**CMOS SINGLE PHOTON Avalanche diodes AND TIME-To-DIGITAL CONVERTERS FOR Time resolved fluorescence analysis**

**CMOS SINGLE PHOTON AVALANCHE DIODES AND TIME-TO-DIGITAL CONVERTERS FOR Time-Resolved Fluorescence analysis**

By

Dariusz Palubiak

B.Sc. McMaster University, 2005

M.S. McMaster University, 2008

A Thesis

Submitted to the School of Graduate

Studies in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

McMaster University

Hamilton, Ontario, Canada

© Copyright by Dariusz Palubiak, Dec. 2015

Doctor of Philosophy (2015) McMaster University

(Electrical and Computer Engineering) Hamilton, Ontario

TITLE: CMOS single-photon avalanche diodes and time-to-digital converters for time-resolved fluorescence analysis

AUTHOR: Dariusz Palubiak,

B. A. Sc. McMaster University, Hamilton, ON, Canada

M. A. Sc. McMaster University, Hamilton, ON, Canada

SUPERVISOR: Prof. M. Jamal Deen

NUMBER OF PAGES: 181

Abstract

Fluorescence lifetime imaging (FLIM) has the potential to provide rapid screening and detection of diseases. However, time-resolved fluorescence measurements require high-performance detectors with single-photon sensitivity and sub-nanosecond time resolution. These systems should also be compact, reliable, inexpensive, and easily deployable for laboratory and clinical applications. It is with these applications in mind that the development of single photon avalanche diodes (SPAD) and time-to-digital converter (TDC) prototype integrated circuits (IC) in standard digital CMOS have been pursued in this thesis.

SPAD and TDC ICs were designed and fabricated in 130 nm IBM CMOS technology and then intensively studied. Several different SPAD pixels were modeled and designed, and the electro-optical performance was characterized and comparatively studied. By repurposing existing design layers of a standard CMOS process, the fabricated SPAD pixel test structures achieved up to 20× improvement of dark count rate (DCR) compared to previous designs. Optical measurements also showed up to 10× improvement in the detection limits for low-level light. Detailed dark noise characterization was performed at various temperatures using free-running and time-gated modes of operation. Optimal operating conditions were found for minimal afterpulsing effects. The SPAD’s capability to accurately measure fast fluorescence decays was also demonstrated in a practical setting with the lifetime measurements of two fluorophores, Rhodamine 6G and Ruby crystal, which have fluorescence lifetimes of approximately 4 ns and 3 ms, respectively.

A fast and accurate TDC prototype circuit for time-correlated single-photon counting (TCSPC) applications was designed, fabricated and characterized. With a coarse-fine delay line architecture, the TDC size was reduced without compromising its linearity and jitter performance. Extensive characterization of the fabricated SPAD and TDC ICs shows that the measured performance met the stated design goals.

Acknowledgements

First and foremost, I’d like to express my deepest gratitude and appreciation to Distinguished University Professor Dr. M. J Deen for supporting my research over the years and for providing his continual encouragement and motivation. It has been an honor and privilege to be part of Prof. Deen’s research team, and I am very grateful for having been able to receive his guidance. Prof. Deen has been a great mentor and role model for me, and his enthusiastic support and insight has not only made this work possible, but has also made a great impact in my life.

Also, I would like to thank my committee members Prof. Shiva Kumar and Prof. Ravi Selvaganapathy for their advice I received during my committee meetings and for taking the time to review my thesis.

I would like to profoundly thank Dr. Ognian Marinov for his interest in my research and for the helpful discussions and valuable assistance offered during the course of my studies. His experience and knowledge of microelectronics has greatly aided me in obtaining the results presented in this thesis. I am also thankful to Dr. Q. Fang for granting me access to his lab and for his help and suggestions for my measurements. Also I would like to thank Dr. Peng, Dr. Nicolici, and Dr. Chen for their encouragement during my studies.

Next I would like to Cheryl Gies for all her kind support throughout my time as a graduate student at McMaster University and the technical staff of the ECE department of McMaster University, including Tyler Ackland, Terry Greenlay, Dan Manolescu, Ron Harwood and Steve Spencer.

My thanks also go to my colleagues that I have had the pleasure to work with during my graduate studies. I am greatly indebted to Dr. Munir El-Desouki who helped to get me started on this research project, Dr. Waleed Shinwari who always gave me his valuable advice and help, and Dr. Zhiyun Li who developed the time-gated circuit studied in this work. I would also like to thank all the members of Prof. Deen’s research group, including Dr. Mohammedreza Dadkhah, Hani Alhemsi, David Cheng, and Mrwan Alayed. Many thanks are also due to members of Dr. Fang’s research group, including Anthony Tsikauras, Dr. Hanna Budz, Eric Mahoney and Du Le.

I also thank the Canadian Microelectronics Corporation (CMC) for arranging the fabrication of the test chips.

Last and not least, I am grateful for the endless love, support, patience and encouragement that I have received from my parents and from my dear fiancée Mira. This thesis is dedicated to them.

List of Abbreviations

|  |  |
| --- | --- |
| 3T | 3 Transistor |
| 4T | 4 Transistor |
| AC | Alternating Current |
| ADC | Analog to Digital Converter |
| AP | Afterpulsing Probability |
| APD | Avalanche Photodiode |
| APS | Active Pixel Sensor |
| AQAR | Active Quench, Active Reset |
| AQR | Active Quench and Reset |
| BPF | Bandpass Filter |
| BV | Breakdown Voltage |
| CB | Conduction Band |
| CC | Coupling Capacitor |
| CS | Common Source |
| CCD | Charge-Coupled Device |
| CIS | CMOS Image Sensor |
| CFD | Constant Fraction Discriminator |
| CLK | Clock |
| CMC | Canadian Microelectronics Corporation |
| CMOS | Complementary Metal-Oxide-Semiconductor |
| CMP | Chemical-Mechanical Polishing |
| CP | Charge Pump |
| CV | Coefficient of Variation |
| CW | Continuous Wave |
| dB | Decibel |
| DC | Direct Current |
| DCR | Dark Count Rate |
| DG | Delay Generator |
| DLL | Delay Locked Loop |
| DNA | Deoxyribonucleic Acid |
| DNL | Differential Non-Linearity |
| DNW | Deep N-Well |
| DRC | Design Rule Checking |
| DSM | Deep Sub-Micron |
| ESD | Electro-Static Discharge |
| iCCD | Intensified Charge Coupled Device |
| EM CCD | Electron Multiplying Charge Coupled Device |
| ENOB | Effective Number of Bits |
| FCS | Fluorescence Correlation Spectroscopy |
| FET | Field Effect Transistor |
| FF | Fill Factor |
| FLIM | Fluorescence Lifetime Imaging Microscopy |
| FoM | Figure-of-Merit |
| FR | Free-Running |
| FWHM | Full-Width Half-Maximum |
| FW(100/M) | Full-Width at one hundredth of Maximum |
| GND | Ground |
| GOI | Gated Optical Intensifier |
| GR | Generation-Recombination |
| HDL | Hardware Description Language |
| HV | High Voltage |
| IAT | Inter-Arrival Time |
| IBM | International Business Machines Corporation |
| IC | Integrated Circuit |
| IRF | Instrument Response Function |
| I-V | Current-Voltage |
| I/O | Input/Output |
| LIDAR | Light Detection and Ranging |
| LLL | Low-Level Light |
| LO-HI | Low-to-High |
| LoC | Lab-on-Chip |
| LSB | Least Significant Bit |
| LSM | Least Square Method |
| MCP | Microchannel Plate |
| MIM | Metal-Insulator-Metal |
| MOS | Metal-Oxide-Semiconductor |
| MOSFET | Metal-Oxide-Semiconductor Field-Effect Transistor |
| MOSIS | Metal-Oxide-Semiconductor Implementation Service |
| MPD | MicroPhoton Devices |
| MPW | Multi-Project Wafer |
| NIR | Near Infra-Red |
| ND | Neutral Density |
| NMOS | n-channel MOSFET |
| NW | N-Well |
| OEIC | Optoelectronic Integrated Circuit |
| OP | Oxide Protect |
| PET | Positron Emission Tomography |
| PDE | Photon Detection Efficiency |
| PCB | Printed Circuit Board |
| PCR | Photon Counting Rate |
| PD | Phase Detector |
| PDF | Probability Density Function |
| PEB | Premature Edge Breakdown |
| PGA | Pin-Grid-Array |
| PMOS | p-channel MOSFET |
| PMF | Probability Mass Function |
| PMT | Photo-multiplier Tube |
| PPG | Pulse Pattern Generator |
| PQ | Passive Quenching |
| PQAR | Passive Quench, Active Reset |
| PQPR | Passive Quench, Passive Reset |
| PVT | Process Voltage Temperature |
| PW | Pulse Width |
| QE | Quantum Efficiency |
| R6G | Rhodamine 6G |
| REF | Reference |
| RF | Radio Frequency |
| RIE | Reactive Ion Etching |
| RLD | Rapid Lifetime Determination |
| RMS | Root Mean Square |
| SE | Single Ended |
| SF | Source Follower |
| SiPD | Silicon Photodiode |
| SPAD | Single Photon Avalanche Diode |
| SMS | Single Molecule Spectroscopy |
| SMU | Source Measurement Unit |
| SNR | Signal-to-Noise Ratio |
| SoC | System-on-a-Chip |
| SPA | Semiconductor Parameter Analyzer |
| SRH | Shockley-Read-Hall |
| STI | Shallow Trench Isolation |
| SYNC | Synchronizer |
| TAC | Time to Amplitude Converter |
| TCSPC | Time Correlated Single Photon Counting |
| TDC | Time to Digital Converter |
| TG | Time Gated |
| TGSPC | Time Gated Single Photon Counting |
| TSPC | True-single-phase clock |
| TTR | Transit Time Response |
| TTS | Transit Time Spread |
| TTTR | Time-Tagged-Time-Resolved |
| TRIG | Trigger |
| UV | Ultraviolet |
| ToF | Time of Flight |
| VB | Valence Band |
| VCDL | Voltage Controlled Delay Line |
| VDL | Vernier Delay Line |
| VLSI | Very Large Scale Integration |
| VLQC | Variable Load Quenching Circuit |

