single gate window is often shifted over an entire SPAD array in order to produce a histogram. As the gate window must be shifted and have a large number of measurement cycles in each setting, this significantly reduces the maximum achievable imaging rate. The multi-time-gated method is capable of allowing groups of SPADs to operate on separate gate windows, with each SPAD being responsible for covering only a portion of the total histogram. This could allow different sections of the histogram to be recorded simultaneously, improving the imaging rate of the detector.

As a proof of concept, both a 6 × 1 1D array and a 4 × 4 2D multi-time-gated SPAD array were designed in the TSMC 65 nm process with the shifted gate windows being generated on-chip by circuitry closely integrated with the SPAD array. While this structure would generally add more circuitry to the design, which reduces the fill factor, the proposed design shares a single delay line to create the timing pulses for the time-gated front-end circuits of the SPADs, the shifting of the gate windows, and coarse time-to-digital conversion. Therefore, the proposed structure adds minimal circuitry per SPAD compared to a standard time-gated pixel of the same form.

Lastly, while it was not implemented in this work, it should be noted that due to the shared circuitry that implements the multi-time-gated operation and coarse time-to-digital conversion, the 2D array, in particular, is highly scalable. Any example of a potential implementation for a larger SPAD array is shown in Figure 4-14. Several small arrays of the proposed design could be tiled and share a single fine interpolating TDC to improve the

timing resolution. Since in DOT, the event rate is kept low to avoid pile-up distortion, minimal events could be missed by the TDC. This structure could therefore help to mitigate the degradation of the fill factor in order to maximize the PDE of the total array while achieving high timing performance and fast operation through the multi-time-gated operation.

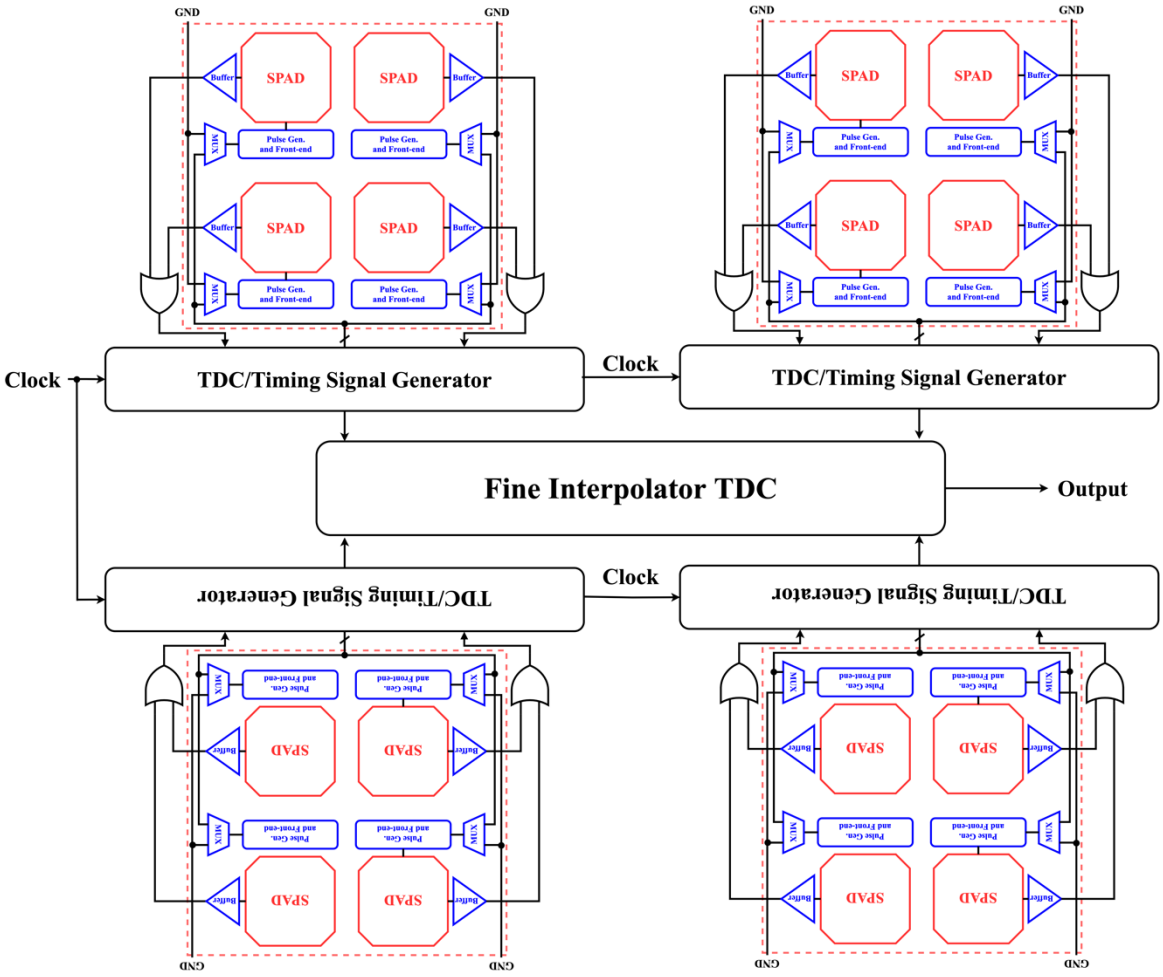

Figure 4-14: A scalable multi-time-gated architecture that shares a fine interpolating TDC between several smaller arrays.

Chapter 5

Measurement Results of CMOS SPADs

5.1. Fabricated Test Chip and Printed Circuit Board

In this chapter, the measurement results are presented for SPAD designs in the TSMC 65 nm CMOS process. The top-level layout of the fabricated chip is shown in Figure 5-1 and consists of several test structures. Here, we perform a detailed performance characterization of the p+/n-well SPAD in a passive quench configuration. The p+/n-well SPAD structure was additionally used on the same chip to form the initial prototype of a 1D multi-time-gated SPAD array. Due to shortcomings with the current design, several time-gated pixels were generating a SPAD output in every gate window. As such, the full 1D array could not be characterized. The potential explanations for this error will be discussed later, based on the performance results of individual pixels within the array that were functional. The total chip area was 1.5 mm x 1.5 mm and was bonded to a 68 pin PGA package.

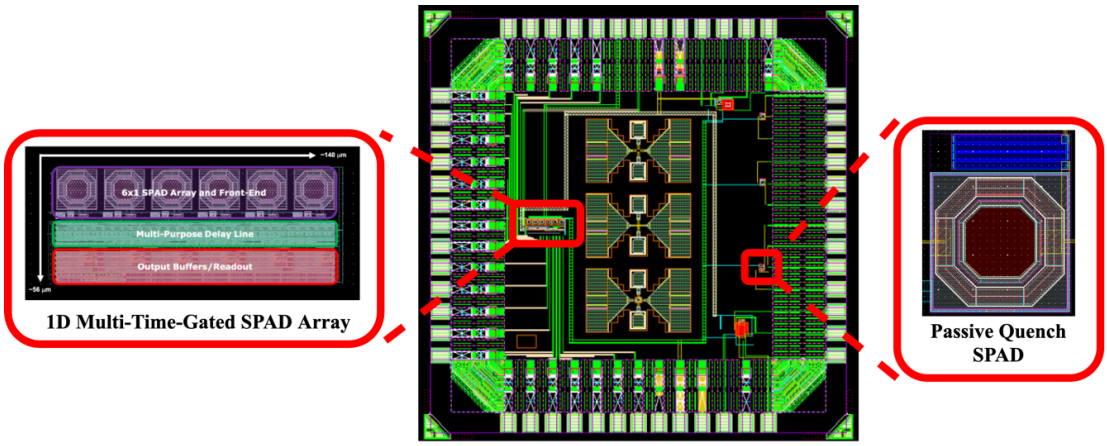

Figure 5-1: Chip-level layout that was fabricated in the TSMC 65 nm CMOS process.

For all experiments pertaining to the passive quench p+/n-well SPAD and the 1D multi-time-gated SPAD array, the printed circuit board (PCB) shown in Figure 5-2 was

used. The test PCB consisted of DC barrel jacks for connecting to a 3.3 V supply voltage, as well as for the DC biasing of the SPADs. The 3.3 V supply was used as the IO voltage for the test chip, and the 1 V core voltage was generated from a low-dropout regulator. Aside from the DC voltages, edge-mounted SMA connectors were used for sending signals in and out of the board. The input clock to the chip comes from an SMA connector on the left-hand side of the board, terminated by a 50 Ω resistor to minimize reflections from an impedance mismatch. Also, since there are several other test structures on this chip, jumpers were used to select which test structures to route each connector, to reduce the number of connectors required on the PCB.

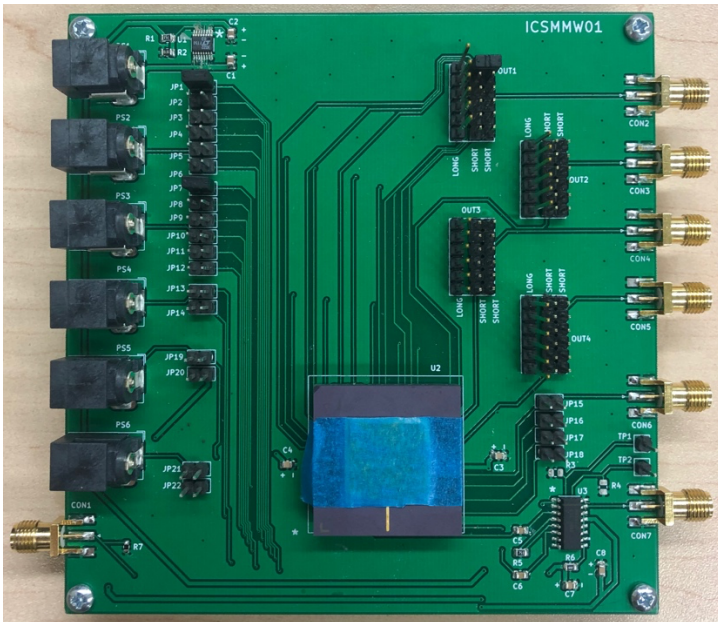

Figure 5-2: Photograph of the test PCB used to assess the functionality and performance of the design.

5.2. Passive-Quench SPADs

The first results that will be shown are from the p+/n-well SPAD in a passive quench configuration, as shown in Figure 5-3. In this circuit, the quench resistor is placed on the anode of the SPAD, and the output is an active-high pulse whose amplitude is equal to the excess bias above the breakdown voltage. While it leads to a longer dead-time of the SPAD, a large 50 kΩ resistor was used to ensure that the current through the SPAD during an

avalanche is reduced enough to ensure proper quenching. While this is acceptable for determining the general performance characteristics of the SPAD, it should be noted that a smaller quench resistor may be a more optimal choice for real applications, or an active quench configuration could be used.

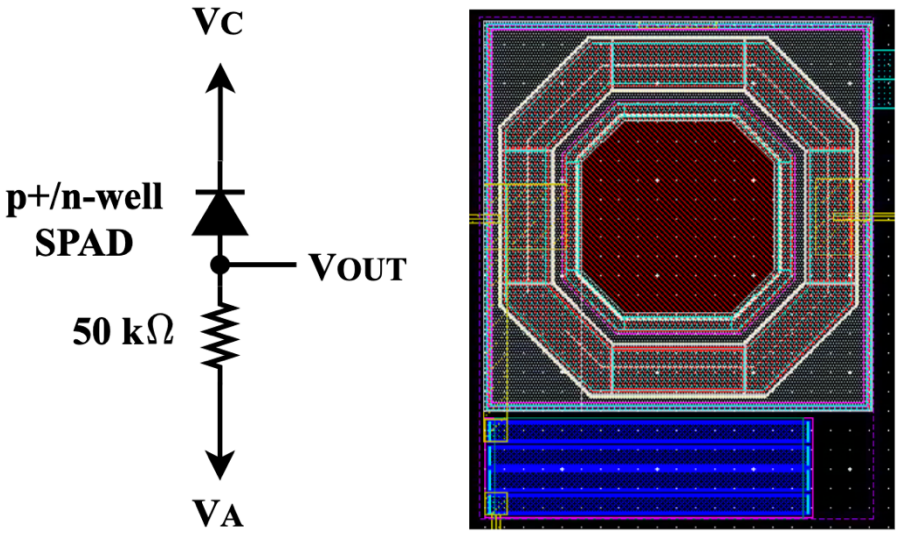

Figure 5-3: Schematic and layout of the passively quenched p+/n-well SPAD in the TSMC 65 nm process.

5.2.1. Breakdown Voltage

The first and most fundamental step to evaluating the performance of a SPAD is to accurately determine the breakdown voltage. As aspects of the SPAD performance such as dark noise and afterpulsing are often measured in different temperatures, the temperature dependence of the SPAD breakdown voltage should be measured. In this way, a constant excess bias can be maintained during the temperature-dependent measurements. Due to the increased phonon scattering at higher temperatures, electrons and holes are less likely to have the required energy for impact ionization [165]. Therefore, it is generally expected that the breakdown voltage should increase with temperature and have an approximately linear relationship within our measurement range. As such, the SPAD bias can be easily compensated through a linear decrease or increase in the bias voltage when measuring at low or high temperatures, respectively.

To measure the temperature dependence of the breakdown voltage, the test PCB was placed inside an Espec thermal chamber and connected to an external Agilent B1500A semiconductor device analyzer, as shown in Figure 5-4. The semiconductor device analyzer was responsible for applying a bias voltage to the SPAD which increased in small steps, and measuring the DC current. The IV relationship was saved to a spreadsheet, and the breakdown voltage was taken as the point where the current increased by at least 10 times between two adjacent steps. For the complete measurement, 4 SPADs were used and measured at temperatures ranging from -30 °C to 30 °C in 15 °C increments. For each point on the graph, 10 measurements of the IV characteristics were averaged to avoid the impact of any variation during the measurement.

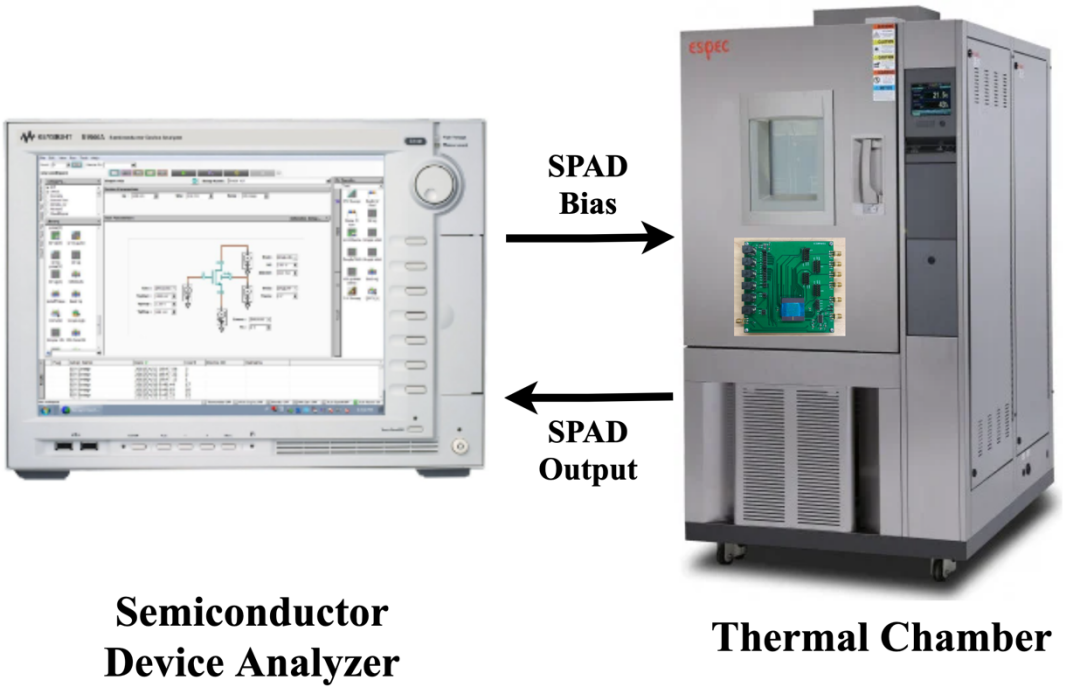

Figure 5-4: Diagram of the experimental setup for the breakdown voltage measurements.

Room temperature within our lab setting was ~25 °C, at which the average breakdown voltage for these 4 SPADs was measured to be 9.88 V. The results of the breakdown voltage measurement against temperature are presented in Figure 5-5. By performing a best linear fit to the data points for each of the 4 SPADs, the average slope was taken as the temperature coefficient of 4.9 mV/°C. The magnitude of this breakdown voltage is only

slightly higher, and the temperature coefficient is very consistent with a previous n+/p-well SPAD we have designed in the same process [162]. The reason for the slightly higher breakdown voltage could be from 2 likely sources. Firstly, the p+/n-well junction may be purposefully less doped than the n+/p-well equivalent in this process; or secondly, the lower doping on this set of chips could be attributed to process variations.

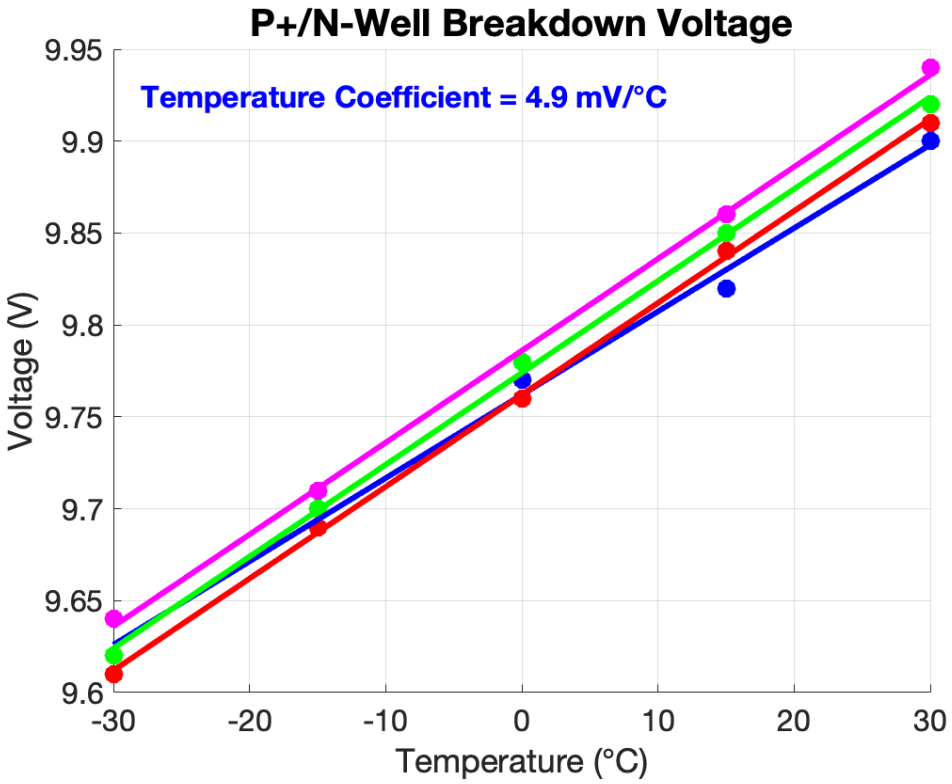

Figure 5-5: Results of the breakdown voltage measurement at different temperatures for 4 SPADs.