Table of Contents

[Chapter 1 Introduction 1](#_Toc439283111)

[1.1. Thesis Motivations 1](#_Toc439283112)

[1.2. Research Contributions 6](#_Toc439283113)

[1.3. Thesis Overview 7](#_Toc439283114)

[Chapter 2 Background and Review 9](#_Toc439283115)

[2.1. Fluorescence lifetime imaging (FLIM) 9](#_Toc439283116)

[2.1.1. Time-correlated single photon counting (TCSPC) FLIM 11](#_Toc439283117)

[2.1.2. Time Gated Single-Photon Counting (TGSPC) FLIM 14](#_Toc439283118)

[2.2. Existing single-photon detector technologies 16](#_Toc439283119)

[2.2.1. Photo-multiplier Tube (PMT) 17](#_Toc439283120)

[2.2.2. Charge-Coupled Devices (CCD) 19](#_Toc439283121)

[2.2.3. Single-photon avalanche diode (SPAD) 21](#_Toc439283122)

[2.3. CMOS SPADs 24](#_Toc439283123)

[2.3.1. Dark Count Rate (DCR) 25](#_Toc439283124)

[2.3.2. Afterpulsing 27](#_Toc439283125)

[2.3.3. Breakdown Voltage (BV) 28](#_Toc439283126)

[2.3.4. Photon Detection Efficiency (PDE) 29](#_Toc439283127)

[2.3.5. Timing Resolution 31](#_Toc439283128)

[2.4. Time-to-Digital Converters (TDC) 32](#_Toc439283129)

[2.4.1. Key TDC Specifications 33](#_Toc439283130)

[2.4.2. TDC Architecture Review 38](#_Toc439283131)

[Chapter 3 CMOS SPAD Design in standard 130 nm technology 43](#_Toc439283132)

[3.1. Standard deep-submicron CMOS Technology 43](#_Toc439283133)

[3.1.1. Technology features of standard DSM CMOS technology 44](#_Toc439283134)

[3.1.2. SPAD Guard-ring Structures in DSM CMOS 51](#_Toc439283135)

[3.2. SPAD structures fabricated in 130 nm standard CMOS 52](#_Toc439283136)

[3.2.1. Structural Characterization and Fabrication Details 53](#_Toc439283137)

[3.2.2. I-V and Breakdown Voltage Measurements 57](#_Toc439283138)

[3.2.3. SPAD pixel circuit modeling 62](#_Toc439283139)

[3.3. Passively-Quenched Pixel Designs and Measurements 65](#_Toc439283140)

[3.3.1. Unbuffered SPAD pixels 66](#_Toc439283141)

[3.3.2. Free-running (FR) Source-follower SPAD (SF-SPAD) pixel 71](#_Toc439283142)

[3.3.3. Free-running (FR) Common-Source SPAD (CS-SPAD) pixel 76](#_Toc439283143)

[3.3.4. Time-Gated SPAD (TG-SPAD) Pixel 79](#_Toc439283144)

[Chapter 4 Dark Count Rate and Afterpulsing Performance of Free-running and Time-gated SPADs 82](#_Toc439283145)

[4.1. Dark Noise: Dark Count Rate (DCR) 83](#_Toc439283146)

[4.1.1. Dark Noise Mechanisms 83](#_Toc439283147)

[4.1.2. Characterization Methods 86](#_Toc439283148)

[4.1.3. Experimental DCR results for free-running and time-gated pixels 90](#_Toc439283149)

[4.2. Afterpulsing Characteristics of Free-running and Time-gated SPADs 99](#_Toc439283150)

[4.2.1. Afterpulsing mechanisms 100](#_Toc439283151)

[4.2.2. Afterpulsing Characterization 102](#_Toc439283152)

[4.2.3. Experimental results for free-running and time-gated pixels 104](#_Toc439283153)

[Chapter 5 CMOS SPAD Optical Characterization And Fluorescence Lifetime Measurement Results 110](#_Toc439283154)

[5.1. Optical Characterization Results 111](#_Toc439283155)

[5.1.1. Dynamic Range 111](#_Toc439283156)

[5.1.2. Photon Detection Efficiency (PDE) 116](#_Toc439283157)

[5.1.3. Timing Resolution 123](#_Toc439283158)

[5.2. Fluorescence Lifetime Measurements 126](#_Toc439283159)

[5.2.1. Rhodamine 6G Lifetime 127](#_Toc439283160)

[5.2.2. Ruby Crystal Lifetimes 133](#_Toc439283161)

[Chapter 6 CMOS Time-to-Digital Converter Design, Simulation and Measurements 138](#_Toc439283162)

[6.1. TDC Architecture and Design 139](#_Toc439283163)

[6.1.1. Voltage-Controlled Delay Lines (VCDL) 143](#_Toc439283164)

[6.1.2. Synchronizer Logic 146](#_Toc439283165)

[6.1.3. Delay-Locked-Loop (DLL) 147](#_Toc439283166)

[6.1.4. Data Read-out scheme 149](#_Toc439283167)

[6.2. Measured TDC Performance 152](#_Toc439283168)

[6.2.1. Measurement Set-up 153](#_Toc439283169)

[6.2.2. Transfer Characteristic 154](#_Toc439283170)

[6.2.3. Non-linearity 158](#_Toc439283171)

[6.2.4. Jitter 159](#_Toc439283172)

[Chapter 7 Conclusion 162](#_Toc439283173)

[7.1. Summary and Discussion 162](#_Toc439283174)

[7.2. Recommendations for Future Work 165](#_Toc439283175)

[References 168](#_Toc439283176)

List of Figures

Figure 1‑1: Illustration of various different FLIM systems and their applications (a) FLIM using commercially available instruments (b) Miniaturized FLIM systems ………………………………………………….2

Figure 1‑2: Evolution of single-photon detector technology ……………………………………………………...3

Figure 2‑1: (a) Energy level diagram of a fluorescent molecule. Excitation of the molecule into an excited state by the absorption of a photon promotes of the weakly bound electrons to a higher energy level. The excited electron returns to the ground state either by the emission of a fluorescence photon (green) or a non-radiative transition (grey). (b) Representation of a typical fluorescence absorption and emission spectrum illustrating the Stoke’s shift phenomenon……………………………………………………10

Figure 2‑2: (a) Principle of operation of TCSPC. In excitation period *N1*, the second arriving photon is missed due to dead-time, resulting in pulse pile-up at the detector. (b) A generalized TCSPC measurement set-up. A TDC measures the inter-arrival time (IAT) between laser pulses and photon arrivals and the results are stored in a histogram……………………………………………………………………………………12

Figure 2‑3: (a) Principle of operation of TGSPC. (b) Generalized TGSPC measurement set-up…………………15

Figure 2‑4: (a) Simplified structure of a conventional PMT and examples of the signals measured at the anode. [26]. (b) Photon Detection Efficiency (PDE) comparison of different photocathode materials [28]…..18

Figure 2‑5: (a) Schematic representation of charge transfer in a CCD [27]. (b) Principle of operation of an EM CCD [121] ……………………………………………………………………………………………...20

Figure 2‑6: Principle of SPAD operation (a) Avalanche breakdown process in a reverse biased pn junction. (b) Load-line representation of SPAD operation [72]………………………………………………………22

Figure 2‑7: Summary of CMOS SPAD DCR temperature dependence..…………………………………………26

Figure 2‑8: Summary of CMOS SPAD DCR excess bias dependence….……………………………………….26

Figure 2‑9: Summary of CMOS SPAD PDE performance………………………………………………………30

Figure 2‑10: (a) Input-output characteristic of an ideal 6-bit TDC with *TLSB* = T*CLK*­/64 =­ 156.25 and T*CLK* = 1000 ps. (b) Associated quantization error values.…………………………………………………….……...33

Figure 2‑11: (a) Input-output characteristic of a 6-bit TDC including non-linearity. (b) Plot of the associated quantization error……………………………………………………………………………………….33

Figure 2‑12: Plot of DNL/INL for 6-bit TDC. (a) max. INL < 1 LSB, and (b) max. INL > 1 LSB…..…………….36

Figure 2‑13: Corresponding quantization error pmf of 6-bit TDC characteristics shown in Fig. 2-12. Quantization rms error for max. INL < 1 LSB in case (a) is much closer to the ideal case than for max. INL > 1 LSB in case (b)…..………....………………………………………………………………………………...36

Figure 2‑14: Representation of TDC output code histogram [202]………………………………………………37

Figure 3‑1: Cross-section views of a triple-well DSM CMOS technology: (a) Inter-metal dielectric stack [289] and (b) transistor structures [228]. (c) Three possible photodiode structures are available for SPADs: n+/p-substrate, deep n-well/p-substrate and p+/n-well. Arrows point in the direction of the electric field which promotes the drift current due to the minority carriers… ………………………………….…….45

Figure 3‑2: (a) Layout view of SPAD test structure with important dimensions in microns. (b) Unbuffered SPAD pixels – Left pixel: Silicided n+ junction. Right pixel: Non-silicided n+ junction. (c) Cross-section view of SPAD and nMOS transistor. (d) The wire-bonded die resides in the cavity of a 68 pin-grid-array (PGA68) ceramic package....………………………………………………………...………………....54

Figure 3‑3: (a) Test structure used to evaluate SPAD breakdown voltages. (b) Measured breakdown voltages of SPAD and parasitic junction. (c) Illustration of the voltage headroom limits for proper SPAD operation………………………………………………………………………………………………..59

Figure 3‑4: (a) Measured breakdown voltage as a function of temperature for 7 randomly chosen SPADs. (b) I-V curves for SPAD3 between -40 °C and 60 °C…………………………………………………….…….60

Figure 3‑5: Comparison of breakdown voltages of non-silicided (SPAD1) and silicided (SPAD2) devices (a) Measured I-V curves at room temperature for 5 different chips. (b) Temperature variation of breakdown voltage for SPAD1 and SPAD2………………………………………………………………………...61

Figure 3-6: Circuit model used for simulating passively-quenched SPAD [256]………...………………………62

Figure 3‑7: (a) Simulated and measured cathode voltage, anode current and VDD current during passive quenching and recharge. (b) Simulation of cathode voltage and comparator output during avalanche re-triggering for different delays ΔT………………………………………………………………………63

Figure 3‑8: (a) Unbuffered SPAD test structure. Switches S1 and S2 enable selection between two different load capacitances for the SPAD. (b) Typical cathode waveform and illustration of comparator outputs for two different thresholds……………………………………………………………………………………..67

Figure 3‑9: (a) Measured SPAD pulse amplitude histograms for as a function of excess voltage (b) Measured average pulse amplitudes as function of *VEX* of eight different SPAD chips. Saturation effects are prominent for SPAD5 and SPAD8.…………………………………………………………….……….68

Figure 3‑10: Measured pulse width histograms of unbuffered SPAD pixels for (a) *CSPAD* = 1 and 9.5 pF and (b) *VEX* = 1.5 and 2.5 V.………………………………………………………………………………….…69

Figure 3‑11: Mean and standard deviation of pulse width histogram as a function of *VEX*. (a) Measured results of eight chips for CSPAD = 1 pF. (b) Measured results of SPAD1 for two different CSPAD values……70

Figure ‑12: (a) SPAD front-end with nMOS source follower and current-source load. (b) simulated source-follower gain and bias current variation versus temperature for different process corners…………….72

Figure 3‑13: (a) Simulated and measured waveforms for SF-SPAD pixel. PW2 > PW1 due to an afterpulse occurring during the recharge time. (b) Measured pulse amplitude distributions for different temperatures.……………………………………………………………………………………………73

Figure 3‑14: Average pulse amplitudes of SF-SPAD pixel as a function of (a) excess voltage, and (b) temperature……………………………………………………………………………………………..74

Figure 3‑15: Average pulse widths of SF-SPAD pixel as a function of (a) excess voltage and (b) temperature…75

Figure 3‑16: (a) Schematic of SPAD pixel using common-source amplifier with current source load. (b) Simulated and measured waveforms with relative positions of V*TH1* and V*TH2* labeled.......……………………….76

Figure 3‑17: (a) Left: Simulated I/O characteristic of CS front-end shows that relative positions of V*TH1* and V*TH2* are unchanged at different temperatures Right: Measured and simulated output pulse widths at different excess voltages. (b) Simulated and measured SPAD output pulses that illustrate the effects of afterpulsing on the output pulse width (PW). PW1 > PW2 due to an afterpulse occurring during the recharge time...………………………………………….………………………………………………77

Figure 3‑18: (a) Measured pulse width distribution in the dark and with light at -30 °C and room temperature at *VEX* = 3.6 V. (b) Measured pulse width (with standard deviation error bars) as a function of temperature at two different excess voltages...………………………………………………………………………78

Figure 3‑19: (a) Schematic and layout of TG SPAD pixel. (b) Simulated and measured waveforms. In the first time gate, a photon is detected, resulting in an output pulse being produced. In the second time gate, no photon is detected and therefore no output pulse is produced.………………………………………….80

Figure 4‑1: (a) Illustration of DCR mechanisms [90],[155]. (b) DCR as a function of temperature for a commercially available SPAD [260]……………………………………………………………………83

Figure 4‑2: (a) Output pulses measured by the oscilloscope for an SPAD pixel with SF front-end (b) Corresponding histogram of pulse counts in a 5 ms interval…………………………………………….87

Figure 4‑3: (a) Calculation of avalanche inter-arrival times for unbuffered SPAD pixel. (b) Resulting IAT histogram showing raw histogram data, data after smoothing and resulting exponential fit……………89

Figure 4‑4: IAT distributions at *VEX* = 1.25 V displayed on a log-log scale for (a) unbuffered SPAD pixel and (b) CS-SPAD pixel…………………………………………………………………………………………90

Figure 4‑5: (a) Measured distribution of dark counts for unbuffered SPAD during a 0.5 ms interval. Inset: Average count values (with standard deviation error bars) as a function of excess voltage. (b) Distribution of IATs for the same SPAD. Inset: Calculated CV values as a function of excess voltage……………………..91

Figure 4‑6: (a) Measured and corrected DCR as a function of excess voltage extracted from dark count histograms (0.5 ms time interval) and from exponential fitting of IAT histograms. (b) Measured DCR of eight different chips for the unbuffered SPAD test structure………………………………………………….93

Figure 4‑7: (a) Measured distribution of dark counts for CS-SPAD pixel during a 0.5 ms interval. Inset: Average count values (with standard deviation error bars) as a function of excess voltage. (b) Distribution of IATs for the SPAD. Inset: Calculated CV values as a function of excess voltage……………………………..93

Figure 4‑8: (a) Measured total DCR of SPAD-CS chips as a function of *VEX*. (b) Corresponding CV as a function of *VEX*………………………………………………………..………………………………………….94

Figure 4‑9: Temperature measurements of DCR for SF-SPAD pixel. (a) DCR as a function of temperature for *VEX* = 0.4 V – 1.6 V. The solid lines represent the total measured DCR while the dashed lines represent the primary DCR component obtained from exponential fitting of the IAT histograms. (b) Corresponding Arrhenius plot and extracted activation energies as a function of excess voltage.……………………....95

Figure 4‑10: Temperature measurements of DCR for CS-SPAD pixel. (a) DCR as a function of temperature at *VEX* = 1.3 V for seven measured chips. SPAD1 and SPAD2 represent best-case and worse-case DCR measurements. Average DCR of seven pixels and the standard deviation error bars are also shown. (b) Corresponding Arrhenius plot and extracted activation energies……………………………………….97

Figure 4‑11: Measured DCR of silicided (red) and non-silicided (blue) SPADs (a) DCR as a function of *VEX*. Afterpulses are subtracted from Primary DCR. Exponential dependence on *VEX* is indicative of tunneling effects. (b) IAT distributions for *VEX* = 0.5 V and 1.2 V. Afterpulsing effects seen as deviations between total and primary DCR for lower IATs……………………………………………………..…………..98

Figure 4‑12: (a) Schematic representation of trapping and subsequent release of an electron by a deep level. (b) The probablility density of afterpulse generation in a silicon SPAD operating at room temperature. Increasing the hold-off times reduces the afterpulsing probablity [153]………………………………100

Figure 4-13: Measured IAT distribution and exponential fitting results for MPD SPAD……………………….104

Figure 4‑14: (a) Measured and fitted IAT distributions for unbuffered SPAD pixel at T = -30 °C. (b) Measured afterpulsing probability and count rate as a function of excess voltage. A halogen lamp was used to provide background illumination to increase the count rate and reduce the measurement time as a result…………………………………………………………………………………………………..105