Due to the slightly higher breakdown voltage, it is expected that the p+/n-well junction has a wider depletion width in this TSMC 65 nm process. As such, it is expected that the PDP of the SPAD should be larger due to the increased size of the depletion region in which photons can be absorbed. Additionally, as the junction would appear to be slightly less doped, the contribution to the dark noise from band-to-band tunneling may be slightly reduced and will be assessed with an Arrhenius plot in the next subsection.

5.2.2. Dark Count Rate

Although the PDP and timing jitter performance of a SPAD are improved when we increase the excess bias, the DCR will also increase exponentially [27]. As such, the excess voltage dependence of the DCR for the p+/n-well passive quench SPAD was measured using the setup shown in Figure 5-6. The test PCB was placed within an Espec thermal chamber, where the bias was supplied by an Agilent E3646A DC power supply, and the SPAD output was connected to a Lecroy Waverunner 625Zi mixed-signal oscilloscope. The SPAD excess bias voltage was increased from 0.3 V to 0.7 V in 0.1 V increments. For each setting, ~10^5 samples were taken such that the tail of the exponential interarrival time (IAT) histogram was stable. Although this SPAD was anticipated to be operated at room temperature for the remaining measurements, the thermal chamber was used for this experiment such that the excess voltage dependence was measured for temperatures between -30 °C to 30 °C in 15 °C increments. In doing so, the activation energy for the SPADs was later extracted to identify the dominant mechanism contributing to the dark noise of the SPAD.

Figure 5-6: Diagram of the experimental setup for the excess voltage and temperature-dependent dark count rate measurements.

By plotting the DCR results on a semi-log scale as shown in Figure 5-7, the DCR was confirmed to follow an exponential relationship with excess voltage. In the worst case, with the highest temperature of 30 °C and the highest excess voltage of 0.7 V, the SPAD exhibited a dark count rate of 44.92 kHz. This is comparable with results published in recent literature for the same CMOS process. In [166], the same p+-n-well junction exhibited a DCR of almost 1 MHz with an excess voltage of 1.5 V. However, the focus of that design was on obtaining low timing jitter with optimized front-end circuitry. Additionally, the DCR performance of this SPAD is greatly improved from our previous n+/p-well design that obtained a DCR of ~13.8 MHz at a 0.3 V excess bias [162]. Previously, to create a triple junction SPAD for wavelength distinction, the top junction was formed from an n+/p-well SPAD with an STI guard ring. While the STI can effectively act as a guard ring for preventing premature edge breakdown, it introduced defects near the depletion region of the SPAD, which acted as generation-recombination centers for charge carriers, and greatly increased the DCR. The disadvantage of the p-well guard ring in this design is that it occupies greater space within the SPAD pixel that reduces the fill factor. However, when designing SPADs in standard processes that are not optimized for low noise photodetection, it is a necessary trade-off to ensure a functional device.

Another important result that was extracted from the DCR measurement is the afterpulsing behaviour of the SPAD. During an avalanche generated from either the detection of a photon or from dark noise, the carriers moving through the SPAD's depletion region may fill traps resulting from crystal impurities or defects. These traps have finite lifetimes and may probabilistically release carriers after a period related to the trap lifetime. If the SPAD is biased above breakdown when the carrier is released, then a correlated pulse known as an afterpulse may occur. Afterpulsing is an undesirable phenomenon as these counts are either from dark noise or from photons that were already detected, and will distort single-photon timing measurements. Additionally, if the afterpulsing exhibited is large, it will effectively lower the count rate of the SPAD to real photon events.

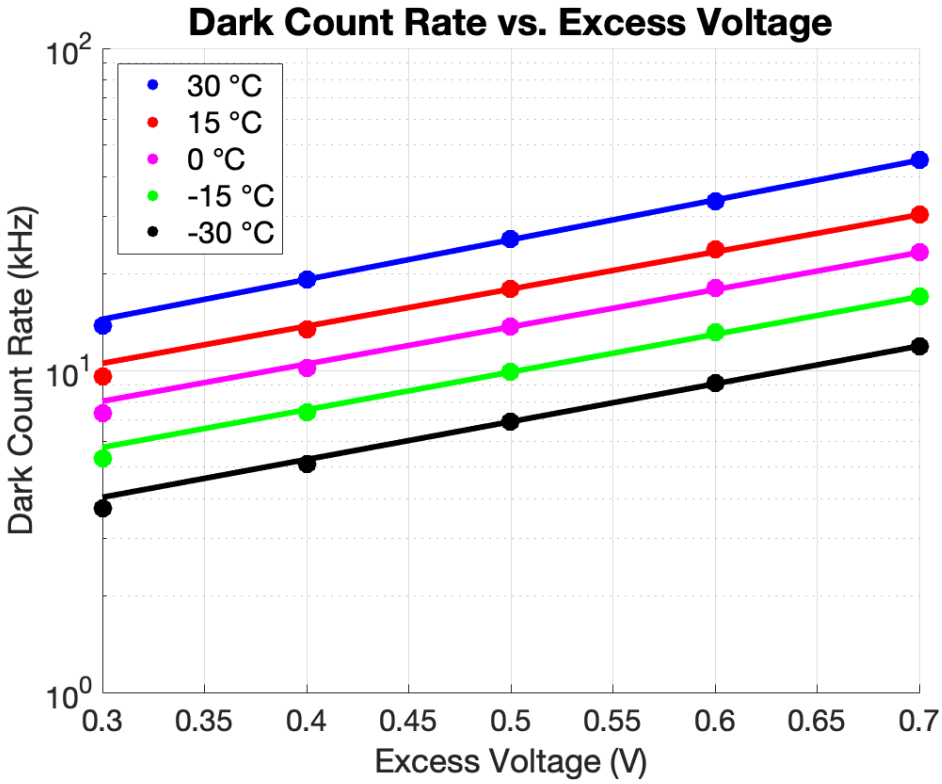

Figure 5-7: Dark count rate vs. excess voltage for different temperatures.

The first order IATs of primary dark counts resulting from band-to-band and trap assisted tunneling and thermal generation follow a Poisson process as follows [167]:

$$f(\tau) = \lambda \exp(-\lambda\tau), \tag{5-1}$$

where λ is the mean IAT, being the inverse of the primary DCR. As such, by applying an exponential fit to the IAT distribution, afterpulses are observed as an increase in the total DCR at low interarrival times since afterpulses occur very shortly after the original pulse due to the short lifetimes of the traps. Therefore, the afterpulsing can be determined by integrating the difference between the exponential fit and the measured IAT histogram. Figure 5-8 shows the IAT distributions for a SPAD with a 0.3 V excess voltage at -30 °C, 0 °C, and 30 °C. In these measurements, afterpulsing was only observed at very low temperatures, being 0.616% in the -30 °C case. The afterpulsing is therefore considered negligible for these SPADs within the temperature range of these measurements.

While the SPAD exhibited negligible afterpulsing, another effect that should be noted in the 0 °C and 30 °C cases is the saturation of the SPAD due to its finite dead time. When the DCR is low, the chance of another dark count occurring within the recharge time of the SPAD is low. However, as the DCR increases at higher temperatures and excess voltages, the recharge time of the SPAD takes up a larger percentage of the total time. Therefore, the probability of other dark counts occurring within the recharge time is increased. The result is that the dark counts occurring within the recharge time may have smaller amplitude or not be registered as counts by the measurement setup. This will cause the apparent dead time of the SPAD to be increased as these dark counts are not measured as counts, but reset the dead time of the SPAD [25]. This effect was observed as a small decrease of the counts in the 0 °C and 30 °C cases at short IATs but was removed to apply the primary DCR fit.

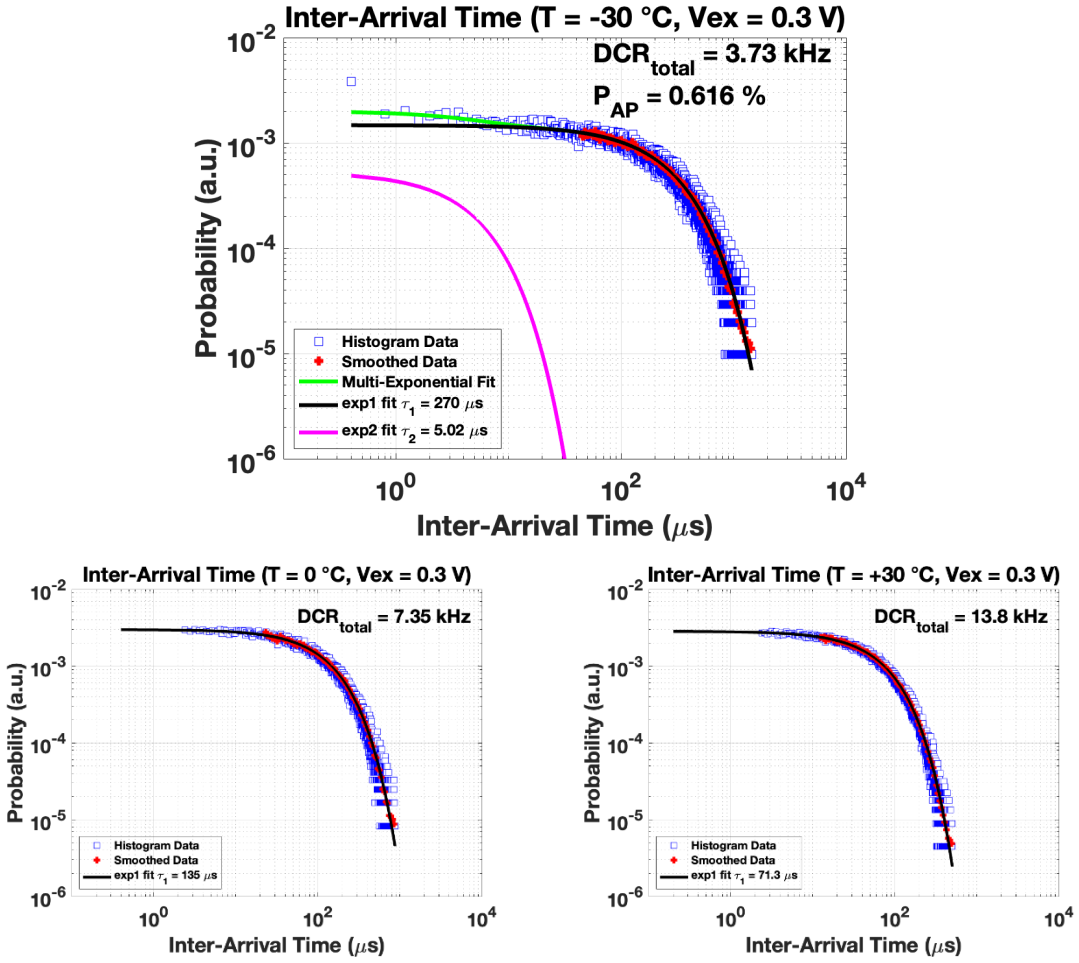

Figure 5-8: A sample of the interarrival time distributions for a 0.3 V excess bias at -30 °C, 0 °C, and 30 °C.

As mentioned previously, measuring the DCR of the SPAD in different temperatures can give insight into the primary dark noise mechanism of the SPAD through the Arrhenius relationship [27]:

$$DCR \propto T^2 \exp\left(\frac{-E_A}{kT}\right), \tag{5-2}$$

where T is the temperature in Kelvin, E_A is the activation energy in eV, and k is Boltzmann's constant. If the DCR of the SPAD was dominated by direct thermal generation or diffusion, the activation energy should be approximately equal to the 1.1 eV band-gap of silicon. Since the defects at the exact midpoint between the conduction and valence bands are the most efficient generation-recombination centres for trap-assisted thermal generation, an activation energy of ~0.55 eV (~E_{gap}/2) would be expected if trap-assisted thermal generation were the dominant mechanism [168], [169]. An activation energy below the mid-gap energy indicates that the dominant noise mechanism is from field-assisted generation mechanisms and band-to-band and trap-assisted tunnelling due to defects outside the mid-gap. Under a high-electric field, the field-assisted generation mechanisms such as Poole-Frenkel effects are dominated by tunnelling effects [27].

Figure 5-9 shows the Arrhenius plot for a SPAD at the 5 measured excess voltages. The plot shows two distinct regions for the activation energy. At low temperatures, the tunnelling effects were seen to be quite dominant, with all excess voltages giving an activation energy of ~ 0.17 eV. At higher temperatures, the activation energy was observed to slightly increase, to ~0.24 eV – 0.25 eV. This indicated that tunnelling is still dominant even at higher temperatures, albeit less so than in our previous n+/p-well SPAD that exhibited an activation energy of 0.11 eV at a 0.4 V excess bias. Typically, with tunnelling dominated noise mechanisms, it should be observed that the SPADs activation energy will decrease with an increasing excess voltage. Here, it was observed that the 0.3 V excess bias SPAD gives slightly lower activation energy. However, this can be attributed to the counting loss of the SPAD due to its long dead time in the passive quench configuration, which is worsened particularly at high temperatures and high excess voltages.

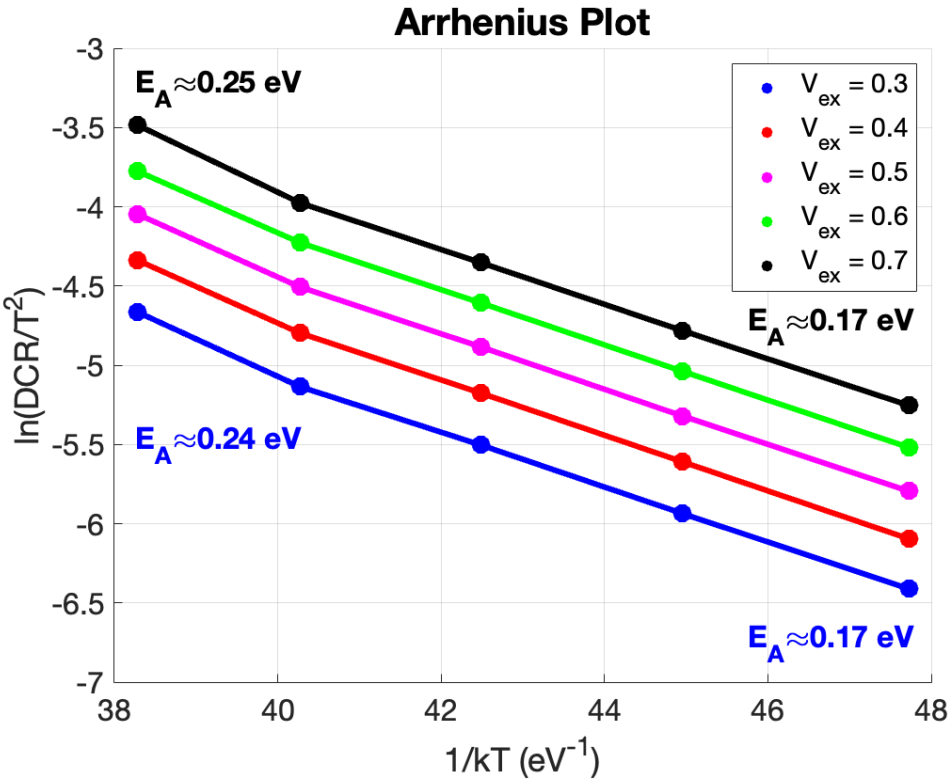

Figure 5-9: The Arrhenius plot obtained from the temperature dependent DCR measurement.

5.2.3. Timing Jitter

To measure the timing jitter of the passive-quench SPAD, it is required to illuminate the SPAD with a periodic pulsed laser and measure the delay for the SPAD to generate an output pulse. The relative delay of the SPAD output with respect to the laser pulse can then be collected for a large number of counts and plotted on a histogram where the FWHM gives the jitter of the complete measurement setup. The setup used for this experiment is shown in Figure 5-10.

The SPAD bias was supplied by an Agilent E3646A DC power supply, and the delay between SPAD counts and the laser sync was collected by using a 50% comparison threshold on a Lecroy Waverunner 625Zi mixed-signal oscilloscope. To provide the trigger for the pulsed laser, a Hewlett Packard 3325B function generator provided a 10 MHz clock to a PicoQuant PDL 800-B laser driver connected to an LDH-P-C-690 laser head with a center wavelength of 685 nm. An important consideration for this measurement is that the

SPAD is ensured to operate in a photon starved regime. When performing a measurement of this type, if the laser power is increased to be too high, the SPAD can be nearly guaranteed to start avalanching near the very beginning of the laser pulse due to the high flux of incoming photons. This will not accurately measure the single-photon timing jitter of the SPAD as it may be significantly underestimated. Normally, it is desired to keep the count rate of the SPAD at less than 1% of the laser repetition rate to avoid pile-up distortion in TCSPC measurements. However, when measuring a standard CMOS SPAD at higher excess voltages, the DCR can become quite large and create a large noise floor to overcome in the timing jitter histogram. Therefore, the laser power in this experiment was increased until the count rate of the SPAD was ~2.5% in order to ensure the SPAD pulses from photon events were at least 10 times larger than the DCR in the worst case (i.e., high excess voltage).

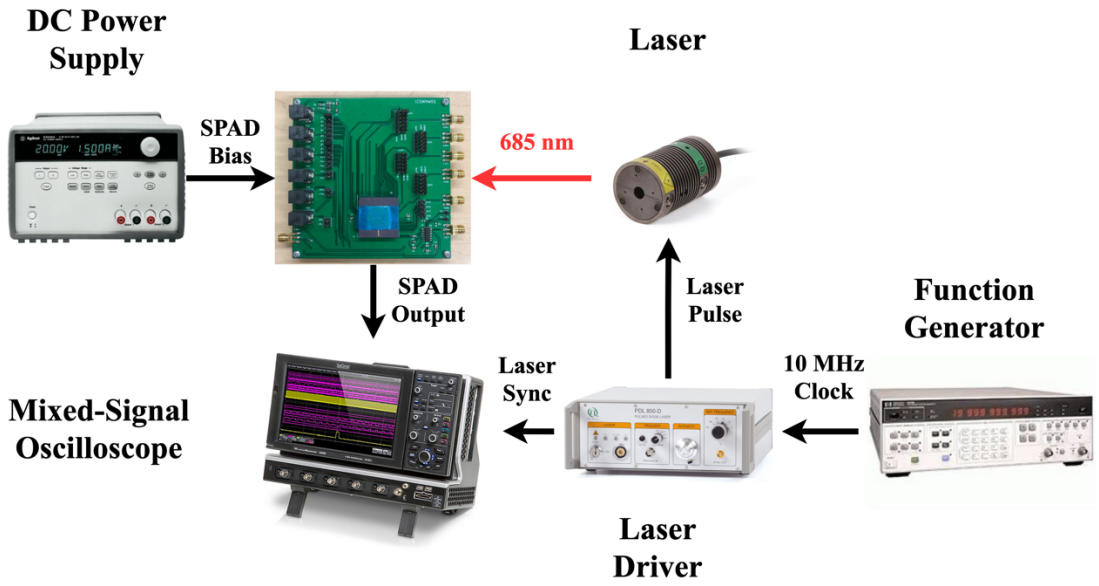

Figure 5-10: Experimental setup used for the timing jitter measurement of the passively quenched SPAD.

The results of the timing jitter measurement for excess voltages of 0.3 V, 0.5 V, and 0.7 V are shown in Figure 5-11. It should be noted that these histograms represent the timing jitter of not just the SPAD, but the complete measurement setup. However, as the jitter in the period of the 10 MHz laser sync signal was ~ 25 ps and the laser pulse is ~ 30 ps wide, the jitter is dominated by the SPAD. As expected, the timing jitter of the SPAD

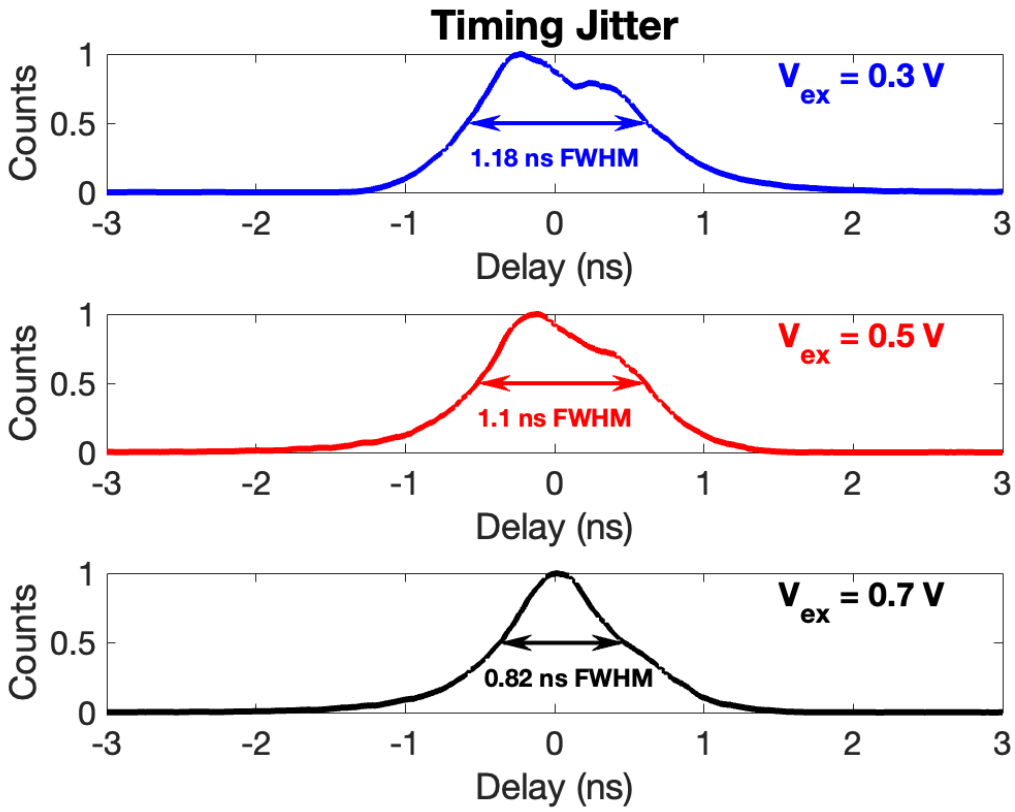

Figure 5-11: Timing jitter histograms for excess biases of 0.3 V, 0.5 V, and 0.7 V.

reduces with increasing excess voltage from 1.18 ns at 0.3 V excess bias, to 0.82 ns at 0.7 V excess bias. This is because a higher excess bias on the SPAD will increase the electric field in the depletion region, reducing the statistical variations in the avalanche build-up time [34]. One point to be noted is the slight difference in shape that is observed in the distribution as the excess bias increases. While the variation of the avalanche build-up time from photons absorbed in the depletion region is Gaussian, the effect of avalanches generated by the diffusion of photogenerated carriers outside the depletion region causes the histogram to have an exponential tail. In the cases of low excess bias, as shown in detail in Figure 5-12, the contribution to the timing jitter from diffused carriers is much larger than at high excess bias. Since the laser wavelength of 685 nm does not align with the peak PDP wavelength of the SPAD, 31 % of the total SPAD counts from photons are a result of diffused carriers. This was obtained by integrating the difference between the exponential

tail of the timing jitter distribution and the total histogram. As the excess bias increases however, the contribution to the timing jitter from depletion region generated carriers becomes more dominant. Since our previous passive quenched n+/p-well SPAD exhibited high DCR, it was only measurable in a time-gated configuration and is thus not directly comparable [162]. In the time-gated mode, it obtained < 200 ps FWHM timing jitter at a 0.3 V excess bias, being much less than this p+/n-well passive quench design. However, the timing jitter of this SPAD was anticipated to be quite large due to the long rise time of the SPAD's output pulse in the unbuffered passive quench configuration, particularly when connected to the large capacitive load of the oscilloscope probe and when using a large quench resistor. With the optimized choice of quench resistor and smaller load capacitance, the fast rise time of the SPAD output can result in increased timing accuracy during sampling, and thus lower jitter.

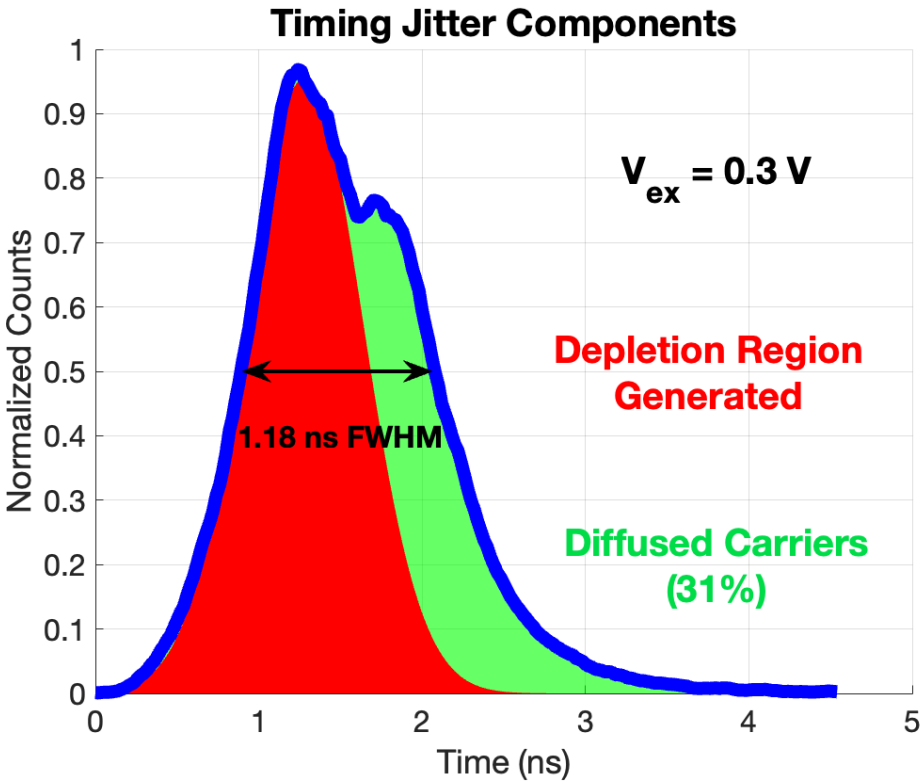

Figure 5-12: Illustration of the timing jitter histogram components obtained with 0.3 V excess bias.

5.2.4. Photon Detection Probability

To measure the PDP of the SPAD with respect to wavelength for different excess voltages, the setup depicted in Figure 5-13 was used. The SPAD bias was supplied by the Agilent E3646A DC power supply, and the SPAD counts were measured by a Lecroy Waverunner 625Zi oscilloscope. As narrow beam pulsed lasers with known optical power were not available for the various wavelengths we intended to measure, the SPAD was illuminated by continuous light from a xenon lamp that was passed through optical bandpass filters and neutral density filters. In this way, we assume the illumination from the lamp is constant on the area of the SPAD, with controlled optical power and wavelength. As the SPAD output is a digital pulse corresponding to the detection of a single photon, it was necessary to determine the number of incident photons on the SPAD. Using a Newport 818-SL wavelength calibrated silicon photodetector (SiPD) coupled to a Newport 1830-C optical power meter, the number of incident photons on the SPAD area per unit of time could be estimated by the following equation:

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

where P_{SiPD} is the power measured by the wavelength calibrated SiPD, λ is the wavelength of the light, h is Planck's constant, c is the speed of light, A_{SPAD} is the active area of the SPAD (~90 μm^2), and A_{SiPD} is the active area of the SiPD (~1 cm^2). The units of the $\frac{\lambda}{hc}$ term is in $[\frac{1}{Joules}]$, and when multiplied by the power, give the number of photons per second that are incident on the SiPD. Since the SPAD active area is smaller than the SiPD, the $\frac{A_{SPAD}}{A_{SiPD}}$ term corrects for the size mismatch. After the total number of photons incident on the SPAD is estimated, the DCR at that excess voltage is subtracted from the total counts, and the ratio of SPAD pulses from photons to the total number of incident photons gives the PDP.

Although this method of measurement can give an estimation of the PDP for the SPAD, it should be noted that there are some limitations, particularly in the passive quench configuration. Ideally, a pulsed laser with known optical power should be used, where the

chance of the SPAD generating an output for a given laser pulse is very small. In this way, there is a negligible probability that a second photon that would have generated an output pulse is missed during the quench and reset of the SPAD. However, under the condition of constant light incident on the SPAD, and with a long dead-time in the passive quench configuration, this probability can be increased and results in the PDP of the SPAD being underestimated, particularly at high excess bias when the DCR and PDP are higher, and the SPAD dead-time occupies a greater proportion of the total time. This is another benefit in operating the SPAD in a fast active quench or time-gated mode, where saturation effects are greatly reduced.

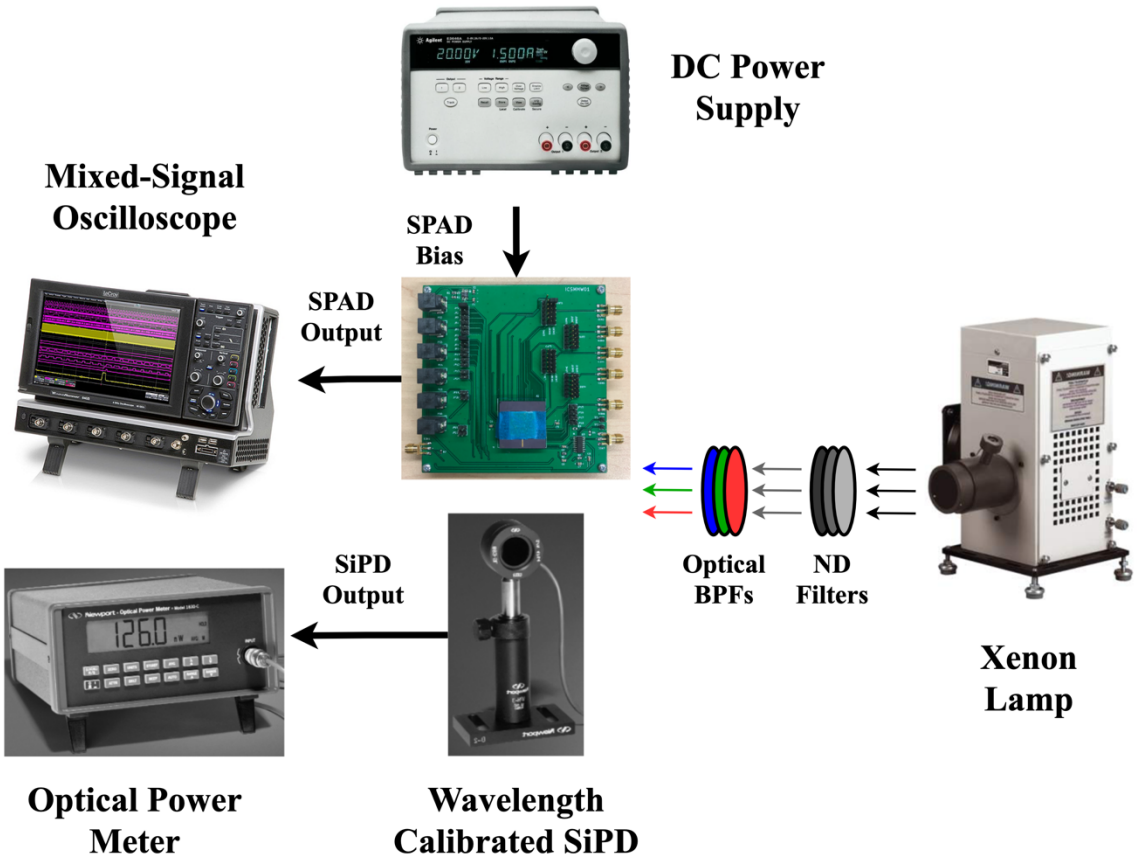

Figure 5-13: Experimental setup for the PDE measurement.

The results of the PDP measurement for several wavelengths between 400 nm and 940 nm are shown in Figure 5-14. Note that the lower range of the wavelengths that could be tested was limited by the wavelength calibrated range of our SiPD, and the upper

wavelength is limited by the cut-off wavelength of silicon, as silicon becomes transparent above 1100 nm [170], with negligible PDP at our maximum optical BPF wavelength of 940 nm. The PDP of the SPAD was tested under 0.3 V, 0.5 V, and 0.7 V excess voltages, with peak PDPs of 13.23%, 16.47%, and 18.13%, respectively. In all cases, the peak PDP occurred at a wavelength of 420 nm. The preferential absorption of photons with a lower wavelength was anticipated, as the p+/n-well junction is close to the surface of the chip. This additionally is consistent with our previous n+/p-well design that exhibited a peak PDP at 440 nm [162]. Another p+/n-well junction in the same TSMC 65 nm process achieved a peak PDP at 470 nm of almost 50% with a 1.5 V excess bias [166]. Due to the short 10 ns deadtime in that work, it was possible to extract the PDP from the SPAD at higher excess voltages while minimizing saturation effects.

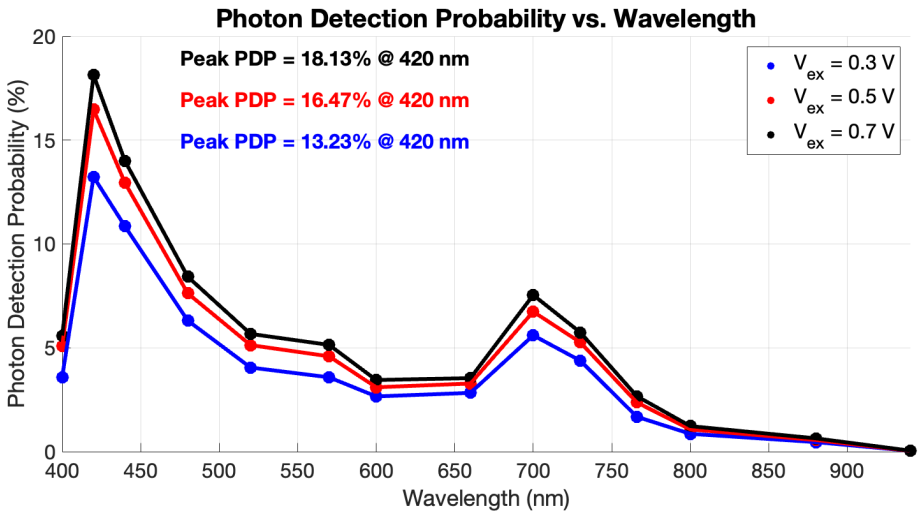

Figure 5-14: Results of the PDE measurement for a range of wavelengths.

To verify the underestimation of the PDP at increased excess bias in the passive quench configuration, the PDP was estimated for the SPAD at the 420 nm peak PDP wavelength for excess voltages between 0.1 V and 1 V in 0.1 V steps. The results of the estimated PDP and DCR are shown in Figure 5-15. The results show that even though the PDP of the SPAD should typically increase with excess voltage, at ~0.8 V the PDP stopped increasing due to the saturation effects. Above this point, not only does the increased dead time from the high DCR effectively prevent the SPAD from detecting photons, but the increased PDP

will mean that photons that may have been detected can be missed during the recharge time of the SPAD from a previous photon detection. One option to reduce the saturation effects is to lower the optical power of the xenon lamp for higher excess voltages. However, in these cases, the photon counts would no longer be 5-10 times larger than the dark counts, and the SPAD could not effectively operate as a TCSPC device. Also, changing the optical power for different excess voltages may introduce additional uncertainty to the measurement as the xenon lamp exhibited "hot-spots" due to the light intensity not being truly constant across its area when incident on the SPAD or calibrated SiPD.

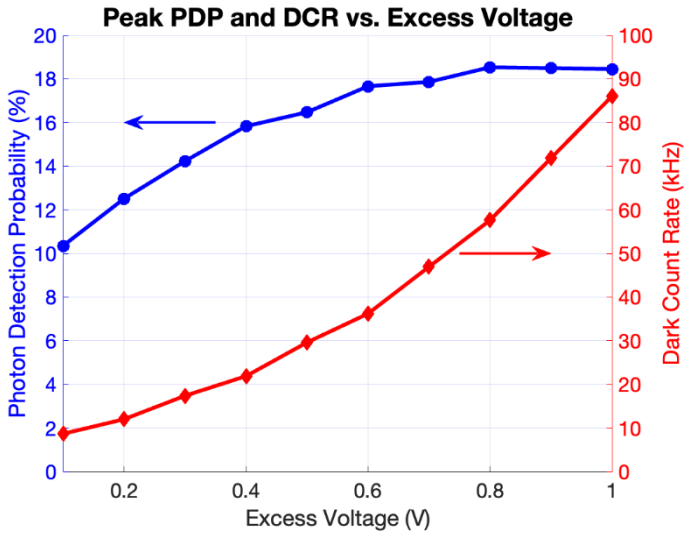

Figure 5-15: Measurement results for the peak PDP and DCR variation with excess voltage on a single plot.

5.3. Multi-Time-Gated SPADs

5.3.1. Time-Gate Window

When assessing the performance of a time-gated SPAD, it is important that the width of the time-gate is experimentally verified. Due to process variations when fabricating a design in CMOS technology, particularly in more advanced nodes, the time-gate window can vary significantly compared to the final post-layout simulation. To verify the time-gate windows, the setup shown in Figure 5-16 was used.