Figure 4‑15: Measured and fitted IAT histograms of two FR pixels: (a) SPAD1 and (b) SPAD2 at –30 °C. All data are for VEX = 1.3 V………………………………………………………………………………..106

Figure 416: (a) IAT probability distributions at different temperatures for SPAD1 (Left) and SPAD2 (Right) at *VEX* = 1.3 V. CV of histogram data is shown in the insets. (b) Calculated afterpulsing probabilities (Left) calculated hold-off time required for 1% afterpulsing (Right) for FR SPAD pixels as function of temperature at *VEX* = 1.3 V. Errors in the calculations arise from small uncertainties in temperature, breakdown voltage, and from variations in quality-of-fit to afterpulsing data………………………..107

Figure 4‑17: (a) Measured distributions of afterpulsing in time-gated mode for illuminated SPAD at –30 °C with *VEX* = 1.3 V. 7 ns hold-off time results in 34.5 % afterpulsing probability (Left). No detectable afterpulses for 157 ns hold-off time (Right). The inset shows the superposition of the first 31 time-gates following an avalanche. The measured full-width at half-maximum (FWHM) of the gate width is 2.9 ns. (b) Measured afterpulsing probability versus hold-off time for TG SPAD. Left: Fitted results at –30 °C show temporal behavior between exponential and power law. Right: At room temperature the behavior is exponential…………………………………………………………………………………………....108

Figure 5-1: (a) Instrumentation used for optical characterization of SPADs. (b) Right: Measured optical spectrum of xenon lamp and filtered light. Left: Measured optical powers for λ between 520 – 580 nm………...112

Figure 5-2: (a) The measured input photon flux striking the detector along with the uncorrected SPAD counting rate is plotted as a function of the optical attenuation. (b) The corrected SPAD counting rate as a function of optical intensity is shown along with a linear fit…………………………………………………….113

Figure 5‑3: (a) The measured input photon flux striking the detector is plotted along with the uncorrected SPAD counting rate at different excess voltages for decreasing levels of optical attenuation. (b) The corrected SPAD counting rate with background DCR subtracted is plotted as a function of optical intensity. Measurement taken at λ = 560 nm……………………………………………………………………..114

Figure 5‑4: Measured dynamic range of non-silicided SPAD. (a) Plot of the measured input photon flux striking the detector along with the uncorrected SPAD counting rate as a function of the optical attenuation (b) Plot of corrected SPAD counting rate as a function of optical intensity on the SPAD. Measurement taken at λ = 560 nm.……………………………………………...…………………………………………..115

Figure 5‑5: (a) Measured PDE of the MPD SPAD compared to the quoted value. (b) Left: Measured IAT histogram used to evaluate PDE of MPD SPAD at λ = 400 nm. Effects of afterpulsing are eliminated from the measurement by taking the primary counting rate as 1/τ2. Right: CV of measured IAT histograms used in PDE evaluation as a function of wavelength……………………………………...117

Figure 5‑6: (a) Measured PDE at different wavelengths as a function of optical attenuation. The PDE is overestimated at the higher optical attenuations. (b) Measured pulse count histograms with 1 ms integration time at 580 nm……………………………………………………………………………..118

Figure 5‑7: (a) Measured PDE of silicided CS-SPAD pixel at three different excess voltages. The measurements were performed with an optical intensity that ensured accurate PDE calculations. Inset shows a histogram of CV values obtained from the IAT distributions measured at each wavelength and bias point. (b) Transmittance of light passing through the stack of dielectric layers in 130 nm IBM CMOS technology with ±20% process variation in thickness of layers [289]……………………………………………..119

Figure 5‑8: (a) Measured PDE of non-silicided SPAD test structures pixel at five different excess voltages……………………………………………………………………………………………......121

Figure 5‑9: (a) The measusred PDE as a function of excess voltage for non-silicided SPAD pixel at three wavelengths. (b) The corresponding plot of DCR versus PDE for the same pixel……………………..122

Figure 5‑10: (a) Measured PDE performance of TG-SPAD pixel. The measured PDE (green) was normalized (red) by the duty cycle and compared with free-running PDE performance (blue). Both pixels were silicided. (b) Comparison of PDE performance between silicided and non-silicided SPAD pixels as well as MPD SPAD. The y-axis is shown on a log scale…………………………………….……………..123

Figure 5‑11: (a) The experimental set-up used to measuring the timing response of SPADs. (b) Measured timing response of MPD SPAD (left) had good agreement with the manufacturer’s specifications. The measured timing jitter of CS-SPAD pixel (right) showed excellent performance in comparison………………...124

Figure 5‑12: Measured timing jitter distributions of CS-SPAD pixels for two excess voltages, *VEX* = 1 V and 1.2 V. The laser wavelengths are (a) λ = 470 nm and (b) λ = 510. nm……………………………….…….125

Figure 5‑13: (a) Measured timing jitter distributions of unbuffered SPAD test structure for four excess voltages and (b) corresponding FWHM values as a function of excess voltage at λ = 510 nm………………....126

Figure 5‑14: (a) Experimental set-up used to measure florescence lifetime of R6G. (b) A picture of the laboratory set-up ………………………………………………………………………………………………….128

Figure 5‑15: (a) Measured IRF from MPD SPAD. (b) Measured fluorescence decays of R6G in ethanol (left) and methanol (right) obtained with MPD SPAD. Variations in fluorescence lifetime were apparent for different concentrations (10-4 to 10-6 M) R6G solutions in ethanol and methanol …………………...129

Figure 5‑16: (a) Measured IRFs of the non-silicided SPAD pixels obtained with the scattering solution at two different excess voltages. (b) Measured fluorescence decay data of R6G solutions in ethanol (left) and methanol (right) at 10-5 M and 10-4 M concentrations, respectively. Best fit exponential decays are also shown with their corresponding lifetimes……………………………………………………………..132

Figure 5‑17: (a) Measured IRFs of the silicided SPAD pixels obtained with the scattering solution at two different excess voltages. (b) Measured fluorescence decay data of R6G solutions in ethanol (left) and methanol (right) at 10-5 M and 10-4 M concentrations, respectively. Best fit exponential decays are also shown with their corresponding lifetimes………………...………………………………………………………..132

Figure 5‑18: (a) The measurement set-up used to measure the fluorescence lifetime of ruby crystal. (b) A photograph of the laboratory set-up…………………………………………………………………...133

Figure 5‑19: Measured data from the set-up in Fig. 5-10 with reference MPD SPAD. The intensity signals are shown for two different integration times. The chopper reference signal is also shown. (a) The acquisition time of the oscilloscope was set to 128 ms to record 2.5 excitation cycles using a 1 GS/s oscilloscope sample rate.(b) When the chopper reference signal is low the chopper blade blocks the incident light and the fluorescence decay can be obtained by fitting the decay portion of photon count rate with an exponential function…………………………………………………………………………………..135

Figure 5‑20: Measurement results for ruby lifetime obtained with the non-silicided SPAD pixel. (a) The measured SPAD output is plotted along with the extracted photon counting rates for two different integration times. (b) Fitting result is shown for the portion of the exponential decay in photon count rate………….…..136

Figure 5‑21: Measurement results for ruby lifetime obtained with the silicided SPAD pixel. (a) The measured SPAD output is plotted along with the extracted photon counting rates for two different integration times (b) Fitting result is shown for the portion of the exponential decay in photon count rate……………..136

Figure 5‑22: (a) Comparison of measured photon counts between MPD SPAD and silicided and non-silicided CMOS SPADs. (b) Corresponding plot of fluorescence decays on a y-axis log scale from which the fluorescence lifetime of ruby crystal was obtained..…………………………………………………..136

Figure 6‑1: Coarse-fine TDC timing diagram. The fine interpolator calculates the time between rising edges of START2 and STOP2. This result *T2* is used to obtain the residue time *TR*, which is added to the coarse-interpolation result *T1* to digitize the timing interval *TM*……………………………………………….140

Figure 6‑2: Architecture of a single channel of the proposed coarse-fine interpolating TDC. Only the circuits in the hatched portion of the TDC core need to be replicated for a multi-channel realization…………….142

Figure 6‑3: Voltage Controlled Delay Line (VCDL) (a) Schematic of VCDL delay cell used for coarse and fine interpolation, comprised of a differential stage (diff. stage), differential to single-ended (diff. to SE) conversion stage, and edge aligner stage. (b) Block diagram of VCDL. A single-ended to differential (SE-to-DIFF) converter is used to convert the single-ended input signal to differential format. A Voltage-to-Current (V-to-I) converter generates a bias current from *VCTRL*. (c) Layout of coarse VCDL.………………………………………………………………………………………………...144

Figure 6‑4: Post-layout simulation results of delay transfer characteristic of delay element and charge pump mismatch for (a) coarse and (b) fine interpolators. The range of VCDL control voltages that produce TC and TLSB at different process corners are indicated for 120 MHz reference clock frequency. In this range, the charge pump mismatch is at its lowest value, resulting in lower charge pump phase offset...146

Figure 6‑5: (a) Synchronizer logic diagram and (b) post-layout simulation results for case when *TM* = 5.1 ns and *TLSB*=156.25 ps (corresponding to a 100 MHz reference clock). In this case, *TM*/*TC* > 8, and according to *SYNC* logic, STOP2=ϕS*=*ϕ1(9). The *SYNC* delay between ϕ1(9)and STOP2is Δ2……………………..146

Figure 6‑6: (a) Timing diagram of the TDC read-out. Once the DLLs are locked, the TDC outputs valid data M1, M2, M3, etc. (b) Post-layout simulation results of TDC read-out. Locking times of DLL1 and DLL2 are approximately 7 μs…………………………………………………………………………………….150

Figure 6‑7: (a) Photograph of the TDC chip highlighting the TDC core area (b) Top level diagram of TDC chip and photograph of the printed circuit board (PCD) used for testing the TDC prototype chips. (c) TDC core layout with main components highlighted.…………………………………………………..…...151

Figure 6‑8: (a) Measurement set-up used for TDC characterization at 1 MHz sample rate. (b) Measured TDC input signals, CLK and HIT (left) and their respective cycle-to-cycle jitter (right)……………………153

Figure 6‑9: (a) Measured histogram of delays from set-up in Fig. 6-8(a). (b) Measured average delays show a deviation from the ideal behavior……………………………………………………………………...154

Figure 6‑10: Comparison between measured, corrected and ideal transfer characteristics of a 6-bit TDC. Figure inset shows an example of non-monoticitiy in the measured characteristic due to timing jitter. Measured TDC characteristics was obtained using a 120 MHz reference clock frequency corresponding to *TLSB* = 130.2 ps………………………………………………………………………………………………..155

Figure 6‑11: Measured TDC characteristics of 10 chips using a 120 MHz reference clock frequency corresponding to *TLSB* = 130.2 ps. (a) Comparison between measured, corrected and ideal transfer characteristics of a 6-bit TDC. Figure inset shows an example of non-monoticitiy in the measured characteristic due to timing jitter. (b) Measured TDC quantization error (top) and associated histogram (bottom). The average rms resolution of the 10 measured chips was 0.61 LSB…………………………………………………….157

Figure 6‑12: Top and bottom graphs are the DNL and INL, respectively, of 10 measured chips at room temperature……………………………………………………………………………………………158

Figure 6‑13: TDC jitter measured at 1 MHz sample rate. (a) The TDC jitter across the entire measurement range was 0.385 LSB rms. (b) Close-up view of the jitter due to a coarse code transition. In this shorter time range, the rms jitter was 0.8 LSB………………………………………………………………………159

Figure 6‑14: TDC jitter measured at 60 MHz sample rate using 100 MHz reference clock……………………..160

List of Tables

Table I – Different reported SPAD pixels with low reported afterpulsing (AP) performance…………………..28

Table II – Summary of CSPAD breakdown voltages and DCR performance at room temperature for different technologies…………………………………………………………………………………………….29

# Introduction

## Thesis Motivations

Fluorescence lifetime imaging microscopy (FLIM) is a powerful tool used in biological imaging, whereby the image contrast is provided by fluorescence lifetime rather than fluorescence intensity or wavelength [1]-[3]. The great advantage of FLIM is that it involves non-ionizing radiation, is minimally invasive and non-destructive, and can therefore be applied to living cells and tissues. This is especially valuable in biomedicine for minimally invasive surgery and disease detection studies [4]-[7]. FLIM is routinely used not only in biology and medicine [8]-[11], but also in other diverse scientific fields such as in forensics [12], combustion research [13], microfluidic systems [14], temperature sensing [15], and art conservation [16].

One of the main challenges of fluorescence lifetime sensing is that fluorescence decays of most biological samples are in the nanosecond range. Therefore, measuring fluorescence on such short time scale requires special instruments that can detect the very weak optical signals with very high (sub-nanosecond) timing resolution [17]-[29]. Fig.1-1(a) shows examples of several different FLIM systems that utilize commercially available instruments and the corresponding time-resolved fluorescence images of biological samples. In the resulting images, the contrasting fluorescence lifetime characteristics can be used to discriminate between regions of healthy and sick tissue for cancer detection [6],[30]. The fluorescence lifetime is an intrinsic molecular contrast parameter that can be readily measured, but only with very high-performance detection systems [24]-[41] In most cases, the emitted light from a fluorescent sample is too weak to create an analog voltage representing the optical flux using conventional image sensors. Single photon-counting detectors, on the other hand, provide digital pulses for every photon detection event and are therefore well suited for low-light level (LLL) detection and can achieve high signal-to-noise ratio (SNR) [27].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 1‑1: Illustration of various different FLIM systems and their applications (a) FLIM using commercially available instruments (b) Miniaturized FLIM systems.

Figure 1‑2: Illustration of various different FLIM systems and their applications (a) FLIM using commercially available instruments (b) Miniaturized FLIM systems.

The quest towards miniaturization of FLIM instruments necessitates reduction of the power consumption, decrease in module size, and lower overall system cost. Fig.1-1(b) shows several examples of miniaturized FLIM systems that have been recently developed for fluorescence lifetime analysis [42]-[60]. Solid-state detectors are ideal for this purpose because of their lower power consumption, smaller size, and most importantly for mass-produced devices, significantly lower costs. Developments in solid-state single-photon detector design, in parallel with advances in optics, and analog and digital signal processing electronics, have enabled the increasing use of FLIM over the years for these reasons.

Fig.1-2 outlines the evolution of single-photon detector technology. Photomultiplier tubes (PMT) and intensified charge coupled devices (iCCD) have for a very long time been the best single-photon sensors available for biomedical imaging, especially in terms of single photon sensitivity and picosecond temporal response [26]- [29]. However, despite their high levels of performance, these detectors are expensive, bulky, fragile, and consume large power.

|  |
| --- |
|  |

Figure 1‑2: Evolution of single-photon detector technology.

|  |
| --- |
|  |

Figure 1‑4: Evolution of single-photon detector technology.

The single photon avalanche diode (SPAD) has emerged as a versatile, cost effective and easy to use detector for FLIM measurements, as well as in other applications where high timing resolution and single-photon detection capability is simultaneously required. SPADs are manufactured either in dedicated silicon processes that are optimized especially for SPAD performance, or in complementary-metal-oxide-semiconductor (CMOS) technology [61],[62]. Although the electro-optical characteristics of SPADs manufactured in standard digital CMOS technology are not optimized, their monolithic integration with high-speed electronic circuits offers significant advantages over the competing silicon technologies.

The low-cost of manufacturing and complete system-on-chip integration capability has encouraged the growth of CMOS SPADs for a wide variety of applications in biomedical imaging. CMOS SPADs have been developed for to measure 2D fluorescence lifetime images [42]-[51] and to realize fully integrated lab-on-chip (LoC) devices [47],[49],[50],[51],[57],[58]. CMOS SPAD arrays have also been developed to detect individual scintillation events in ToF PET [63]-[65], take 3D images with centimetric depth resolution using light detection and ranging (LIDAR) [66]-[73], and to record Raman spectra [52]-[56] using high-speed, time gating techniques [74]-[82]. The downscaling of CMOS technology into the deep-submicron regime (DSM) has made it possible to drastically reduce the size, cost and power consumption of time-resolved, single-photon imaging systems, while simultaneously improving their functionality, versatility and robustness.

Many of the single-photon detector applications utilize the time-correlated single-photon counting (TCSPC) technique to achieve the very high levels of temporal-resolution (down to several picoseconds) [24]-[29]. TCSPC requires time-to-digital converter (TDC) circuits to digitize the time elapsed between a detected photon and a stable reference clock signal edge. Traditionally, the vast majority of available TDCs functioned as stand-alone custom ICs or were used in applications where the requirements on area and power consumption were relaxed, whereas the main limitation of FLIM instruments has been a slow frame rate due to the large SNR required in order to separate the desired signal from uncorrelated background noise [3],[45]. To achieve higher frame rates, greater levels of parallelism can be utilized. An ‘in-pixel’ approach has each SPAD and its associated front-end circuit connected to a dedicated TDC, resulting in very high level of measurement throughput [43],[44],[71]-[73].

At present, single photon imaging applications are demanding larger arrays, faster acquisition rates and better time resolution, necessitating the design of high-performance multichannel TDCs. However, the requirements of multi-channel TDCs are more stringent in terms of area and power consumption, robustness to process, voltage, and temperature (PVT) variations, as well as power supply noise and electrical crosstalk immunity. Architectural considerations also have important implications in detector performance because the time uncertainty for single photon detection is limited by SPAD jitter and TDC non-linearity. While the in-pixel TDC approach achieves highest levels of parallelism, the readout interface between pixels is characterized by data rates on the order of several gigabits per second, leading to higher power consumption [43]-[45]. The TDC linearity performance must also be sacrificed to achieve reasonable pixel fill-factor and attain smaller circuit area for in-pixel time-digitization. TDCs may also produce significant electrical crosstalk noise that can couple to SPADs and affect their jitter and linearity performance [47]. In contrast, designs with TDC sharing between multiple SPADs leads to improved fill-factor, while relaxing the TDC constraints on TDC power, area and linearity [64],[69],[70]. As a result, higher measurement accuracy, and better pixel uniformity can be achieved by sharing one TDC amongst many SPADs.