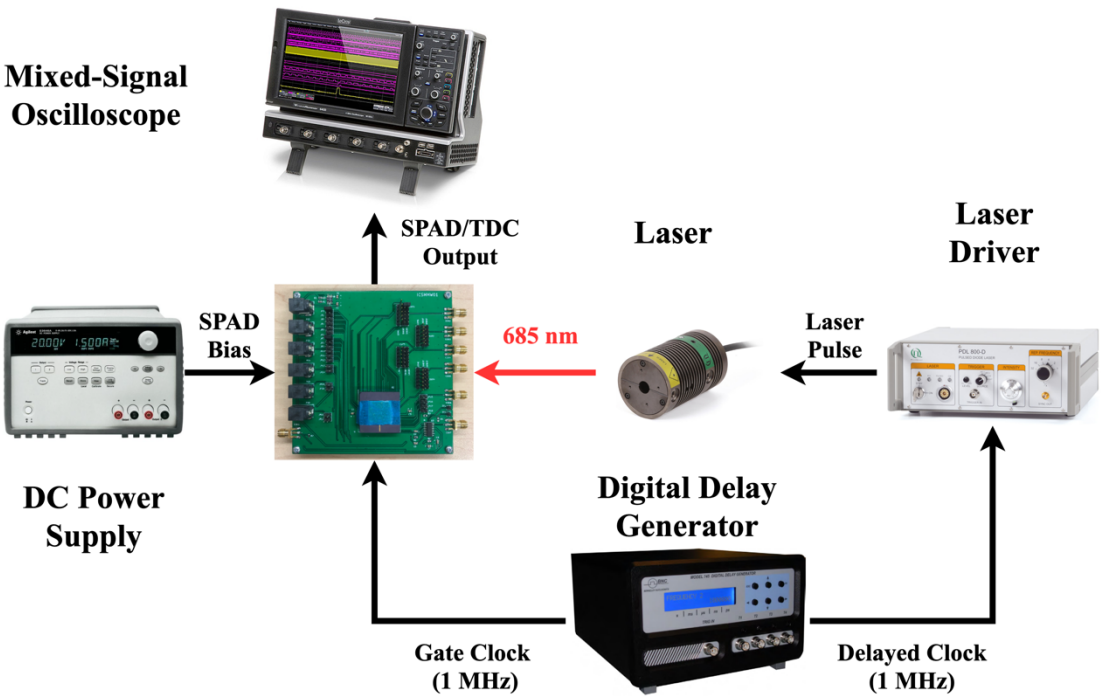

Figure 5-16: Experimental setup used to determine the time-gate windows and the system timing performance.

The SPAD bias and supply voltages for the chip core and IO ring were supplied from an Agilent E3646A DC power supply. The gate clock with a 1 MHz repetition rate was generated by the first channel of a Berkeley Nucleonics Model 745 digital delay generator. To vary the arrival time of the photons within the gate window, the second channel of the delay generator sent delayed replicas of the clock, in 200 ps steps, to the PicoQuant PDL 800-B laser driver connected to an LDH-P-C-690 laser head with a center wavelength of 685 nm. To avoid pile-up distortion in the measurements, the optical power was adjusted such that the total pixel counts were less than 1% of the laser repetition rate at an excess voltage of 0.5 V. Additionally, it was verified that the total counts from photons were more than 10 times larger than the DCR to reduce the influence of the dark counts on the timing measurement results. The counts were recorded by a Lecroy Waverunner 625Zi oscilloscope for 10 million gate windows per point. The number of counts for each point was then used to generate the following plot of the gate windows for several SPADs shown in Figure 5-17.

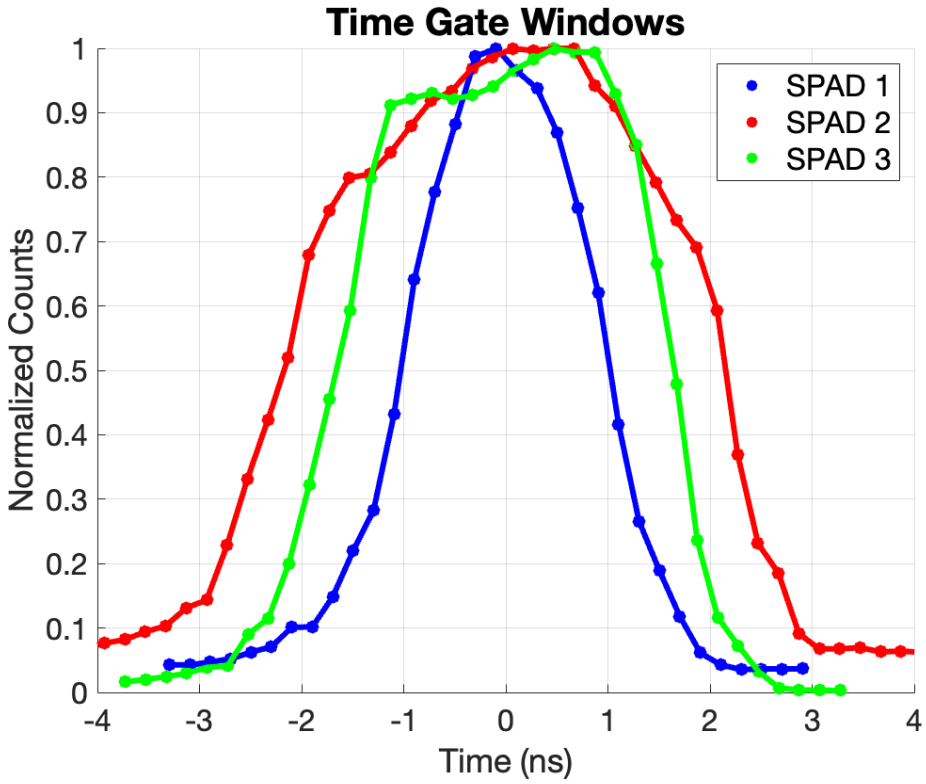

Figure 5-17: Results of the time-gate window measurement for 3 time-gated pixels

For the gate windows shown in Figure 5-17, the SPADs that were tested were in the first, third, and fifth position of the array. Therefore, the gate windows would be shifted from each other for multi-time-gated operation but are centred about 0 ns to illustrate the variation across pixels. The FWHM gate windows for SPAD 1, SPAD 2, and SPAD 3 were 2.01 ns, 4.26 ns, and 3.25 ns, respectively. This level of variation is outside of what was observed from the post-layout process corner simulations. This could be due to the SPAD simulation model, which was kept constant in the corner simulation, but would also have a variation on the physical chip. The shortest gate window shown of 2.01 ns gives an indication of the potential reason for certain pixels generating an output during every gate window. If the gate window is short, then the precharge time of the SPAD is reduced, which may not allow enough time for the SPAD cathode to be charged above the switching point of the time-gated readout circuit. This issue will be discussed in detail in Section 5-4.

From the results in Figure 5-17, it can also be concluded that a future design iteration should ideally implement a DLL. As the multi-purpose delay line in the multi-time-gated design serves the function of generating the gate windows for the SPADs, as well as performing the coarse time-to-digital conversion, the performance of the pixels is highly dependent on the delay. A DLL could lock the delay across the multi-purpose delay line to a reference clock period to achieve the desired gate window widths across PVT variations. To use a DLL, the delay elements within the multi-purpose delay line would have to be converted from standard buffers to delay controlled buffers by using a method such as current starving as used in the TDC in Chapter 3 but without the gating MOSFETs.

5.3.2. Dark Count Probability

To measure the dark noise of the SPAD in a time-gated configuration, the dark count probability (DCP) is measured as opposed to the DCR. The DCP of a time-gated SPAD indicates the probability that a pulse will be measured during a gate window with no incident light on the SPAD. The measurement setup is shown in Figure 5-18. The SPAD bias, and supply voltages for the chip core and IO ring were provided by an Agilent E3646A DC power supply, and SPAD outputs were counted by a Lecroy Waverunner 625Zi oscilloscope. A 1 MHZ clock for the gate windows was generated from a Berkeley Nucleonics Model 745 digital delay generator. The oscilloscope collected 100 ms frames of the SPAD output data, during which time the total counts of the SPAD output for 100,000 gate windows were measured. Two thousand such frames were recorded and histogrammed on the oscilloscope, and the mean value of the ratio of SPAD counts to the total gate windows per frame was taken as the DCP.

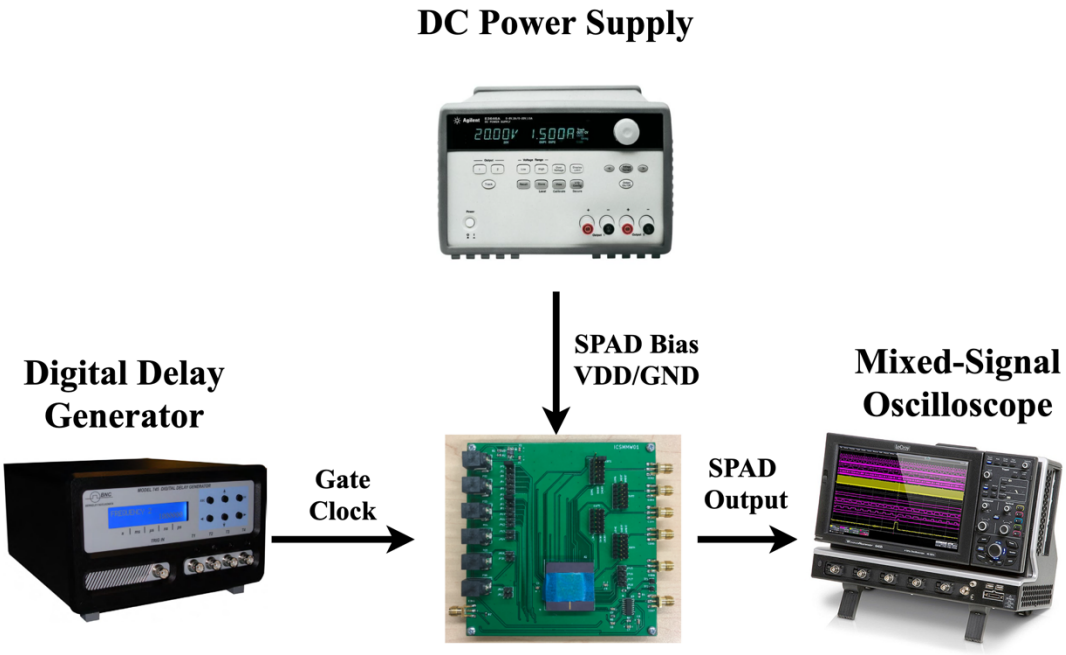

Figure 5-18: Experimental setup for the dark count probability per gate window measurement.

The results of the gate window measurement are shown in Figure 5-19 for 5 SPADs. The main result that is required from this measurement is that the DCP is low enough that SPAD counts from actual photons are dominant in practical TCSPC measurements. As an example, if the laser power in a TCSPC experiment was adjusted such that the total count rate was 1% of the gate windows, then even in the worst case (SPAD 2 at V_{ex} = 0.7 V), our SPAD counts from actual photons could be > 20 times the DCR. From the DCP, it is also possible to estimate the effective DCR (i.e., DCR_{eff}) of the SPAD if the gate window is known by the following expression:

$$DCR_{eff} = \frac{DCP}{T_{ON}} \tag{5-4}$$

where T_{ON} is the width of the gate window. Given the median gate window width from the measured pixels of 3.25 ns and the median DCP of 2.238 x 10^{-4}, the effective DCR can be estimated as 68.87 kHz for a 0.7 V excess bias. The increased DCR compared to the passive quench case is likely due to the insufficient precharge of the SPAD before the gate window is opened or the leakage of the SPAD's cathode voltage during the gate window through the reset MOSFET in its front-end circuitry leading to excess counts.

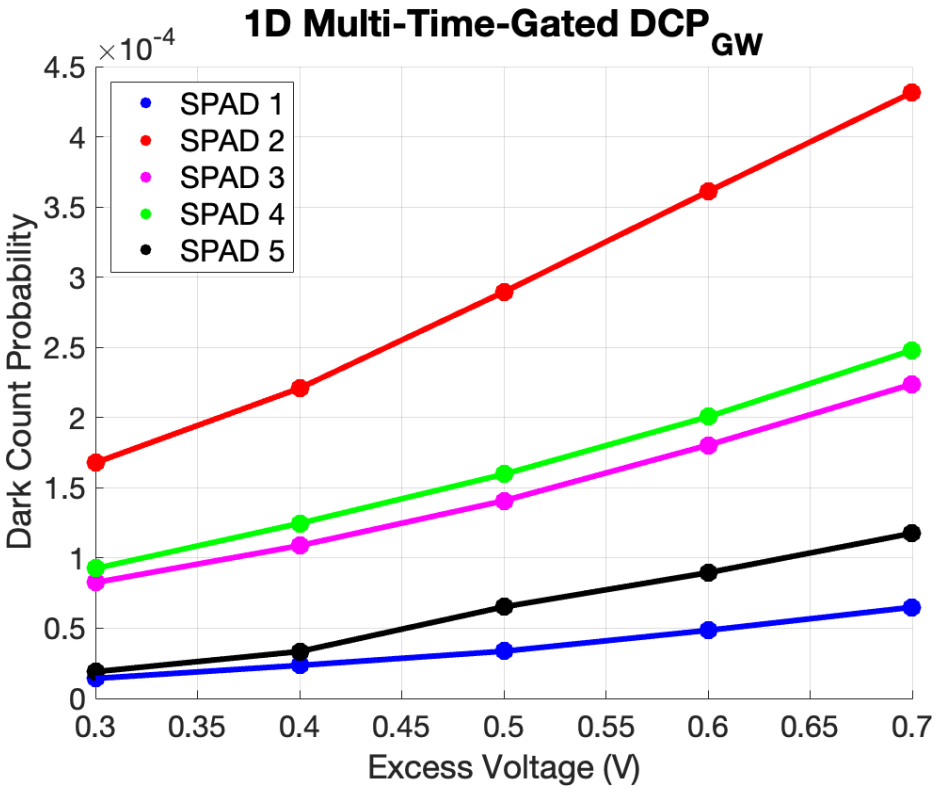

Figure 5-19: Experimental results of the dark count probability per gate window.

5.3.3. Quantization Performance and Jitter

To measure the timing performance of a pixel within the 1D-MTG SPAD array, the same setup was used for the time-gate window measurement. However, 100 ps steps were used for the delay of the gate clock to ensure that several points could be measured for each step on the TDC quantization characteristics. For each of these delay settings, 200 TDC results were sampled, and the average value was used to determine the TDC code associated with that delay. Additionally, the pixel precision for each of the steps was determined as the standard deviation of the TDC output code. This precision includes the effect of the TDC and SPAD jitter, as well as the jitter of the measurement setup.

The results of TDC quantization characteristics and system timing jitter are shown in Figure 5-20 for the pixel whose gate window was 4.26 ns in Figure 5-17 (i.e., Figure 5-20 shows the timing performance within the FWHM gate window of that pixel). The first point that should be noted is, for small delays, the TDC code 000 is never achieved. The reason

for this is due to the fact that during the 000 phase of the delay line, the SPAD is still in the process of being pre-charged above the breakdown voltage and cannot yet generate pulses. This was expected based on previous implementations of this time-gated front-end circuit. The next point that can be noted is the large variation in the step-widths of the quantization characteristics. For a resolution of 440 ps, the TDC achieved a 0.33 LSB_{rms} DNL within the stable region of the gate window. For low-resolution TDCs, as shown in the 4-bit result of Chapter 3, the delay variation should be a smaller fraction of the step width when dummy elements are used to maintain equal capacitive loads at each stage of the delay line. This indicated there were large local variations in the delay elements for the design. A future iteration could use delay elements such as those used in the feedback time amplification TDC of Chapter 3, as they demonstrated much lower local variations.

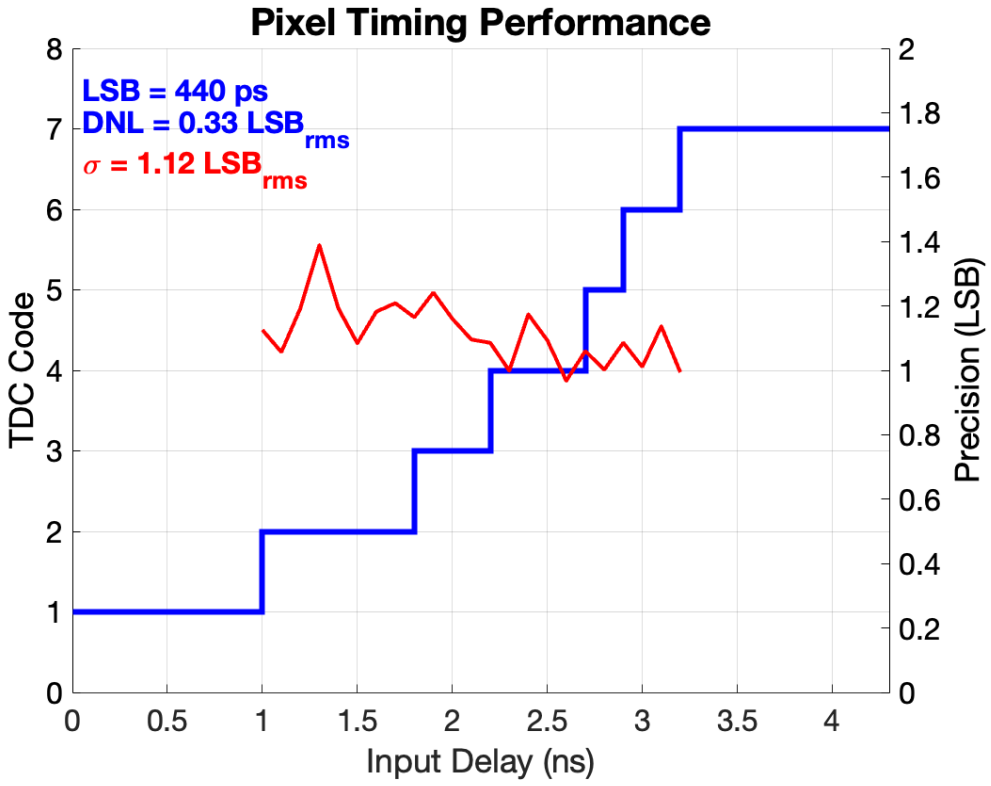

Figure 5-20: Illustration of the achieved timing performance for individual pixels in the multi-time-gated array of SPADs.

The results show that the TDC and SPAD achieve a precision of 1.12 LSB_{rms}. This precision is expected to be dominated by the SPAD based on the results shown for the jitter

of the TDC in Chapter 3, although it cannot be verified as the SPAD and TDC in this design are integrated directly on the chip. This also shows how the jitter of the SPAD is reduced in the time-gated configuration compared to the passive quench approach due to the lower capacitive load on the SPAD, which gives a sharper rise time of the pulse, and the quick detection of the avalanche by the time-gated readout circuitry. The achieved jitter in the current design iteration is additionally limited by the quantization error of the coarse TDC. It should be noted that in addition to reducing the jitter of the SPAD, the system jitter can be improved by reducing the dark noise of the SPAD. Dark counts have an equal probability of occurring within any point of the gate window, leading to additional variation of the system output for a fixed delay. Increasing the ratio of photon counts to dark counts can help to mitigate this issue, but it must be ensured that the total count rate stays below ~1% of the gate windows to mitigate pile-up distortion in TCSPC measurements.

5.4. Conclusions

In this chapter, the measurement results were presented for a p+/n-well SPAD in the TSMC 65 nm process. The SPAD was tested in both a passive quench and time-gated configuration. The SPAD was first characterized in terms of its breakdown voltage, which was determined to be 9.88 V at room temperature with a 4.9 mV/°C temperature coefficient. The passive quench design achieved a maximum DCR of 44.92 kHz at 30 °C with a 0.7 V excess bias, with negligible afterpulsing and the main contributor to the DCR being from tunneling. In the passive quench configuration, the SPAD achieved a timing jitter of 0.82 ns at a 0.7 V excess bias, being quite large due to the large quench resistor and the large capacitive load of the oscilloscope probe, which greatly lengthens the rise time of the SPAD pulse. The SPAD also demonstrated a peak PDP of 18.13% at 420 nm for a 0.7 V excess bias. The preferential detection of shorter wavelengths was expected and consistent with other works due to the closeness of the p+/n-well junction to the surface of the chip.

The same p+/n-well SPAD was also assessed in the multi-time-gated array. Due to issues in the current prototype where certain pixels were giving output at the start of every gate window, the complete multi-time-gated array could not be measured. However, we

were able to extract performance parameters from several pixels. The time-gated pixels exhibited a 3.25 ns median gate window, being smaller than anticipated from the post-layout simulation and with a larger variation than expected from the process corner simulation. This could indicate the possibility of certain pixels generating a constant output due to the insufficient pre-charge of the SPAD cathode, which would additionally explain the increased median DCR of 68.87 kHz compared to the passive quench configuration. Lastly, the quantization and timing jitter performance of the time-gated pixel was extracted. The pixel demonstrated a 440 ps timing resolution with a 0.33 LSB_{rms} DNL. The jitter of the pixel was determined to be 1.12 LSB_{rms} and dominated mainly by the SPAD. From these results, it can be recommended for a future design iteration to lock the multi-purpose delay line within the multi-time-gated array to a reference clock in order to maintain consistent time-gate window widths, consistent resolution, and to ensure the TDC jitter is minimized.

In the current prototype of the multi-time-gated SPAD array, a large number of the pixels exhibited constant output on every clock cycle. As such, those pixels are unable to detect impinging photons and timestamp their arrival as each pixel can only have a single output pulse per gate. For the improvement of a future design iteration, we have identified two possible causes of this behaviour which are illustrated in Figure 5-21.

First, the multi-time-gated pixel could generate an output pulse on every clock edge if the SPAD bias is not sufficiently pre-charged through the MOSFET M1. In the ideal case, the SPAD cathode should be fully charged during the period in which P1 is low. However, if the cathode is not charged to within a threshold voltage of V_{DD}, then as soon as P3 goes low, the MOSFETs M3 and M4 will turn on, and an output pulse will be generated. After this, even if a photon is detected later in the gate window, since the voltage was already below the switching point of M4, no additional pulse will be generated. Second, the pixel could produce an output in every gate window if the leakage of the SPAD's cathode through the reset MOSFET M2 during the gate window is larger than a threshold voltage (i.e., the leakage reduces the cathode voltage below the switching point of M4).

In the current prototype, it was determined the most likely cause of the constant output of certain pixels was the insufficient pre-charge of the SPAD's cathode. This is more likely since, in these pixels, the SPAD output occurred near the very start of the gate window. Additionally, the large variation in the gate window measurement with gates approaching 2 ns would also make the insufficient pre-charge appear more likely as the P1 pulse duration would be reduced. However, a combination of the two mechanisms is also a possibility. The SPAD cathode could be charged only slightly above the M4 switching point, after which only a small amount of leakage is required to generate a false output.

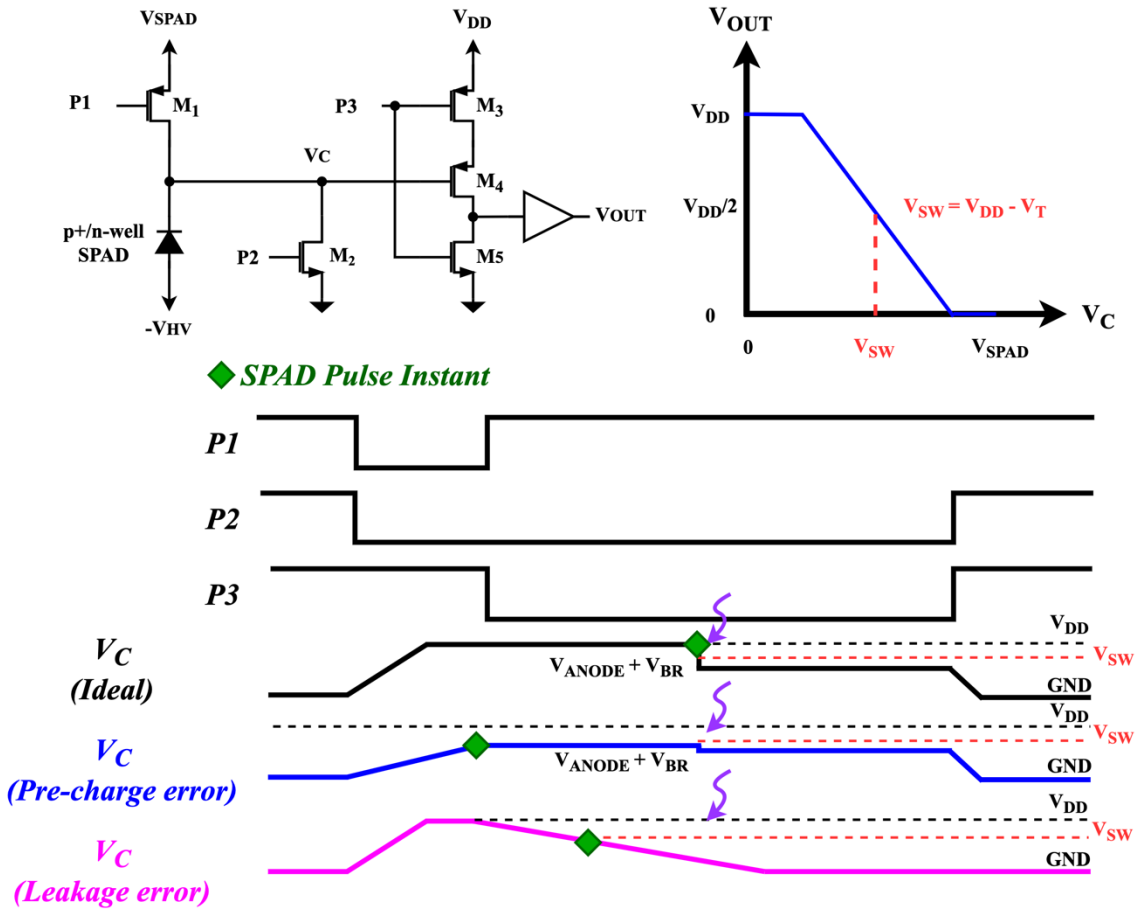

Figure 5-21: Illustration of the constant time-gated pixel output resulting from the insufficient pre-charge during the P1 phase of the gate window or leakage through the reset MOSFET during the gate window.

To solve these design issues, a few steps can be taken. First, this design aimed to make the P1 pulse short to reduce the probability of avalanches building up before the gate window is opened, which would cause a larger power consumption if the photon was

detected while M1 is on. However, the width of the P1 pulse can be increased to further ensure the SPAD cathode is fully charged, or the width of the M1 MOSFET can be increased. Second, if we wish to minimize the leakage of the SPAD's cathode voltage, the MOSFET M2 can be made narrow and long to maximize its resistance in the off state. Third, the aspect ratio of the MOSFET M4 can be reduced to lower its switching point. In this way, the pre-charge and leakage constraints are slightly relaxed, and even if these issues persist, the front-end may not generate an output pulse until a photon is detected by the SPAD. However, reducing the switching threshold may increase the timing jitter of the SPAD, as the avalanche is not detected as early when the rise time is sharpest. In all these potential solutions, it should be ensured that the MOSFET sizes are not increased too much, as the increased capacitance may worsen the timing jitter of the SPAD pixel. Additionally, a larger parasitic capacitance on the SPAD's cathode can result in increased afterpulsing due to the larger amount of charge that will pass through the SPAD to create the same voltage difference during an avalanche, which may become trapped in defects and later released to cause afterpulses if the gating frequency is high.

Chapter 6
Conclusions and Future Work

6.1. Conclusions

This thesis research focussed on the design and measurement of SPADs and TDCs in standard CMOS technology for biomedical imaging applications. When integrated together to form dSiPMs and SPAD imagers, SPADs and TDCs are capable of detecting individual photons and timestamping them with resolution on the order of picoseconds. The high level of performance achieved when SPADs and TDC are closely integrated on the same chip, and their performance advantages over previous detector technologies such as avalanche photodiodes and photomultiplier tubes have led SPAD-based sensors to be a heavily researched topic in recent years. Specifically, the application of SPAD-based sensors to biomedical imaging such as PET, FLIM, and DOT was a key research stream as it provides the potential for developing next-generation high-performance medical imaging systems at a low cost and in a compact size.

Prior to creating our original designs, the thesis first provided a detailed review of fundamental concepts, performance metrics, and state-of-the-art results of recently published TDCs in Chapter 2. It was concluded that while the highest performance TDCs use methods such as Vernier ring oscillators, pulse shrinking ring oscillators, multipath ring oscillators, $\Delta\Sigma$ modulation, and time amplification; TDCs integrated with SPAD arrays generally opt for the use of simpler structures such as the delay line interpolation of a DLL or basic ring oscillators. This is partly because high-performance TDCs often require too large of an area to be effectively integrated with a SPAD array without comprising the fill factor.

The TDC design presented in Chapter 3 aimed to obtain high resolution within a small silicon area using feedback time amplification. By performing the conversion in three

*Part of this work was accepted for publication in IEEE Reviews in Biomedical Engineering as: R. Scott, W. Jiang, M. J. Deen, CMOS Time-to-Digital Converters for Biomedical Imaging Applications.

stages, the quantization error from an initial coarse measurement can be time amplified and passed to a second identical TDC stage to improve the resolution by a factor equal to the time amplification factor. In this design, the first stage of the TDC is reused for the third conversion, and the time amplifier is shared through a multiplexing scheme to reduce the TDC's area. While the TDC fabricated in a standard TSMC 65 nm CMOS process achieves a fine resolution of 4.14 ps over a 3.2 ns dynamic range within a small silicon area of 0.016 mm^2, the measured results of the current prototype had a high INL of 15.3 LSB_{rms}. Future iterations of this TDC should focus on improving the linearity and reducing the jitter through methods such as incorporating a DLL, using larger transistor sizes for the delay cells to minimize the effect of process variations, taking extra care to ensure matching between timing-sensitive signals, and modifying the delay elements to minimize error from charge injection in the gated delay lines.

Following this, the design is presented for a multi-time-gated SPAD array using a custom p+/n-well SPAD in the same TSMC 65 nm CMOS process in Chapter 4. Previous works on time-gated SPADs have proposed shifting a gate window over an array of SPADs with small steps in order to generate histograms for TCSPC measurements. However, as only one event can be measured by a SPAD per gate window, to generate a histogram with a large number of bins, many measurements must be taken for each step of the gate window, which limits the measurement speed. The multi-time-gated design aims to simultaneously provide shifted gate windows to an array of SPADs such that each bin of the histogram can be generated at the same time for high-speed imaging. Also, it was noted that the delay line used for generating the delayed gate windows and for generating the gating signals for each individual SPAD pixel has an identical structure of a basic delay line TDC. As such, the multi-purpose delay line was also used to achieve coarse time-to-digital conversion.

Due to the current implementation of the multi-time-gated array not being fully functional, the p+/n-well SPAD was first characterized in a basic passive quench configuration in Chapter 5. In the passive quench configuration, the SPAD demonstrated a breakdown voltage of 9.88 V at room temperature with a 4.9 mV/ ºC temperature coefficient. This low breakdown voltage, and the low activation energy estimated from our

Arrhenius plot, indicate that the main contributor to the SPADs dark noise is from tunnelling due to the thinner depletion region in more advanced technology nodes where the doping concentrations are higher. The maximum DCR of the SPAD was measured as 44.9 kHz at 30 ºC for a 0.7 V excess bias. At the 0.7 V excess bias, the SPAD also demonstrated a peak PDP of 18.1 % at a 420 nm wavelength and a 0.82 ns timing jitter. For individual pixels of the multi-time-gated design, the median gate window was then determined to be 3.25 ns, over which a median DCP of 2.2×10^{-4} was observed for a 0.7 V excess bias. The multi-time-gated pixel was also characterized in terms of its timing performance, demonstrating a 440 ps timing resolution with ~ 1 LSB_{rms} timing jitter. Based on these results, it was concluded that the likely cause of the faulty pixels is due to the insufficient pre-charge of the SPAD cathode at the start of the gate window or the leakage of the SPAD's cathode voltage through the reset MOSFET in the time-gated front-end circuitry. Based on these conclusions, we provided possibilities for improvement in a future design iteration.

Based on these results and the parameters summarized in Table 1-1, the p+/n-well SPAD used in this thesis would be most suitable to the PET application, as its peak PDP aligns with the peak emission of typical LYSO scintillators at 420 nm. However, due to requirements of PET imaging, time-gating is not generally supported, and the multi-time-gated topology would be most suitable to an application such as DOT/NIROT. If a deeper junction was used, such as the p-well/deep n-well or deep n-well/p-substrate, then the peak PDP could be designed to be at a longer wavelength more suitable to DOT. Using separate SPADs within the array to cover different portions of the timing range can also help in reducing the measurement time, which is important in DOT since it can minimize motion artifacts from the patient. Additionally, the feedback time amplification TDC could in future be integrated as a fine interpolator for the multi-time-gated array to achieve resolutions below 10 ps while maintaining a compact size, which is desirable to the DOT application.

6.2. Future Work

The ability to timestamp events with picosecond resolution and high throughput using miniaturized circuits has led TDCs to be integrated with the most sensitive photodetectors known as SPADs in dSiPMs and SPAD imagers. In recent years, we have seen vast improvements in SPAD-based sensor technology. However, we have identified several research challenges for TDCs and SPADs that we foresee for the coming years, so that SPAD-based sensor performance can continue to be optimized and the technology can be further embraced in the commercial market.

A. Application Targeted Designs

Although it is desirable for SPAD-based sensors to be diverse and applicable to various applications, many of the recent works have yet to achieve the desired level of performance for their targeted application. As such, before high-performance general dSiPMs and SPAD imagers are possible, we expect to see more specialized sensor structures designed for specific applications. Specifically, for PET imaging, we expect to see TDCs targeting higher resolutions. In both FLIM and NIROT, we expect to see SPAD arrays with a larger number of TDCs in order to increase throughput for high-speed operation. Common across many applications, designers will attempt to keep pushing the area and power lower to provide high fill factors and minimize the TDC's impact on the system power consumption.

B. SPAD Dark Noise

Even under conditions with no light, SPADs may produce output pulses known as dark counts. Dark noise in SPADs primarily results from avalanches resulting from: thermally generated carriers described by Shockley-Read-Hall recombination theory; and trap-assisted and band-to-band tunneling through the depletion region. Much work in recent years has investigated the physical modelling of the dark noise phenomena (e.g. [27]), and minimizing of dark noise in CMOS SPADs through the design of efficient guard ring structures [28], [171]–[173]. While the timing performance of TDCs is improved in advanced CMOS processes, the main limitation in advanced standard CMOS is the large increase in the SPAD's dark noise. In addition to narrower depletion regions, the PDE of

SPADs in advanced CMOS processes is often reduced since the SPADs must operate from lower excess voltages in order to minimize the dark noise. Therefore, the PDE may be reduced despite the improvements in fill-factor. In addition to optimizing SPADs for reduced dark noise in advanced processes, approaches such as differential sensing SPADs can be used to mitigate common-mode noise [135].

C. Detector Level Models

Sharing TDCs between a group of SPADs was commonly implemented in many works, but the impact of these design choices often goes unjustified as there was a lack of modelling on the effects of TDC sharing. In recent years, this topic was addressed by groups encouraging modelling early in the system design process such that power consumption can be minimized and fill factor can be maximized while ensuring high-performance. In [11], simulations were performed to assess the effectiveness of the best linear unbiased estimator (BLUE) at calculating the ToF for a dSiPM coupled to a scintillator. In the case of typical LYSO scintillators, it was shown that only the first two photon timestamps were required in order to achieve the best possible BLUE timing estimate. This indicated that for standard LYSO crystals, it is more important to focus on having high-resolution TDCs with low jitter than to integrate a large number of TDCs into the system. However, if employing the use of next-generation scintillators with prompt photon generation, a greater number of TDCs provides a larger benefit.

Although the aforementioned work takes into account the effect of all the main jitter sources of the system, it does not include any impact of the detector's dead time, which increases with a lower number of TDCs. In [174], the effect of the TDC dead times was included in order to provide an analytical approach for finding the optimal number of TDCs to maximize the SNR of the ToF estimation. Using numerical Monte Carlo simulations, the approach was verified. The results show that for an array of 64 SPADs, 95% of the maximum SNR can be achieved using only 15 TDCs, lowering the total TDC contribution to the detector's power consumption by more than 75%.

While these analyses show promise, they have yet to be embraced in the majority of detector designs that were published and verified extensively on physical implementations.

Additionally, further models need to be developed that are more comprehensive. Expanding on these works to include effects such as dark noise, afterpulsing, PVT variations, and chip timing skew will lead to more robust dSiPM pixel architectures.

D. System-Level Models

While the aforementioned models have focused on optimizing pixel structure at the detector level, these models have not yet considered the loss of spatial resolution that results from TDC sharing at the system level. For example, in [2] groups of 720 SPADs share a pair of interleaved TDCs, resulting in a 720:1 spatial compression loss. As SPAD-based sensors will often either operate in a photon starved mode or only require the first photons of gamma events, sharing TDCs has shown great benefits as it allows for more complex TDC structures to improve timing resolution. As designs continue to increase in complexity, factors such as the sharing of TDCs may show consequences at the system level (e.g., in a complete PET ring) due to spatial compression losses when grouping many SPADs. Factors such as this may be considered so that designs can be optimized for not only the individual detectors, but for system-level imaging. Potential solutions to this issue also include encoding the address of the first conducting SPAD into the data that is output from the TDC. In this way, many SPADs may share a single TDC, but the spatial information is retained. However, this would require additional circuit area and bandwidth compared to pure spatial compression.

E. 3D-Stacked Sensors

Regardless of whether TDCs are integrated with each SPAD or shared between groups, 3D dSiPMs still appear promising to improve performance (see Figure 6-1). With TDCs being integrated in a separate tier than the SPADs, it is possible to achieve fill factors above 70% [20]. This would greatly improve the detector PDE, allowing weaker photon events to be detected. With extra silicon area being available, more advanced TDC structures can be used that offer improved timing resolution. The 3D structure with a large number of TDCs also yields the ability to minimize timing skew from routing the SPADs to TDCs by replicating the exact same 3D structure for each individual SPAD or miniature SPAD array. It should be noted that with 3D stacking, SPADs and TDCs no longer need to be

implemented in the same process. SPADs could be implemented in a technology optimized for photodetection, while the TDCs could be implemented in an advanced high-speed digital process that can obtain fine timing resolution in a smaller area.