The main goal of this work is to develop SPADs and TDC ICs in a mainstream standard digital deep submicron (DSM) CMOS technology for low-cost, time-resolved fluorescence analysis. This research is therefore devoted to the design and electro-optical characterization of SPADs and the investigation of the potential improvements of sensor performance over previously reported implementations [53],[61],[83]-[85]. Other important research goals include finding optimal operation conditions in terms of temperature and voltage in order to obtain the best performance for SPADs in a standard (non-imaging) CMOS technology [86]. For this purpose, accurate SPAD characterization and analysis were pursued. In addition, demonstration of the capability of the fabricated SPAD to accurately measure fast fluorescence decays in a practical laboratory environment was an important goal.

The other main goal of the research was to design, simulate, fabricate and characterize high-performance TDC prototype ICs in DSM CMOS suitable for high-speed, multi-channel, low-power FLIM applications. For this purpose, a TDC prototype chip was developed for future implementations of SPAD cameras which utilize the TDC sharing scheme. High-speed measurements and sub-nanosecond resolution were the primary design goals. Small nonlinearity of the converter’s transfer characteristic was desired to achieve higher measurement precision and accuracy over the dynamic range, thus eliminating the need for post-processing. The TDC should also be designed to be compact and scalable to multiple channels, and have high fabrication yield. The use of Delay Locked Loop (DLL) circuits for low sensitivity to PVT variations and automatic self-calibration would also result in stable performance over different operating temperatures and supply voltages. The achievements of compact size, high-speed and robustness to PVT variations make the TDC prototype a suitable building block for miniaturized, multi-channel TCSPC systems such as FLIM.

## Research Contributions

Based on a comprehensive literature survey of CMOS SPAD/TDCs implemented by various groups and on a detailed analysis on their comparative performance [61], the specific contributions of this work are the following:

* SPAD design in standard CMOS: A set of single-pixel SPAD test structures were designed and fabricated to explore the single-photon detection performance of SPADs fabricated in a low-cost, standard digital 130 nm IBM CMOS technology available from Canadian Microelectronics Corporation (CMC) [83],[84]. These test structures include passively-quenched SPADs with the largest active areas reported for a standard CMOS technology. The key limitations of previously implemented SPADs were identified in terms of the device structure and improved devices were fabricated. The improvements of the proposed structures were validated by detailed measurements and analysis. The comparative performance of the different pixel test-structures was also studied.
* Experiments and analysis of SPADs: Extensive characterization and analysis of afterpulsing behavior was done for standard CMOS SPADs under both gated and free-running operation [86]. A critical analysis was done on the afterpulsing characteristics of different SPAD pixel structures with their temperature and voltage variations taken into account. Minimization of the afterpulsing behavior at low temperatures was accomplished by using the time-gated mode of operation.
* Practical Application of SPADs: Demonstration of SPAD’s performance was demonstrated for practical laboratory applications. The accurate determination of fluorescence lifetimes in the nanosecond range for Rhodamine solutions, and in the millisecond range for Ruby crystals, was performed for this purpose.
* High-performance TDC prototype: A compact, high-speed, sub-nanosecond resolution TDC prototype IC was developed for FLIM applications. The compact circuit size and high-speed performance of the proposed TDC was achieved by utilizing a dual-interpolation architecture. The effect of PVT variations was reduced with an integrated delay-locked loop (DLL). Characterization of the fabricated TDC chips with different voltage supply and temperature conditions validated the circuit’s robust performance in the measurement of high-speed timing data.

## Thesis Overview

In Chapter 2, fluorescence lifetime detection systems are introduced and an overview of the existing well-established technologies that are used for single-photon detection is provided. Then, the current state-of-the-art single-photon imaging systems employing CMOS technology are considered with emphasis on the key detector characteristics. An outline of TDC specifications is covered and the key architectures are reviewed. Then, the design goals for this research project are introduced.

In Chapter 3, the design and test of SPAD pixels with front-end circuitry implemented in a standard CMOS process is described. The different SPAD pixel designs are comparatively studied. Important technological considerations for the successful fabrication of improved SPAD structures in CMOS are reviewed along with a summary of previously reported works. The performance of the SPAD pixels in an experimental setting is presented along with a demonstration of the pixel’s sensitivity to temperature and voltage changes.

In Chapter 4, the issue of intrinsic detector noise in terms of the dark count rate and afterpulsing probability is discussed. The mechanisms responsible for DCR and afterpulsing in a CMOS process are investigated and the experimental methods, as well as the measured results, are presented. The afterpulsing performance of free-running and time-gated pixels is studied at different temperatures. Optimal conditions are obtained for low afterpulsing probability.

Optical characterization results of the SPAD pixels and a comparison of the detection limits of the different pixel designs are presented in Chapter 5. Results from the study of dynamic range, photon detection efficiency, and timing resolution are described. Then, a demonstration of the developed SPAD pixels being used in fluorescence lifetime measurements is presented.

In Chapter 6, a high-speed, compact and precise TDC designed for SPAD imaging systems is presented. A prototype CMOS chip containing a single TDC channel was fabricated in a standard digital IBM 130 nm CMOS process and was extensively characterized. Full details of the design, characterization and comparisons are described.

In Chapter 7, the thesis is concluded with a summary of the work and a proposal for a number of different areas of research that could be investigated in future to further advance this work.

# Background and Review

This section begins with a review of the target application of the SPADs and TDCs that were developed in this work, namely fluorescence lifetime imaging (FLIM). The competing technologies traditionally used in the field of time-resolved single-photon detection will be discussed. Then, the recent progress of SPADs and TDCs in CMOS technology is reviewed. The main performance characterization metrics and key design issues, challenges and implementation details are highlighted.

## Fluorescence lifetime imaging (FLIM)

Fluorescence is defined as the light emitted by a molecule following the absorption of electromagnetic energy (photons). As shown in Fig. 2-1(a), the emitted light results from the radiative transitions of the lowest electronically excited singlet state, S1, back to the electronic ground state, S0. The singlet states are the energy levels that can be populated by the weakly bound electrons of the fluorescent molecule without undergoing a change in electron spin. The fluorescence decay lifetime *τ* is the time required by a population of *N* excited molecules to decrease exponentially to *N*/*e* via the loss of energy through fluorescence (emission of photons) or through a non-radiative quenching processes. A quenching process internally dissipates or transfers to the molecular environment the energy gained as a result of photon absorption [1]-[3]. The fact that the non-radiative decay rate constant depends on the local molecular environment is the main feature that makes FLIM such a powerful tool for mapping biochemical interactions on the molecular scale in biomedical imaging [11]. Whereas steady-state fluorescence intensity measurements are sensitive to light scattering, background noise and variations in fluorophore concentration, fluorescence decay lifetime measurements are much less so, resulting in improved contrast and sensitivity of the resulting images. Further, the use of optical filters allows excitation and fluorescence photons to be easily separated due to the Stoke’s shift phenomenon, as illustrated in Fig. 2-1(b).

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑1: (a) Energy level diagram of a fluorescent molecule. Excitation of the molecule into an excited state by the absorption of a photon promotes of the weakly bound electrons to a higher energy level. The excited electron returns to the ground state either by the emission of a fluorescence photon (green) or a non-radiative transition (grey). (b) Representation of a typical fluorescence absorption and emission spectrum illustrating the Stoke’s shift phenomenon.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑2: a) Energy level diagram of a fluorescent molecule. Excitation of the molecule into an excited state by the absorption of a photon promotes of the weakly bound electrons to a higher energy level. The excited electron returns to the ground state either by the emission of a fluorescence photon (green) or a non-radiative transition (grey). b) Representation of a typical fluorescence absorption and emission spectrum illustrating the Stoke’s shift phenomenon.

The timescale of fluorescence decays is in the range of picoseconds to nanoseconds. Therefore, very fast photodetectors and electronics are required in biological imaging. The photodetection systems in FLIM measurements are generally categorized as either operating in time-domain or frequency-domain [17]-[29]. In the time domain, the detector measures the fluorescence as a function of time delay following a pulsed excitation. In the frequency domain, the lifetime information is derived from comparisons between a modulated excitation signal and the resulting modulated fluorescence. Both approaches can provide equivalent information, but specific implementations present different trade-offs with respect to cost, complexity, and instrumentation requirements [17]-[22]. Frequency-domain methods are generally less expensive to implement, since expensive laser sources are not needed and the electronic circuitry requirements are somewhat relaxed. However, achieving the optimum signal-to-noise ratio (SNR) in the frequency domain requires careful optimization of measurement parameters that depends on the lifetime of the fluorophore under investigation [20]-[23].