Additionally, sensors with a 3D structure allow other digital circuits to be integrated in the system to offer real-time embedded processing or compression. In the recent LiDAR work in [144], a large 256 × 256 SPAD array was integrated in the top tier in a 90 nm process. Every 4 × 4 group of SPADs was connected to a photon processing unit in a 40 nm process bottom tier. Due to the extra silicon area available, a high fill factor of 51% was achieved in the photodetection tier, while also allowing the sensor to operate in 6 different modes. This work has demonstrated that in the 3D structure, a high-fill factor can be achieved even with additional circuitry for multi-mode operation. Therefore, it allows the possibility for the design of SPAD-based sensors that are more generalized and able to meet the requirements for various applications. However, this comes at an increased cost due to 3D integration and the potential use of specialized processes.

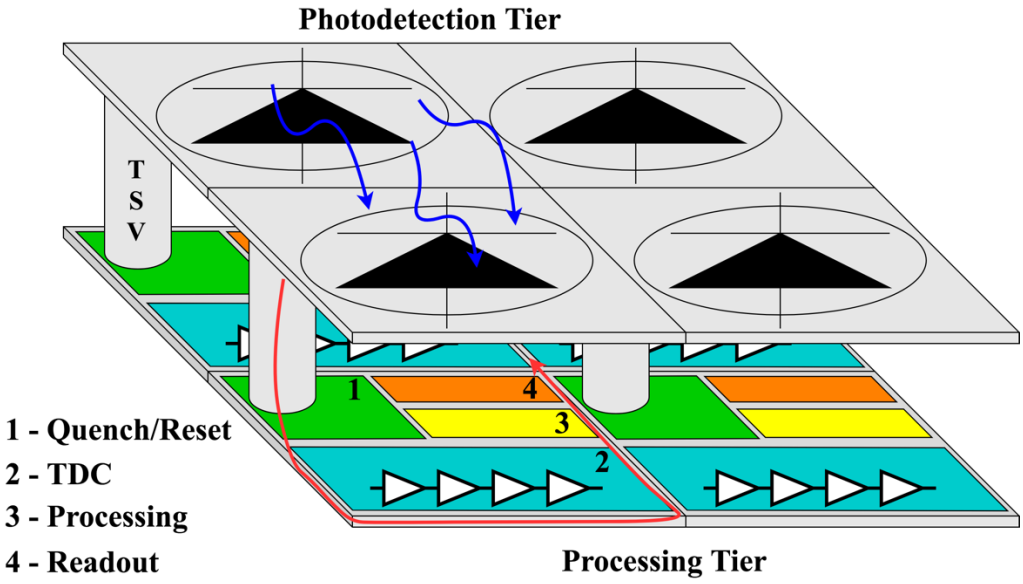

Figure 6-1: Illustration of a 3D pixel structure for a SPAD-based sensor. (© 2021 IEEE)

F. Readout Optimization

In future SPAD-based sensors aiming for high throughput operation, large numbers of TDCs may be implemented on chip to quickly construct a full frame; but these designs have

to consider optimized readout approaches. Since the *start* signal normally comes from a SPAD pulse, and the system clock serves as the *stop* signal, in the worst case, a large number of TDCs will have data available at the system clock edge. This can lead to system-level pile-ups as the data is sent off-chip. Some previous works, such as [129], have opted for the use of embedded memories to store conversion results. This allows the system to continue operation as data will be addressed and read off-chip by interfacing with the RAMs rather than the detection circuitry [20]. If a full-frame image is not needed from each pixel of the detector, the event-driven operation has shown to be more effective in reducing the required bandwidth and providing higher frequency operation.

G. Embedded Processing

In addition to event-driven readout circuitry, embedded processing is an approach for maintaining scalability and reducing bandwidth requirements in larger systems. Current embedded processing has included the creation of histograms from the measured TDC data [4], [175]. Accumulating TDC timestamps in bins associated with specific time intervals is highly useful if it is desired to fit the photon ToF data to a distribution, and also reduces the amount of data to be transferred. In recent years, a new TDC topology was developed that is capable of quickly building histograms while mitigating pile-up distortion by allowing multiple events to be timestamped at the same time by using an XOR tree to encode SPAD pulses into rising and falling edges along a delay line [144], [176]. In [175], the detector's histogram mode was reported to increase the count rate to 85 times higher than the standard TCSPC mode.

Recently, a new approach known as partial histogramming (PH) was applied in the design of [177] that targeted LiDAR applications. By noticing that in this application, the histogram normally has a single dominant peak and very few counts as we move further from the peak, the full histogram does not need to be stored. Using this approach, 3 exposure periods are used where the peak of the TDC timestamps are found with greater precision on each iteration. After the peak is found, a histogram with fewer bins and higher resolution can be formed by taking a series of bins that are centered around the peak. A 14.9:1 compression ratio was achieved using this method compared to the traditional

histogram readout. Beyond only data compression, motion detection triggered LiDAR was implemented in [178] that reduced the system power consumption by 70%. Further work on embedded processing such as gamma event detection for PET imaging or embedded fluorescence lifetime estimation for FLIM would greatly benefit the applications of biomedical imaging in the future, as they provide a way to remove the readout bottleneck and allow for high-speed imagers with a larger number of high-resolution timing elements.

H. FPGA-Based TDCs

Due to the quickly evolving modern FPGA technologies, TDCs implemented in FPGAs are likely to become more common approaches for dSiPM and SPAD imager applications. Shorter development cycles, lower costs, and ease of integration with complex digital systems are key advantages that would arise from FPGA TDCs [120]. Tapped delay line TDCs were the most common approach in FPGAs, achieving resolutions in the range of several picoseconds with implementations in more advanced technologies at a lower cost, compared to most of the ASIC-based works [179]–[183].

While FPGA-based TDCs are capable of achieving resolutions comparable to that of ASIC TDCs, much of the research was based on linearization techniques, as FPGA delay elements are based on standard cells that were not custom-designed for equal delays. Post-conversion calibration is very common in FPGA TDCs, and while it was proven to be effective, it can add to the TDC dead time and reduce the throughput [120]. As such, linearizing TDCs while having minimal impact on the throughout would appear to be a valuable stream of research. Also, most modern FPGAs include a microprocessor, so the potential for automatically generated TDCs becomes possible [120]. Analyzing the current linearity performance of a delay line and then subsequently rerouting and reassessing the performance could be done iteratively until the desired performance is reached.

I. Artificial Intelligence

As we have seen large advancements in AI in recent years, and its adoption into many areas of engineering, it is natural that we expect to see it being used to advance SPAD-based sensor technology. Typically, in a design cycle for an IC, the schematic design, layout, post-layout simulation, fabrication, and verification lead to lengthy design cycles that normally will require multiple iterations to reach the desired level of performance.

Commercially viable fabrications on the first attempt could be achievable through design optimization using AI and machine learning techniques, while providing a method of obtaining optimal performance, shorter development time, and lower costs.

In [184], the voltage-to-time converter (VTC) and TDC nonlinearities were calibrated out in post-processing from a coarse-fine ADC that employed a TDC-based fine stage. The machine learning based calibration was capable of correcting the VTC gain errors and TDC delay line nonlinearities. It was shown in [185] that a backpropagation machine learning algorithm could be used to calibrate the digitally-controlled delay lines to optimize the TDC's linearity. Aside from the optimization of circuits, we expect that as detector performance improves, more AI-based models will be used on the software side after detector data was collected. Recently, an artificial neural network was used to extract the fluorescence lifetimes from each pixel of the SPAD array in [186]. With higher quality data, more accurate biological images can be reconstructed, leading to earlier and more precise diagnoses.

References

[1] 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, no. 05, pp. P05024----P05024, May 2013, doi: 10.1088/1748-0221/8/05/P05024.

[2] L. H. C. C. Braga *et al.*, "A Fully Digital 8×16 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, Jan. 2014, doi: 10.1109/JSSC.2013.2284351.

[3] 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, doi: 10.1109/JSSC.2013.2293777.

[4] N. Krstajić, J. Levitt, S. Poland, S. Ameer-Beg, and R. Henderson, "256 × 2 SPAD line sensor for time resolved fluorescence spectroscopy," *Opt. Express*, vol. 23, no. 5, p. 5653, Mar. 2015, doi: 10.1364/OE.23.005653.

[5] A. Farina *et al.*, "Time-Domain Functional Diffuse Optical Tomography System Based on Fiber-Free Silicon Photomultipliers," *Appl. Sci.*, vol. 7, no. 12, p. 1235, Nov. 2017, doi: 10.3390/app7121235.

[6] M. Alayed and M. J. Deen, "Time-Resolved Diffuse Optical Spectroscopy and Imaging Using Solid-State Detectors: Characteristics, Present Status, and Research Challenges," *Sensors*, vol. 17, no. 9, p. 2115, Sep. 2017, doi: 10.3390/s17092115.

[7] M. Kanoun, L. Arpin, V.-P. Rheaume, M.-A. Tetreault, Y. Berube-Lauziere, and R. Fontaine, "A 10-bit, 3 ps rms precision time-to-digital converter for diffuse optical tomography measurements," in *2014 21st IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Dec. 2014, no. April 2015, pp. 554–557, doi: 10.1109/ICECS.2014.7050045.

[8] S. R. Cherry, J. A. Sorenson, and M. E. Phelps, *Physics in nuclear medicine*. Elsevier/Saunders, 2012.

[9] W. Jiang, Y. Chalich, and M.J. Deen, "Sensors for Positron Emission Tomography Applications," *Sensors*, vol. 19, no. 22, p. 5019, Nov. 2019, doi: 10.3390/s19225019.

[10] 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, Nov. 2010, doi: 10.3390/s101110484.

[11] M.-A. Tetrault, A. Corbeil Therrien, W. Lemaire, R. Fontaine, and J.-F. Pratte, "TDC Array Tradeoffs in Current and Upcoming Digital SiPM Detectors for Time-of-Flight PET," *IEEE Trans. Nucl. Sci.*, vol. 64, no. 3, pp. 925–932, Mar. 2017, doi: 10.1109/TNS.2017.2665878.

[12] M.-A. Tetrault, A. C. Therrien, E. D. Lamy, A. Boisvert, R. Fontaine, and J.-F. Pratte, "Dark Count Impact for First Photon Discriminators for SPAD Digital Arrays in PET," *IEEE Trans. Nucl. Sci.*, vol. 62, no. 3, pp. 719–726, Jun. 2015, doi: 10.1109/TNS.2015.2420795.

[13] M. Kfouri *et al.*, "Toward a Miniaturized Wireless Fluorescence-Based Diagnostic Imaging System," *IEEE J. Sel. Top. Quantum Electron.*, vol. 14, no. 1, pp. 226–234, 2008, doi: 10.1109/JSTQE.2007.911765.

[14] 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, doi: 10.3390/s90100430.

[15] K. Suhling *et al.*, "Fluorescence lifetime imaging (FLIM): Basic concepts and some recent developments," *Med. Photonics*, vol. 27, pp. 3–40, May 2015, doi: 10.1016/j.medpho.2014.12.001.

[16] 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, doi: 10.1109/TED.2015.2503718.

[17] 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, Nov. 2008, doi: 10.1109/JSSC.2008.2005818.

[18] M. Gersbach *et al.*, "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, Jun. 2012, doi: 10.1109/JSSC.2012.2188466.

[19] D. Tyndall *et al.*, "A High-Throughput Time-Resolved Mini-Silicon Photomultiplier With Embedded Fluorescence Lifetime Estimation in 0.13 μm CMOS," *IEEE Trans. Biomed. Circuits Syst.*, vol. 6, no. 6, pp. 562–570, Dec. 2012, doi: 10.1109/TBCAS.2012.2222639.

[20] J. M. Pavia, M. Scandini, S. Lindner, M. Wolf, and E. Charbon, "A 1 × 400 Backside-Illuminated SPAD Sensor With 49.7 ps Resolution, 30 pJ/Sample TDCs Fabricated in 3D CMOS Technology for Near-Infrared Optical Tomography," *IEEE J. Solid-State Circuits*, vol. 50, no. 10, pp. 2406–2418, Oct. 2015, doi: 10.1109/JSSC.2015.2467170.

[21] P. Lecoq, E. Auffray, and A. Knapitsch, "How photonic crystals can improve the timing resolution of scintillators," *IEEE Trans. Nucl. Sci.*, vol. 60, no. 3, pp. 1653–1657, 2013, doi: 10.1109/TNS.2013.2260768.

[22] R. Datta, T. M. Heaster, J. T. Sharick, A. A. Gillette, and M. C. Skala, "Fluorescence lifetime imaging microscopy: fundamentals and advances in instrumentation, analysis, and applications," *J. Biomed. Opt.*, vol. 25, no. 07, p. 1, May 2020, doi: 10.1117/1.JBO.25.7.071203.

[23] Y. Chalich, "Design of CMOS SPADs Towards High-Performance Imagers," McMaster University, 2019.

[24] 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, Jan. 2014, doi: 10.1109/LED.2013.2288983.

[25] L. Neri *et al.*, "Note: Dead time causes and correction method for single photon avalanche diode devices," *Rev. Sci. Instrum.*, vol. 81, no. 8, p. 86102, Aug. 2010, doi: 10.1063/1.3476317.

[26] A. Gallivanoni, I. Rech, and M. Ghioni, "Progress in quenching circuits for single photon avalanche diodes," *IEEE Trans. Nucl. Sci.*, vol. 57, no. 6, pp. 3815–3826, Dec. 2010, doi: 10.1109/TNS.2010.2074213.

[27] Y. Xu, P. Xiang, and X. Xie, "Comprehensive understanding of dark count mechanisms of single-photon avalanche diodes fabricated in deep sub-micron CMOS technologies," *Solid. State. Electron.*, vol. 129, pp. 168–174, Mar. 2017, doi: 10.1016/j.sse.2016.11.009.

[28] M.-J. J. Lee and W.-Y. Y. Choi, "Performance Optimization and Improvement of Silicon

Avalanche Photodetectors in Standard CMOS Technology," *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 2, pp. 1–13, Mar. 2018, doi: 10.1109/JSTQE.2017.2754359.

[29] V. Savuskan, M. Javitt, G. Visokolov, I. Brouk, and Y. Nemirovsky, "Selecting Single Photon Avalanche Diode (SPAD) Passive-Quenching Resistance: An Approach," *IEEE Sens. J.*, vol. 13, no. 6, pp. 2322–2328, Jun. 2013, doi: 10.1109/JSEN.2013.2253603.

[30] G. Bonanno *et al.*, "Precision Measurements of Photon Detection Efficiency for SiPM Detectors," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 610, no. 1, pp. 93–97, Oct. 2009, doi: 10.1016/J.NIMA.2009.05.117.

[31] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Appl. Opt.*, vol. 35, no. 12, p. 1956, Apr. 1996, doi: 10.1364/AO.35.001956.

[32] S. Cova, M. Ghioni, A. Lotito, I. Rech, and F. Zappa, "Evolution and prospects for single-photon avalanche diodes and quenching circuits," *J. Mod. Opt.*, vol. 51–9, no. 10, pp. 1267–1288, 2004, doi: 10.1080/09500340408235272.

[33] D. Hall, Y. H. Liu, and Y. H. Lo, "Single photon avalanche detectors: Prospects of new quenching and gain mechanisms," *Nanophotonics*, vol. 4, no. 1, pp. 397–412, 2015, doi: 10.1515/nanoph-2015-0021.

[34] F. Sun, Y. Xu, Z. Wu, and J. Zhang, "A Simple Analytic Modeling Method for SPAD Timing Jitter Prediction," *IEEE J. Electron Devices Soc.*, vol. 7, no. March, pp. 261–267, 2019, doi: 10.1109/JEDS.2019.2895151.

[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, doi: 10.1016/j.nima.2006.05.060.

[36] A. Bandyopadhyay, M. J. Deen, L. E. Tarof, and W. Clark, "A simplified approach to time-domain modeling of avalanche photodiodes," *IEEE J. Quantum Electron.*, vol. 34, no. 4, pp. 691–699, Apr. 1998, doi: 10.1109/3.663452.

[37] C. L. F. Ma, M. J. Deen, L. E. Tarof, and J. C. H. Yu, "Temperature dependence of breakdown voltages in separate absorption, grading, charge, and multiplication InP/InGaAs avalanche photodiodes," *IEEE Trans. Electron Devices*, vol. 42, no. 5, pp. 810–818, May 1995, doi: 10.1109/16.381974.

[38] D. A. Neamen, "Semiconductor physics and devices," *Mater. Today*, vol. 9, no. 5, p. 57, May 2006, doi: 10.1016/S1369-7021(06)71498-5.

[39] W. Jiang and M. J. Deen, "Random Telegraph Signal in n + /p-Well CMOS Single-Photon Avalanche Diodes," *IEEE Trans. Electron Devices*, vol. 68, no. 6, pp. 2764–2769, Jun. 2021, doi: 10.1109/TED.2021.3070557.

[40] D. P. 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. 409–426, Nov. 2014, doi: 10.1109/JSTQE.2014.2344034.

[41] 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, Oct. 2011, doi: 10.1109/JSEN.2011.2123090.

[42] L. Parmesan, N. A. W. Dutton, N. J. Calder, L. A. Grant, and R. K. Henderson, "A 256x256 SPAD array with in-pixel Time to Amplitude Conversion for Fluorescence Lifetime Imaging Microscopy," *Int. Image Sens. Work.*, pp. 8–11, 2015, [Online]. Available: http://www.imagesensors.org/Past Workshops/2015 Workshop/2015

Papers/Sessions/Session_9/9-04_Parmesan.pdf.

[43] P. Keranen, K. Maatta, and J. Kostamovaara, "Wide-Range Time-to-Digital Converter With 1-ps Single-Shot Precision," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 9, pp. 3162–3172, Sep. 2011, doi: 10.1109/TIM.2011.2122510.

[44] B. K. Swann *et al.*, "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, Nov. 2004, doi: 10.1109/JSSC.2004.835832.

[45] K. Maatta 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, Apr. 1998, doi: 10.1109/19.744201.

[46] E. Raisanen-Ruotsalainen, T. Rahkonen, and J. Kostamovaara, "A high resolution time-to-digital converter based on time-to-voltage interpolation," *Proc. 23rd Eur. Solid-State Circuits Conf.*, pp. 2–5, 1997, doi: 10.1109/ESSCIR.1997.186174.

[47] J. Kalisz, R. Pelka, and A. Poniecki, "Precision time counter for laser ranging to satellites," *Rev. Sci. Instrum.*, vol. 65, no. 3, pp. 736–741, Mar. 1994, doi: 10.1063/1.1145094.

[48] J. Kostamovaara and R. Myllylä, "Time-to-digital converter with an analog interpolation circuit," *Rev. Sci. Instrum.*, vol. 57, no. 11, pp. 2880–2885, Nov. 1986, doi: 10.1063/1.1139008.

[49] E. Raisanen-Ruotsalainen, T. Rahkonen, and J. Kostamovaara, "A BiCMOS Time-to-Digital Converter with Time Stretching Interpolators," *Esscirc*, no. October 1996, pp. 428–431, 1996.

[50] R. Granja, M. Santos, J. Guilherme, and N. Horta, "11.7b Time-To-Digital Converter with 0.82ps resolution in 130nm CMOS Technology," in *2018 14th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)*, Jul. 2018, pp. 29–32, doi: 10.1109/PRIME.2018.8430374.

[51] T. Iizuka, S. Miura, R. Yamamoto, Y. Chiba, S. Kubo, and K. Asada, "A 580 fs-resolution time-to-digital converter utilizing differential pulse-shrinking buffer ring in 0.18 μm CMOS technology," *IEICE Trans. Electron.*, vol. E95-C, no. 4, pp. 661–667, 2012, doi: 10.1587/transele.E95.C.661.

[52] J.-C. C. Lai and T.-Y. Y. Hsu, "Cost-Effective Time-to-Digital Converter Using Time-Residue Feedback," *IEEE Trans. Ind. Electron.*, vol. 64, no. 6, pp. 4690–4700, Jun. 2017, doi: 10.1109/TIE.2017.2669883.

[53] P. Keranen and J. Kostamovaara, "A Wide Range, 4.2 ps(rms) Precision CMOS TDC With Cyclic Interpolators Based on Switched-Frequency Ring Oscillators," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 62, no. 12, pp. 2795–2805, Dec. 2015, doi: 10.1109/TCSI.2015.2485719.

[54] Y. H. Seo, J. S. Kim, H. J. Park, and J. Y. Sim, "A 0.63ps resolution, 11b pipeline TDC in 0.13μm CMOS," *IEEE Symp. VLSI Circuits, Dig. Tech. Pap.*, pp. 152–153, 2011.

[55] 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 Regul. Pap.*, vol. 60, no. 3, pp. 557–569, Mar. 2013, doi: 10.1109/TCSI.2012.2215737.

[56] N. Roy, F. Nolet, F. Dubois, M.-O. Mercier, R. Fontaine, and J.-F. Pratte, "Low Power and Small Area, 6.9 ps RMS Time-to-Digital Converter for 3-D Digital SiPM," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 1, no. 6, pp. 486–494, Nov. 2017, doi:

10.1109/TRPMS.2017.2757444.

[57] R. B. Staszewski, S. Vemulapalli, P. Vallur, J. Wallberg, and P. T. Balsara, "1.3 V 20 ps time-to-digital converter for frequency synthesis in 90-nm CMOS," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 53, no. 3, pp. 220–224, Mar. 2006, doi: 10.1109/TCSII.2005.858754.

[58] 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, doi: 10.1016/j.nima.2006.01.071.

[59] M. Zanuso, P. Madoglio, S. Levantino, C. Samori, and A. L. A. L. Lacaita, "Time-to-Digital Converter for Frequency Synthesis Based on a Digital Bang-Bang DLL," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 57, no. 3, pp. 548–555, Mar. 2010, doi: 10.1109/TCSI.2009.2023945.

[60] M. Kanoun, M. W. Ben Attouch, Y. Bérubé-Lauzière, and R. Fontaine, "A 10-Bit, 12 ps Resolution CMOS Time-to-Digital Converter Dedicated to Ultra-Fast Optical Timing Applications," *Circuits, Syst. Signal Process.*, vol. 34, no. 4, pp. 1129–1148, 2015, doi: 10.1007/s00034-014-9901-7.

[61] 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. 887–894, 1993, doi: 10.1109/4.231325.

[62] B. Razavi, "The Delay-Locked Loop [A Circuit for All Seasons]," *IEEE Solid-State Circuits Mag.*, vol. 10, no. 3, pp. 9–15, 2018, doi: 10.1109/MSSC.2018.2844615.

[63] M. Moazedi, A. Abrishamifar, and A. M. Sodagar, "A highly-linear modified pseudo-differential current starved delay element with wide tuning range," 2011.

[64] J. G. Maneatis, "Low-Jitter Process-Independent DLL and PLL Based on Self-Biased Techniques," in *Phase-Locking in High-Performance Systems*, vol. 31, no. 11, IEEE, 2009, pp. 396–405.

[65] D. M. Santos, S. F. Dow, J. M. Flasck, and M. E. Levi, "A CMOS delay locked loop and sub-nanosecond time-to-digital converter chip," *IEEE Trans. Nucl. Sci.*, vol. 43, no. 3, pp. 1717–1719, Jun. 1996, doi: 10.1109/23.507177.

[66] P. Dudek, S. Szczepanski, J. V. J. V. Hatfield, S. Szczepański, and J. V. 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, Feb. 2000, doi: 10.1109/4.823449.

[67] J. Kong, S. Henzler, D. Schmitt-Landsiedel, and L. Siek, "A 9-bit, 1.08ps resolution two-step time-to-digital converter in 65 nm CMOS for time-mode ADC," in *2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Oct. 2016, pp. 348–351, doi: 10.1109/APCCAS.2016.7803972.

[68] N. U. Andersson and M. Vesterbacka, "A Vernier Time-to-Digital Converter With Delay Latch Chain Architecture," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 61, no. 10, pp. 773–777, Oct. 2014, doi: 10.1109/TCSII.2014.2345289.

[69] V. Ramakrishnan and P. T. Balsara, "A wide-range, high-resolution, compact, CMOS time to digital converter," in *19th International Conference on VLSI Design held jointly with 5th International Conference on Embedded Systems Design (VLSID'06)*, 2006, vol. 2006, p. 6 pp., doi: 10.1109/VLSID.2006.28.

[70] Y. J. Park and F. Yuan, "0.25-4 ns 185 MS/s 4-bit pulse-shrinking time-to-digital converter

in 130 nm CMOS using a 2-step conversion scheme," in *2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2015, pp. 1–4, doi: 10.1109/MWSCAS.2015.7282113.

[71] Y. J. Park and F. Yuan, "A 12.88 MS/s 0.28 pJ/conv.step 8-bit stage-interleaved pulse-shrinking time-to-digital converter in 130 nm CMOS," in *2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2015, pp. 1–4, doi: 10.1109/MWSCAS.2015.7282114.

[72] C.-H. Wu, S.-Y. Huang, M. Chern, Y.-F. Chou, and D.-M. Kwai, "Resilient Cell-Based Architecture for Time-to-Digital Converter," in *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, Jul. 2017, pp. 7–12, doi: 10.1109/ISVLSI.2017.12.

[73] Y.-C. Chang, S.-Y. Huang, C.-W. Tzeng, and J. Yao, "A fully cell-based design for timing measurement of memory," in *2011 IEEE International Test Conference*, Sep. 2011, pp. 1–10, doi: 10.1109/TEST.2011.6139150.

[74] E. Raisanen-Ruotsalainen, T. Rahkonen, and J. Kostamovaara, "A low-power CMOS time-to-digital converter," *IEEE J. Solid-State Circuits*, vol. 30, no. 9, pp. 984–990, 1995, doi: 10.1109/4.406397.

[75] R. Enomoto, T. Iizuka, T. Koga, T. Nakura, and K. Asada, "A 16-bit 2.0-ps resolution two-step TDC in 0.18-μ m CMOS utilizing pulse-shrinking fine stage with built-in coarse gain calibration," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 27, no. 1, pp. 11–19, Jan. 2019, doi: 10.1109/TVLSI.2018.2867505.

[76] T. Iizuka, T. Koga, T. Nakura, and K. Asada, "A fine-resolution pulse-shrinking time-to-digital converter with completion detection utilizing built-in offset pulse," in *2016 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Nov. 2016, pp. 313–316, doi: 10.1109/ASSCC.2016.7844198.

[77] F. Yuan, "CMOS time-to-digital converters for mixed-mode signal processing," *J. Eng.*, vol. 2014, no. 4, pp. 140–154, Apr. 2014, doi: 10.1049/joe.2014.0044.

[78] K. Okuno, T. Konishi, S. Izumi, M. Yoshimoto, and H. Kawaguchi, "A 62-dB SNDR second-order gated ring oscillator TDC with two-stage dynamic D-type flipflops as a quantization noise propagator," in *10th IEEE International NEWCAS Conference*, Jun. 2012, pp. 289–292, doi: 10.1109/NEWCAS.2012.6329013.

[79] K.-C. C. Choi, S.-W. W. Lee, B.-C. C. Lee, and W.-Y. Y. 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 Express Briefs*, vol. 59, no. 3, pp. 143–147, Mar. 2012, doi: 10.1109/TCSII.2012.2184370.

[80] M. Perenzoni, H. Xu, and D. Stoppa, "Small area 0.3 pJ/conv, 45 ps time-to-digital converter for arrays of silicon photomultiplier interfaces in 150 nm CMOS," *Electron. Lett.*, vol. 51, no. 23, pp. 1933–1935, Nov. 2015, doi: 10.1049/el.2015.2761.

[81] B. Wu, S. Zhu, Y. Zhou, and Y. Chiu, "A 9-bit 215 MS/s Folding-Flash Time-to-Digital Converter Based on Redundant Remainder Number System in 45-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 53, no. 3, pp. 839–849, Mar. 2018, doi: 10.1109/JSSC.2017.2782766.

[82] M. Kim, K.-S. Son, N. Kim, C. H. Rho, and J.-K. Kang, "A Two-Step Time-to-Digital Converter using Ring Oscillator Time Amplifier," in *2018 International SoC Design Conference (ISOCC)*, Nov. 2018, pp. 143–144, doi: 10.1109/ISOCC.2018.8649906.

[83] J. P. Caram, J. Galloway, and J. S. Kenney, "Time-to-Digital Converter With Sample-and-

Hold and Quantization Noise Scrambling Using Harmonics in Ring Oscillators," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 65, no. 1, pp. 74–83, Jan. 2018, doi: 10.1109/TCSI.2017.2712518.

[84] S.-H. Chung, K.-D. Hwang, W.-Y. Lee, and L.-S. Kim, "A high resolution metastability-independent two-step gated ring oscillator TDC with enhanced noise shaping," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, May 2010, pp. 1300–1303, doi: 10.1109/ISCAS.2010.5537261.

[85] K.-D. Hwang and L.-S. Kim, "An area efficient asynchronous gated ring oscillator TDC with minimum GRO stages," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, May 2010, pp. 3973–3976, doi: 10.1109/ISCAS.2010.5537663.

[86] J. Yu, F. F. Dai, and R. C. Jaeger, "A 12-Bit vernier ring time-to-digital converter in 0.13 μm CMOS technology," in *IEEE Journal of Solid-State Circuits*, Apr. 2010, vol. 45, no. 4, pp. 830–842, doi: 10.1109/JSSC.2010.2040306.

[87] N. Xing, J.-K. Woo, W.-Y. Shin, H. Lee, and S. Kim, "A 14.6 ps Resolution, 50 ns Input-Range Cyclic Time-to-Digital Converter Using Fractional Difference Conversion Method," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 57, no. 12, pp. 3064–3072, Dec. 2010, doi: 10.1109/TCSI.2010.2073810.

[88] Ping Lu, P. Andreani, and A. Liscidini, "A 90nm CMOS gated-ring-oscillator-based 2-dimension Vernier time-to-digital converter," in *NORCHIP 2012*, Nov. 2012, pp. 1–4, doi: 10.1109/NORCHP.2012.6403120.

[89] 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, doi: 10.1063/1.4733705.

[90] 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, Aug. 2013, doi: 10.1007/s10470-013-0084-0.

[91] H. Park, Z.-Z. Z. Yu, J. Kim, and J. Burm, "Resolution tunable ring oscillator type TDC," in *2016 International SoC Design Conference (ISOCC)*, Oct. 2016, pp. 241–242, doi: 10.1109/ISOCC.2016.7799767.

[92] H. Wang and F. F. Dai, "A 14-Bit, 1-ps resolution, two-step ring and 2D Vernier TDC in 130nm CMOS technology," in *ESSCIRC 2017 - 43rd IEEE European Solid State Circuits Conference*, Sep. 2017, pp. 143–146, doi: 10.1109/ESSCIRC.2017.8094546.

[93] V. Nguyen, D. Duong, Y. Chung, and J.-W. Lee, "A Cyclic Vernier Two-Step TDC for High Input Range Time-of-Flight Sensor Using Startup Time Correction Technique," *Sensors*, vol. 18, no. 11, p. 3948, Nov. 2018, doi: 10.3390/s18113948.

[94] V. Sesta, F. Villa, E. Conca, and A. Tosi, "A novel sub-10 ps resolution TDC for CMOS SPAD array," in *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Dec. 2018, pp. 5–8, doi: 10.1109/ICECS.2018.8617859.

[95] A. Annagrebah, E. Bechetoille, I. B. Laktineh, H. Chanal, P. Russo, and H. Mathez, "A Multi-phase Time-to-Digital Converter Differential Vernier Ring Oscillator," in *2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, Jul. 2019, vol. 2019-July, pp. 344–347, doi: 10.1109/ISVLSI.2019.00069.

[96] S. Tisa, A. Lotito, A. Giudice, and F. Zappa, "Monolithic time-to-digital converter with 20ps resolution," *Eur. Solid-State Circuits Conf.*, pp. 465–468, 2003, doi:

10.1109/ESSCIRC.2003.1257173.

[97] Y. Liu *et al.*, "Multi-stage Pulse Shrinking Time-to-Digital Converter for Time Interval Measurements," in *2007 European Conference on Wireless Technologies*, Oct. 2007, pp. 347–350, doi: 10.1109/ECWT.2007.4404018.

[98] Y. Liu *et al.*, "A 6ps resolution pulse shrinking Time-to-Digital Converter as phase detector in multi-mode transceiver," *2008 IEEE Radio Wirel. Symp. RWS*, pp. 163–166, 2008, doi: 10.1109/RWS.2008.4463454.

[99] C. C. Chen, S. H. Lin, and C. S. Hwang, "An area-efficient CMOS time-to-digital converter based on a pulse-shrinking scheme," *IEEE Trans. Circuits Syst. II Express Briefs*, 2014, doi: 10.1109/TCSII.2013.2296192.

[100] S. S. Mohan, W. S. Chan, D. M. Colleran, S. F. Greenwood, J. E. Gamble, and I. G. Kouznetsov, "Differential ring oscillators with multipath delay stages," in *Proceedings of the IEEE 2005 Custom Integrated Circuits Conference, 2005.*, 2005, vol. 2005, no. c, pp. 498–501, doi: 10.1109/CICC.2005.1568716.

[101] 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, doi: 10.1109/JSSC.2009.2014709.

[102] M. Z. Straayer and M. H. Perrott, "An efficient high-resolution 11-bit noise-shaping multipath gated ring oscillator TDC," in *2008 IEEE Symposium on VLSI Circuits*, Jun. 2008, pp. 82–83, doi: 10.1109/VLSIC.2008.4585960.

[103] 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, p. 035004, Mar. 2013, doi: 10.1088/1674-4926/34/3/035004.

[104] S. Ziabakhsh, G. Gagnon, and G. W. Roberts, "A Second-Order Bandpass ΔΣ Time-to-Digital Converter With Negative Time-Mode Feedback," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 66, no. 4, pp. 1355–1368, Apr. 2019, doi: 10.1109/TCSI.2018.2882892.

[105] S. T. Chandrasekaran, A. Jayaraj, M. Danesh, and A. Sanyal, "A highly digital second-order oversampling TDC," *IEEE Solid-State Circuits Lett.*, vol. 1, no. 5, pp. 114–117, 2018, doi: 10.1109/LSSC.2018.2875818.

[106] W. Yu, K. Kim, and S. Cho, "A 0.22 ps rms Integrated Noise 15 MHz Bandwidth Fourth-Order ΔΣ Time-to-Digital Converter Using Time-Domain Error-Feedback Filter," *IEEE J. Solid-State Circuits*, vol. 50, no. 5, pp. 1251–1262, May 2015, doi: 10.1109/JSSC.2015.2399673.

[107] Y. Cao, W. De Cock, M. Steyaert, and P. Leroux, "Design and Assessment of a 6 ps-Resolution Time-to-Digital Converter With 5 MGy Gamma-Dose Tolerance for LIDAR Application," *IEEE Trans. Nucl. Sci.*, vol. 59, no. 4, pp. 1382–1389, Aug. 2012, doi: 10.1109/TNS.2012.2193598.

[108] M. Gande, N. Maghari, T. Oh, and U.-K. Moon, "A 71dB dynamic range third-order ΔΣ TDC using charge-pump," in *2012 Symposium on VLSI Circuits (VLSIC)*, Jun. 2012, pp. 168–169, doi: 10.1109/VLSIC.2012.6243843.

[109] J. P. Hong *et al.*, "A 0.004mm 2 250μW ΔΣ TDC with time-difference accumulator and a 0.012mm 2 2.5mW bang-bang digital PLL using PRNG for low-power SoC applications," *Dig. Tech. Pap. - IEEE Int. Solid-State Circuits Conf.*, vol. 55, pp. 240–241, 2012, doi: 10.1109/ISSCC.2012.6176992.

[110] D. Kim, K. Kim, W. Yu, and S. Cho, "A Second-Order ΔΣ Time-to-Digital Converter Using

Highly Digital Time-Domain Arithmetic Circuits," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 66, no. 10, pp. 1643–1647, Oct. 2019, doi: 10.1109/TCSII.2019.2925860.

[111] C.-K. K. Kwon, H. Kim, J. Park, and S.-W. W. Kim, "A 0.4-mW, 4.7-ps resolution single-loop ΔΣ TDC using a half-delay time integrator," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 24, no. 3, pp. 1184–1188, Mar. 2016, doi: 10.1109/TVLSI.2015.2438851.

[112] K. Kim, W. Yu, and S. Cho, "A 9 bit, 1.12 ps resolution 2.5 b/stage pipelined time-to-digital converter in 65 nm CMOS using time-register," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 1007–1016, Apr. 2014, doi: 10.1109/JSSC.2013.2297412.

[113] H. Molaei and K. Hajsadeghi, "A 5.3-ps, 8-b Time to Digital Converter Using a New Gain-Reconfigurable Time Amplifier," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 66, no. 3, pp. 352–356, Mar. 2019, doi: 10.1109/TCSII.2018.2853187.

[114] P. Lu, A. Liscidini, and P. Andreani, "A 3.6 mW, 90 nm 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, Jul. 2012, doi: 10.1109/JSSC.2012.2191676.

[115] K. Kim, Y.-H. H. Kim, W. Yu, and S. Cho, "A 7 bit, 3.75 ps 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. 1009–1017, Apr. 2013, doi: 10.1109/JSSC.2013.2237996.

[116] P. Lu, Y. Wu, and P. Andreani, "A 2.2-ps Two-Dimensional Gated-Vernier Time-to-Digital Converter With Digital Calibration," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 63, no. 11, pp. 1019–1023, Nov. 2016, doi: 10.1109/TCSII.2016.2548218.

[117] 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, doi: 10.1109/TBCAS.2015.2434957.

[118] H. H. Wang, F. F. Dai, and H. H. Wang, "A Reconfigurable Vernier Time-to-Digital Converter with 2-D Spiral Comparator Array and Second-Order Δ Σ Linearization," *IEEE J. Solid-State Circuits*, vol. 53, no. 3, pp. 738–749, Mar. 2018, doi: 10.1109/JSSC.2017.2788872.

[119] J. Kalisz, "Review of methods for time interval measurements with picosecond resolution," *Metrologia*, vol. 41, no. 1, pp. 17–32, Feb. 2004, doi: 10.1088/0026-1394/41/1/004.

[120] R. Machado, J. Cabral, and F. S. Alves, "Recent Developments and Challenges in FPGA-Based Time-to-Digital Converters," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 11, pp. 4205–4221, Nov. 2019, doi: 10.1109/TIM.2019.2938436.