Time-domain measurements, on the other hand, have SNR that approaches the ideal value corresponding to Poisson counting statistics [24]-[29]. This feature, coupled with the tremendous advances in optoelectronic integrated circuits (OEIC) and ultrafast laser technology over the last 20 years, are the main reasons why time-domain techniques have become the ‘gold standard’ for FLIM in terms of accuracy, repeatability and reliability [23]-[29]. Time-domain FLIM measurements can be classified either as time-gated single-photon counting (TGSPC), where photons are counted during brief time intervals occurring at different delays relative to the excitation period, or time-correlated single-photon counting (TCSPC), where the photon arrival times within the excitation period are measured and assigned to discrete time bins [1]-[3]. These two techniques will be described in more detail in the following sub-sections.

### Time-correlated single photon counting (TCSPC) FLIM

TCSPC is a time-domain, point scanning, lifetime measurement technique that relies on single-photon sensitive detectors to obtain photon arrival time information for each detected photon [1]-[3],[17]-[19],[24]-[29]. The operational principle of TCSPC FLIM is shown in Fig. 2-2(a). Each detected photon is logged along with a time stamp denoting the photo’s arrival time relative to a repetitive synchronization pulse from the pulsed light source. This process is repeated many times in order to produce a decay histogram. One of the main advantages of TCSPC is that it is very photon efficient, with data from all gathered photons being processed.

The emission waveform shown in the figure is what would be observed in ideal conditions when a large number of fluorescent molecules are excited and all the fluorescent photons are emitted in a single excitation period. However, because of practical limitations in the instruments, the operating conditions are set so that the probability of detecting a single photon in each excitation period is much less than one. In other words, if *RPCR* is the average photon counting rate (PCR) and *RREP* laser pulse repetition rate, then *RREP*/*RPCR* >> 1. If this condition is not met, then photon ‘pile-up’ occurs; a phenomenon whereby the apparent measured fluorescence lifetimes become shorter than the real lifetimes [24],[28]-[31],[41]. The measured decays become distorted to shorter times because many photons arrive during an excitation period, but only the first arriving photon is counted due to the dead-time of the detector.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑2: (a) Principle of operation of TCSPC. In excitation period *N1*, the second arriving photon is missed due to dead-time, resulting in pulse pile-up at the detector. (b) A generalized TCSPC measurement set-up. A TDC measures the inter-arrival time (IAT) between laser pulses and photon arrivals and the results are stored in a histogram.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑4: a) Principle of operation of TCSPC. In excitation period *N1*, the second arriving photon is missed due to dead-time, resulting in pulse pile-up at the detector. b) A generalized TCSPC measurement set-up. A TDC measures the inter-arrival time (IAT) between laser pulses and photon arrivals and the results are stored in a histogram.

As general rule of TCSPC, the PCR in a single TCSPC channel is typically limited to 1-10% of the repetition rate of the laser to reduce the pile-up distortion [1]-[3],[24]-[35]. The high-resolution TDCs required for TCSPC FLIM typically introduce the most processing dead-time during which they are unable to process any photon arrival events immediately following a photon detection event [28]. The single photon detector also has a dead-time, but it is typically much shorter than the TDC’s [31]. Multi-channel TCSPC architectures that utilize multiple SPAD/TDCs to process many photon events concurrently have been shown to considerably improve the maximum PCR that can be measured without pile-up distortion [24]-[35]. The approach taken in this work is to develop prototype TDC chips with low dead-time and compact size that can operate in parallel, thereby reducing pileup effects and realizing high-sensitivity, high-speed, low-power, miniaturized FLIM systems. The design of a prototype TDC chip in IBM 130 nm CMOS technology for such purposes is covered in Chapter 6.

A generalized TCSPC set-up is shown in Fig. 2-2(b) [1]-[3]. A picosecond-pulsed laser source is used to excite the fluorescence in the sample under analysis. The fluorescence photons are delivered to a single-photon detector through an optical system that includes filters to remove the excitation light and attenuators to decrease the photon flux at the detector to obtain a PCR that’s below the pile-up limits of the detection system. A single-photon detector, such as a photomultiplier tube (PMT) or a SPAD can be used to detect the fluorescent photons with picosecond level accuracy and precision.

The arrival time of the first photon detected within the excitation period is measured by a TDC which outputs a digital code representing the photon arrival time (STOP) relative to the laser pulse (START). The least-significant-bit time resolution *TLSB* of the TDC represents the smallest inter-arrival time (IAT) of the histogram where the fluorescence photon arrival times are built up. The envelope of this histogram after many photon arrival events provides the fluorescence decay profile from which the lifetime parameter is estimated [1]-[3],[24],[25].

The number of photon counts in each time-bin is distributed according to Poisson statistics. Therefore, the SNR is proportional to the square root of the total number of counts in the histogram [24]. As such, TCSPC is the most accurate and efficient method in determining fluorescence lifetimes and has been most widely implemented with confocal and two-photon laser scanning microscopies, whereby a focused laser beam is scanned across the sample and the resulting fluorescence lifetime is measured at each position [26]-[29] ,[31]-[35].

State-of-the art TCSPC systems are able to record two-dimensional time-resolved fluorescence images with mega-pixel resolution and are tolerant to dynamic changes in the fluorescence decay lifetime [34],[35]. Commercial TCSPC modules also feature the time-tagged-time-resolved (TTTR) mode used in studying the dynamics of fluorescence lifetimes where the arrival times of the photons are recorded with respect to a fixed reference time [26]-[29]. The availability of the time-tags permits identification of photon bursts in real-time, which is useful for various different analysis, for example, in single-molecule spectroscopy (SMS) and fluorescence correlation spectroscopy (FCS) [36]-[38].

There are several important limitations of TCSPC FLIM. First, the image acquisition time can be very long if the number of pixels of the image is very large, the decay is multi-exponential, very high lifetime accuracy is required, and the count rate available from the sample is low. These conditions are typical in autofluorescence imaging of living cells and tissues [39]. The imaging speed can be increased by using multiple detectors and TCSPC channels in parallel [24]-[35]. However, if all the TCSPC electronics are not integrated on a single IC, but are rather available only as discrete printed circuit boards (PCB), then the cost of a multichannel TCSPC system rises considerably and miniaturization of the system is precluded [28].

Another limitation of FLIM is that the measured fluorescence decay is not the true fluorescence decay due to the practical constraints of using a laser pulse of finite duration. The true fluorescence decay can only be obtained by expressing the measured decay as the convolution of an exponential decay model and the measured instrument response function (IRF), which is comprised of laser jitter the intrinsic timing jitter of the detector [28]. Iterative linear or non-linear least-square methods (LSM) such as Marquardt-Levenberg algorithms are commonly used to accurately extract the real decay lifetimes, whereby the convolution of the decay model and the IRF is compared to the measured data and the residual error is reduced over many iterations [1]-[3], [24],[25],[40]. However, the high-speeds required for pixel read-out and the complexity of the data analysis for the determination of fluorescence lifetime remain the main bottlenecks for the realization compact high-speed, low-power TCSPC FLIM systems [42]-[48].

### Time Gated Single-Photon Counting (TGSPC) FLIM

An alternative approach for measuring fluorescence lifetimes in the time-domain makes use the time-gating (TG) technique, conceptually illustrated in Fig. 2-3(a) [1]-[3],[49]-. In TG mode, the fluorescence emission is measured by counting photons in two or more detection gates that are delayed by different amounts relative to the excitation pulse. By placing the detection gates in the appropriate time-delay positions with respect to the excitation, the lifetime of the decay curve can be extracted by simple photon counting during each time gate. Gated detection is accomplished in practice by turning the detector on for a short period of time during the intensity decay and repeatedly measuring the photon counts using different delays for the gate-on time, as is depicted in Fig. 2-3(b) [49]-[56]. Alternatively, the detector can work in free-running mode with the counters briefly enabled at different delays with respect to the excitation pulse [57]-[60].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑3: (a) Principle of operation of TGSPC. (b) Generalized TGSPC measurement set-up.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑6: a) Principle of operation of TGSPC b) Generalized TGSPC measurement set-up.

In the TG mode of operation, a histogram of the fluorescence decay curve is obtained directly by measuring photon counts in sequential time intervals. Compared with the hardware for TCSPC, the complexity of the time-gating acquisition systems is relatively simpler, consumes less power, and is suitable for array implementation. The measurements are less affected by the pile-up nonlinearities as well [49]. TG FLIM does not require a TDC, so it do not have the same dead-time limitations and may be used with higher photon fluxes to obtain faster imaging rates [21],[23].

Another benefit of TGSPC is that it does not use a computationally expensive evaluation of fluorescence lifetimes. Rather than recording a complete decay curve and estimating the fluorescence lifetime using iterative least-squares methods as in TCSPC, the lifetimes of single-exponential decays can be obtained very rapidly by using rapid lifetime determination (RLD) methods [3],[21],[23],. However, prior knowledge of the decay profile (single or multi-exponential) is required in order to accurately estimate the lifetimes using RLD [21]-[23],[77]-[82]. In addition, TGSPC typically has worse time resolution compared to TCPSC because the lifetime accuracy of the former depends on gate duration rather than detector jitter for the latter. Photon counting efficiency is also lower, since all the fluorescence photons falling outside the gating time window are rejected.