[121] Seog-Jun Lee, Beomsup Kim, and Kwyro Lee, "A novel high-speed ring oscillator for multiphase clock generation using negative skewed delay scheme," *IEEE J. Solid-State Circuits*, vol. 32, no. 2, pp. 289–291, 1997, doi: 10.1109/4.551926.

[122] M. Lee and A. A. Abidi, "A 9 b, 1.25 ps resolution coarse-fine time-to-digital converter in 90 nm CMOS that amplifies a time residue," 2008, doi: 10.1109/JSSC.2008.917405.

[123] Z. Li, M. Deen, S. Kumar, and P. Selvaganapathy, "Raman Spectroscopy for In-Line Water Quality Monitoring—Instrumentation and Potential," *Sensors*, vol. 14, no. 9, pp. 17275–17303, Sep. 2014, doi: 10.3390/s140917275.

[124] A. Torricelli *et al.*, "Time domain functional NIRS imaging for human brain mapping," *Neuroimage*, vol. 85, pp. 28–50, Jan. 2014, doi: 10.1016/j.neuroimage.2013.05.106.

[125] E. Slenders *et al.*, "Confocal-based fluorescence fluctuation spectroscopy with a SPAD array detector," *Light Sci. Appl.*, vol. 10, no. 1, p. 31, Feb. 2021, doi: 10.1038/s41377-021-

00475-z.

[126] A. C. Ulku *et al.*, "A 512 × 512 SPAD Image Sensor With Integrated Gating for Widefield FLIM," *IEEE J. Sel. Top. Quantum Electron.*, vol. 25, no. 1, pp. 1–12, Jan. 2019, doi: 10.1109/JSTQE.2018.2867439.

[127] N. Hirmiz, A. Tsikouras, E. J. Osterlund, M. Richards, D. W. Andrews, and Q. Fang, "Highly Multiplexed Confocal Fluorescence Lifetime Microscope Designed for Screening Applications," *IEEE J. Sel. Top. Quantum Electron.*, vol. 27, no. 5, pp. 1–9, Sep. 2021, doi: 10.1109/JSTQE.2020.2997834.

[128] S. P. Poland *et al.*, "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, doi: 10.1364/BOE.6.000277.

[129] C. Veerappan *et al.*, "A 160x128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter," in *2011 IEEE International Solid-State Circuits Conference*, Feb. 2011, pp. 312–314, doi: 10.1109/ISSCC.2011.5746333.

[130] S. Mandai, V. Jain, and E. Charbon, "A 780 × 800 μm2 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, Feb. 2014, doi: 10.1109/TNS.2013.2294022.

[131] A. Carimatto *et al.*, "11.4 A 67,392-SPAD PVTB-compensated multi-channel digital SiPM with 432 column-parallel 48ps 17b TDCs for endoscopic time-of-flight PET," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, Feb. 2015, vol. 58, pp. 1–3, doi: 10.1109/ISSCC.2015.7062996.

[132] A. Carimatto *et al.*, "Multipurpose, Fully Integrated 128×128 Event-Driven MD-SiPM With 512 16-Bit TDCs With 45-ps LSB and 20-ns Gating in 40-nm CMOS Technology," *IEEE Solid-State Circuits Lett.*, vol. 1, no. 12, pp. 241–244, Dec. 2018, doi: 10.1109/lssc.2019.2911043.

[133] E. Manuzzato *et al.*, "A 16×8 Digital-SiPM Array With Distributed Trigger Generator for Low SNR Particle Tracking," *IEEE Solid-State Circuits Lett.*, vol. 2, no. 9, pp. 75–78, Sep. 2019, doi: 10.1109/LSSC.2019.2934598.

[134] R. K. Henderson *et al.*, "A 192×128 Time Correlated SPAD Image Sensor in 40-nm CMOS Technology," *IEEE J. Solid-State Circuits*, vol. 54, no. 7, pp. 1907–1916, Jul. 2019, doi: 10.1109/JSSC.2019.2905163.

[135] E. Conca *et al.*, "Large-Area, Fast-Gated Digital SiPM With Integrated TDC for Portable and Wearable Time-Domain NIRS," *IEEE J. Solid-State Circuits*, vol. 55, no. 11, pp. 3097–3111, Nov. 2020, doi: 10.1109/JSSC.2020.3006442.

[136] 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, Dec. 2008, doi: 10.1109/JSSC.2008.2006445.

[137] J. Richardson *et al.*, "A 32×32 50ps resolution 10 bit time to digital converter array in 130nm CMOS for time correlated imaging," in *2009 IEEE Custom Integrated Circuits Conference*, Sep. 2009, no. 029217, pp. 77–80, doi: 10.1109/CICC.2009.5280890.

[138] A. Muntean *et al.*, "Blumino: the first fully integrated analog SiPM with on-chip time conversion," *IEEE Trans. Radiat. Plasma Med. Sci.*, pp. 1–1, 2020, doi: 10.1109/TRPMS.2020.3045081.

[139] G. Acconcia, S. Farina, I. G. Labanca, M. Ghioni, and I. Rech, "Fast and compact time-

correlated single photon counting system for high-speed measurement with low distortion," in *Single Molecule Spectroscopy and Superresolution Imaging XIII*, Feb. 2020, vol. 1124608, no. February 2020, p. 7, doi: 10.1117/12.2546291.

[140] M. Crotti, I. Rech, and M. Ghioni, "Four Channel, 40 ps Resolution, Fully Integrated Time-to-Amplitude Converter for Time-Resolved Photon Counting," *IEEE J. Solid-State Circuits*, vol. 47, no. 3, pp. 699–708, Mar. 2012, doi: 10.1109/JSSC.2011.2176161.

[141] P. Peronio, G. Acconcia, I. Rech, and M. Ghioni, "Improving the counting efficiency in time-correlated single photon counting experiments by dead-time optimization," *Rev. Sci. Instrum.*, vol. 86, no. 11, p. 113101, Nov. 2015, doi: 10.1063/1.4934812.

[142] G. Acconcia, S. Farina, I. Labanca, M. Ghioni, and I. Rech, "Beyond pile-up limitation in time-correlated single photon counting measurement with high-speed and low-distortion electronics," in *Advanced Photon Counting Techniques XIV*, May 2020, no. May 2020, p. 26, doi: 10.1117/12.2557999.

[143] T. Al Abbas *et al.*, "Sensor SoC for Microendoscopy," *2019 Symp. VLSI Circuits*, pp. 6–7, 2017.

[144] S. W. Hutchings *et al.*, "A Reconfigurable 3-D-Stacked SPAD Imager With In-Pixel Histogramming for Flash LIDAR or High-Speed Time-of-Flight Imaging," *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 2947–2956, Nov. 2019, doi: 10.1109/JSSC.2019.2939083.

[145] A. Ronchini Ximenes, P. Padmanabhan, M.-J. Lee, Y. Yamashita, D.-N. Yaung, and E. Charbon, "A Modular, Direct Time-of-Flight Depth Sensor in 45/65-nm 3-D-Stacked CMOS Technology," *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 3203–3214, Nov. 2019, doi: 10.1109/JSSC.2019.2938412.

[146] C. Niclass, M. Sergio, and E. Charbon, "A single photon avalanche diode array fabricated in 0.35-µm CMOS and based on an event-driven readout for TCSPC experiments," in *Advanced Photon Counting Techniques*, Oct. 2006, vol. 6372, no. November 2006, p. 63720S, doi: 10.1117/12.685974.

[147] A. Cominelli, G. Acconcia, P. Peronio, I. Rech, and M. Ghioni, "Readout Architectures for High Efficiency in Time-Correlated Single Photon Counting Experiments—Analysis and Review," *IEEE Photonics J.*, vol. 9, no. 3, pp. 1–15, Jun. 2017, doi: 10.1109/JPHOT.2017.2695519.

[148] G. Acconcia, A. Cominelli, I. Rech, and M. Ghioni, "High-efficiency integrated readout circuit for single photon avalanche diode arrays in fluorescence lifetime imaging," *Rev. Sci. Instrum.*, vol. 87, no. 11, p. 113110, Nov. 2016, doi: 10.1063/1.4968199.

[149] M. Maymandi-Nejad and M. Sachdev, "A monotonic digitally controlled delay element," *IEEE J. Solid-State Circuits*, vol. 40, no. 11, pp. 2212–2219, Nov. 2005, doi: 10.1109/JSSC.2005.857370.

[150] R. J. Baker, *CMOS Circuit Design, Layout, and Simulation*, Third Edit. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2010.

[151] S. Tancock, E. Arabul, and N. Dahnoun, "A Review of New Time-to-Digital Conversion Techniques," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 3406–3417, Oct. 2019, doi: 10.1109/TIM.2019.2936717.

[152] S. Naghavi *et al.*, "A Simple and Efficient Charge Injection Error Compensation Structure for MOS Sampling Switches," *J. Circuits, Syst. Comput.*, vol. 27, no. 08, p. 1850130, Jul. 2018, doi: 10.1142/S021812661850130X.

[153] Y. Maruyama, J. Blacksberg, and E. Charbon, "A 1024 x 8, 700-ps Time-Gated SPAD Line Sensor for Planetary Surface Exploration With Laser Raman Spectroscopy and LIBS," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 179–189, Jan. 2014, doi: 10.1109/JSSC.2013.2282091.

[154] I. Nissinen, J. Nissinen, P. Keranen, A.-K. Lansman, J. Holma, and J. Kostamovaara, "A 2 x (4) x 128 Multitime-Gated SPAD Line Detector for Pulsed Raman Spectroscopy," *IEEE Sens. J.*, vol. 15, no. 3, pp. 1358–1365, Mar. 2015, doi: 10.1109/JSEN.2014.2361610.

[155] F. Villa, R. Lussana, D. Tamborini, A. Tosi, and F. Zappa, "High-Fill-Factor 60 x 1 SPAD Array With 60 Subnanosecond Integrated TDCs," *IEEE Photonics Technol. Lett.*, vol. 27, no. 12, pp. 1261–1264, 2015, doi: 10.1109/LPT.2015.2416192.

[156] D. Perenzoni, M., Massara,N., Perenzoni,D., Gasparini,L., Stoppa, "A 160 x 120 Pixel Analog-Counting Single-Photon Imager With Time-Gating and Self-Referenced Column-Parallel A/D Conversion for Fluorescence Lifetime Imaging," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 155–167, Jan. 2016, doi: 10.1109/JSSC.2015.2482497.

[157] I. Nissinen, J. Nissinen, P. Keranen, D. Stoppa, and J. Kostamovaara, "A 16×256 SPAD Line Detector With a 50-ps, 3-bit, 256-Channel Time-to-Digital Converter for Raman Spectroscopy," *IEEE Sens. J.*, vol. 18, no. 9, pp. 3789–3798, May 2018, doi: 10.1109/JSEN.2018.2813531.

[158] D. Portaluppi, E. Conca, and F. Villa, "32 × 32 CMOS SPAD Imager for Gated Imaging, Photon Timing, and Photon Coincidence," *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 2, pp. 1–6, Mar. 2018, doi: 10.1109/JSTQE.2017.2754587.

[159] E. Conca, I. Cusini, F. Severini, R. Lussana, F. Zappa, and F. Villa, "Gated SPAD Arrays for Single-Photon Time-Resolved Imaging and Spectroscopy," *IEEE Photonics J.*, vol. 11, no. 6, pp. 1–10, Dec. 2019, doi: 10.1109/JPHOT.2019.2952670.

[160] H. Ruokamo, L. W. Hallman, and J. Kostamovaara, "An 80 x 25 Pixel CMOS Single-Photon Sensor With Flexible On-Chip Time Gating of 40 Subarrays for Solid-State 3-D Range Imaging," *IEEE J. Solid-State Circuits*, vol. 54, no. 2, pp. 501–510, Feb. 2019, doi: 10.1109/JSSC.2018.2878816.

[161] N. Faramarzpour, M. J. Deen, S. Shirani, and Q. Fang, "Fully Integrated Single Photon Avalanche Diode Detector in Standard CMOS 0.18-um Technology," *IEEE Trans. Electron Devices*, vol. 55, no. 3, pp. 760–767, Mar. 2008, doi: 10.1109/TED.2007.914839.

[162] W. Jiang, Y. Chalich, R. Scott, and M. J. Deen, "Time-Gated and Multi-Junction SPADs in Standard 65 nm CMOS Technology," *IEEE Sens. J.*, pp. 1–1, 2021, doi: 10.1109/JSEN.2021.3063319.

[163] K. Morimoto and E. Charbon, "High fill-factor miniaturized SPAD arrays with a guard-ring-sharing technique," *Opt. Express*, vol. 28, no. 9, p. 13068, Apr. 2020, doi: 10.1364/OE.389216.

[164] F. Zappa, A. Tosi, A. D. Mora, and S. Tisa, "SPICE modeling of single photon avalanche diodes," *Sensors Actuators A Phys.*, vol. 153, no. 2, pp. 197–204, Aug. 2009, doi: 10.1016/j.sna.2009.05.007.

[165] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

[166] F. Nolet *et al.*, "Quenching Circuit and SPAD Integrated in CMOS 65 nm with 7.8 ps FWHM Single Photon Timing Resolution," *Instruments*, vol. 2, no. 4, p. 19, Sep. 2018, doi: 10.3390/instruments2040019.

[167] 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, doi: 10.1109/TED.2015.2475126.

[168] M. J. Deen, J. Ilowski, and P. Yang, "Low frequency noise in polysilicon-emitter bipolar junction transistors," *J. Appl. Phys.*, vol. 77, no. 12, pp. 6278–6288, Jun. 1995, doi: 10.1063/1.359095.

[169] M. J. Deen, S. L. Rumyantsev, and M. Schroter, "On the origin of 1/f noise in polysilicon emitter bipolar transistors," *J. Appl. Phys.*, 1999, doi: 10.1063/1.369256.

[170] P. K. B. Deen, M.J., *Silicon photonics*, vol. 43, no. 1. 2006.

[171] M.-J. J. Lee *et al.*, "Effects of Guard-Ring Structures on the Performance of Silicon Avalanche Photodetectors Fabricated With Standard CMOS Technology," *IEEE Electron Device Lett.*, vol. 33, no. 1, pp. 80–82, Jan. 2012, doi: 10.1109/LED.2011.2172390.

[172] D. Shin, B. Park, Y. Chae, and I. Yun, "The Effect of a Deep Virtual Guard Ring on the Device Characteristics of Silicon Single Photon Avalanche Diodes," *IEEE Trans. Electron Devices*, 2019, doi: 10.1109/TED.2019.2913714.

[173] J. Rhim *et al.*, "Guard-ring dependence of noise characteristics for single-photon avalanche diodes in a standard CMOS technology," in *2017 IEEE 14th International Conference on Group IV Photonics (GFP)*, Aug. 2017, pp. 155–156, doi: 10.1109/GROUP4.2017.8082243.

[174] F. Arvani, T. C. Carusone, and E. S. Rogers, "TDC sharing in SPAD-based direct time-of-flight 3D imaging applications," in *Proceedings - IEEE International Symposium on Circuits and Systems*, May 2019, vol. 2019-May, pp. 1–5, doi: 10.1109/ISCAS.2019.8702586.

[175] A. T. Erdogan, R. Walker, N. Finlayson, N. Krstajic, G. O. S. Williams, and R. K. Henderson, "A 16.5 giga events/s 1024 × 8 SPAD line sensor with per-pixel zoomable 50ps-6.4ns/bin histogramming TDC," in *2017 Symposium on VLSI Circuits*, Jun. 2017, pp. C292–C293, doi: 10.23919/VLSIC.2017.8008513.

[176] T. Al Abbas, N. A. W. Dutton, O. Almer, N. Finlayson, F. M. Della Rocca, and R. Henderson, "A CMOS SPAD Sensor With a Multi-Event Folded Flash Time-to-Digital Converter for Ultra-Fast Optical Transient Capture," *IEEE Sens. J.*, vol. 18, no. 8, pp. 3163–3173, Apr. 2018, doi: 10.1109/JSEN.2018.2803087.

[177] C. Zhang, S. Lindner, I. M. Antolovic, J. Mata Pavia, M. Wolf, and E. Charbon, "A 30-frames/s, 252×144 SPAD Flash LiDAR With 1728 Dual-Clock 48.8-ps TDCs, and Pixel-Wise Integrated Histogramming," *IEEE J. Solid-State Circuits*, vol. 54, no. 4, pp. 1137–1151, Apr. 2019, doi: 10.1109/JSSC.2018.2883720.

[178] F. Mattioli Della Rocca *et al.*, "A 128 × 128 SPAD Motion-Triggered Time-of-Flight Image Sensor With In-Pixel Histogram and Column-Parallel Vision Processor," *IEEE J. Solid-State Circuits*, vol. 55, no. 7, pp. 1762–1775, Jul. 2020, doi: 10.1109/JSSC.2020.2993722.

[179] R. Szplet, D. Sondej, and G. Grzeda, "Subpicosecond-resolution time-to-digital converter with multi-edge coding in independent coding lines," in *2014 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, May 2014, pp. 747–751, doi: 10.1109/I2MTC.2014.6860842.

[180] Q. Cao, Y. Wang, and C. Liu, "A combination of multiple channels of FPGA based time-to-digital converter for high time precision," in *2016 IEEE Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop*

(NSS/MIC/RTSD), Oct. 2016, vol. 2017-Janua, pp. 1–3, doi: 10.1109/NSSMIC.2016.8069649.

[181] R. Szplet, P. Kwiatkowski, Z. Jachna, and K. Rozyc, "An Eight-Channel 4.5-ps Precision Timestamps-Based Time Interval Counter in FPGA Chip," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 9, pp. 2088–2100, Sep. 2016, doi: 10.1109/TIM.2016.2564038.

[182] E. Arabul, A. Girach, J. Rarity, and N. Dahnoun, "Precise multi-channel timing analysis system for multi-stop LIDAR correlation," in *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, Oct. 2017, vol. 2018-Janua, pp. 1–6, doi: 10.1109/IST.2017.8261560.

[183] X. Qin, L. Wang, D. Liu, Y. Zhao, X. Rong, and J. Du, "A 1.15-ps Bin Size and 3.5-ps Single-Shot Precision Time-to-Digital Converter With On-Board Offset Correction in an FPGA," *IEEE Trans. Nucl. Sci.*, vol. 64, no. 12, pp. 2951–2957, Dec. 2017, doi: 10.1109/TNS.2017.2768082.

[184] H. Deng, Q. Fan, R. Zhang, and J. Chen, "Machine-Learning Based Nonlinerity Correction for Coarse-Fine SAR-TDC Hybrid ADC," in *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2020, pp. 265–268, doi: 10.1109/MWSCAS48704.2020.9184523.

[185] S. Li, X. Xu, and W. Burleson, "CCATDC: A Configurable Compact Algorithmic Time-to-Digital Converter," in *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, Jul. 2017, vol. 2017-July, pp. 501–506, doi: 10.1109/ISVLSI.2017.118.

[186] V. Zickus *et al.*, "Fluorescence lifetime imaging with a megapixel SPAD camera and neural network lifetime estimation," *Sci. Rep.*, vol. 10, no. 1, p. 20986, Dec. 2020, doi: 10.1038/s41598-020-77737-0.