In spite of the inherently reduced photon efficiency resulting from the application of narrow time gates necessary to count photons with high temporal resolution, TGSPC has been widely used for wide-field FLIM imaging [1]-[3], [77]-[82]. Wide-field FLIM enables parallel image acquisition and faster lifetime imaging compared to TCSPC since the fluorescence is simultaneously collected for all the pixels by the imaging optics, rather than time-consuming laser scanning. Achieving real-time frame rates is more difficult with wide-field TCSPC implementations since the IAT information of every detected photon in every pixel must be read-out, requiring very high data rates which results in very high power consumption [42]-[48]. However, whereas TCSPC is essentially limited by photon shot-noise, the detectors traditionally used in wide-field FLIM (Gated Optical Intensifiers [18]) introduce excess noise to the measurements [29],[79]-[82]. Therefore, time-gating techniques require complex optimization of the acquisition parameters in order to minimize the error when extracting fluorescence lifetimes, especially when multi-exponential decays are encountered

## Existing single-photon detector technologies

Photodetectors capable of detecting single photons in the visible (VIS) (400 nm – 700 nm) spectral range with sub-nanosecond temporal resolution are very critical in single-photon imaging [87]-[92]. Besides FLIM, other important biomedical applications that need single-photon detectors include single-molecule spectroscopy (SMS), [36],[88],[91],[93]-[95], DNA sequencing [96],[97] and time-of-flight (ToF) positron emission tomography (PET) [98]-[100]. Single-photon counting and timing techniques are also widely used in other applications that require measurement of very weak and very fast optical signals, such as in light detection and ranging (LIDAR) [66]-[73], optical fiber characterization [101], optical communications [102], Raman spectroscopy [53],[54],[56],[103]-[105] and high-energy physics experiments [106]-[109].

Understanding the strengths and weaknesses of existing technologies used for single-photon detection is important to obtain a deeper understanding of the overall technical requirements of different single-photon imaging applications. Therefore, in this subsection, the development of single-photon detectors will be reviewed. An outline of the key performance characteristics of time-resolved single-photon imaging systems is also provided.

### Photo-multiplier Tube (PMT)

The structure of a PMT, illustrated in Fig. 2-4(a), generally consists of a photosensitive material (a photocathode) that converts the incident photons into photoelectrons by utilizing the photoelectric effect [106],[110]. Initially, the photoelectrons have low kinetic energy, but they gain kinetic energy during their transit towards the first dynode (electrode). The bombardment of the dynodes by these high-energy electrons results in the production of multiple secondary electrons from each incident electron through the process of secondary emission. The electrons multiply in number when striking successive dynodes in the chain which are biased at progressively higher voltages. With operating voltages in the range of 100-10kV, PMTs can achieve multiplication gains of 105-1010 [111]. The very high gain results in current pulses at the anode which are easily distinguishable from the noise of the subsequent front-end circuits.

Ideally, each PMT output pulse represents the absorption of a single photon by the photocathode. However, output pulses are also occasionally produced by electrons randomly emitted from the photocathode and dynode materials through the process of thermionic emission and field emission, representing the dark count rate (DCR) of the PMT. The DCR of typical PMTs at room temperature can vary from tens to thousands of counts per second, and is heavily dependent on the cathode material and the design of the dynode chain [26]-[28],[106],[111]. The choice of photocathode material is crucial in determining not only the DCR performance, but the photon detection efficiency (PDE) as well. PDE is determined by the internal quantum efficiency (QE) of the photocathode, which is the probability of a photoelectron emission per incident photon. For single-photon detectors, the PDE is smaller than the QE because not every photoelectron produces a detectable anode current pulse [111]. Photocathodes composed of traditional bialkali and multialkali materials only attain a QE of 20-25% in the 300-500 nm wavelength range, while GaAsP cathodes extend the usable range above 700 nm [24],[28],[106],[111]. Since PMTs rely on external photoelectron emission in a vacuum, they inherently have lower PDE compared to solid-state detectors, which rely on internal photoelectric conversion and avalanche multiplication in a semiconductor material [110].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑4: (a) Simplified structure of a conventional PMT and examples of the signals measured at the anode. [26]. (b) Photon Detection Efficiency (PDE) comparison of different photocathode materials [28].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑8: (a) Simplified structure of a conventional PMT and examples of the signals measured at the anode. [26]. (b) Photon Detection Efficiency (PDE) comparison of different photocathode materials [28].

The timing characteristics of PMTs are described by the electron transit time response (TTR) and the transit time spread (TTS). The TTR is the average time difference between arrival of a photon at the photocathode and the collection of the subsequent current pulse at the anode. TTS is the standard deviation or, full-width at half-maximum (FWHM) of the transit time distribution, also known as time resolution or timing jitter [106],[111]. The TTS is critical for TCSPC, since it represents the uncertainty of the detector’s evaluation of the photon arrival time. Currently, Micro-Channel Plates (MCP) offers the best timing resolution, with TTS below 30 ps FWHM [111],[112]. In place of a dynode chain, MCPs contain dense array of micron-sized holes coated with a secondary emissive dynode material to minimize the distance for electron travel and restrict the range of electron paths [28],[111].

MCPs are used in the most demanding time-resolved measurements and have several other advantages compared to conventional PMTs. These advantages include more compact size, two-dimensional detection capability (spatial resolution) and stability in high magnetic fields [29]. MCPs are key components of Gated Optical Intensifiers (GOI) used routinely in FLIM [18]. However, MCPs can be easily damaged by high light levels, and the maximum counting rate of is about ten to a hundred times lower than that of fast PMTs, which limits the dynamic range of MCP based systems [111].

PMTs and MCPs have very large photosensitive areas that have lower noise per unit area compared to solid-state detectors. Circular photocathodes can range in diameter from 1 mm to over 20 cm with DCR in the range of 0.1–1 kHz [111]. The large photosensitive area greatly simplifies the design and alignment of confocal imaging geometries. Low noise is crucial in attaining high sensitivity which improves the SNR in fluorescence lifetimes measurements [88]. On the other hand, PMTs are bulky and fragile, sensitive to electromagnetic disturbances and mechanical vibrations, require high supply voltages (2–3 kV) and are expensive. These disadvantages clearly preclude their practical usage in miniaturized arrays for mass-produced, ‘in-field’ applications which require detectors with lower power and lower cost.

Another disadvantage of PMTs is the complexity of the external electronics required. These circuits typically consist of a preamplifier, constant-fraction discriminator (CFD), time-to-amplitude converter (TAC) and analog-to-digital converter (ADC) and are usually available only as discrete ICs which further drives up the power, size and cost of the overall system [28]. For these reasons, a number of solid-state solutions have been proposed as a replacement for PMTs [87]-[92].

### Charge-Coupled Devices (CCD)

CCDs operate by means of electron-hole pair generation by the photoelectric effect in a semiconductor [113],[114]. The generated charges are collected into localized packets by the pn junction potential wells associated with pixel arrays of metal–oxide–semiconductor (MOS) capacitors. As illustrated in Fig. 2-5(a), during image read-out, the charge packets are transferred serially between pixels by manipulating the voltages on the gates of the capacitors to allow the charges to couple from one capacitor to the next.

Unlike PMTs, CCDs were not envisioned for time-resolved single-photon detection applications [114]. On the contrary, the operational principle of conventional CCDs precludes detection of single photons with high temporal resolution. CCDs are inherently integrating devices, where the amount of charge accumulated in each pixel is proportional to the number of incident photons within the fixed exposure time. Since photo-generated carriers need to be integrated within internal capacitors to form a voltage signal, very long integration times are required to detect very weak light.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑5: (a) Schematic representation of charge transfer in a CCD [27]. (b) Principle of operation of an EM CCD [121].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑10: (a) Schematic representation of charge transfer in a CCD [27]. (b) Principle of operation of an EM CCD [121].

The main difficulty of using CCDs for LLL imaging, besides the long integration time, is the thermal noise and electronic read-out noise, which sets the limit on the minimum detectable optical power. Thermal noise is inherent in semiconductors and occurs through the thermal generation of minority carriers, resulting in charges accumulating in CCD pixels in the absence of illumination when operated at any temperature above absolute zero (0 K) [114]-[116]. The carriers are generated through thermally activated generation-recombination (GR) defect centers associated with imperfections or impurities within the bulk of the semiconductor crystal or at the surface interface [115]. These defect states introduce energy levels into the forbidden bandgap that promote dark current by acting as ‘steps’ in the transition of electrons and holes between the conduction and valence energy band.

In a sufficiently cooled CCD, dark current is insignificant so the optical sensitivity is limited only by read-out noise, which primarily depends on the input-referred noise spectrum of the output stage source-follower (SF) transistor [114]-[116]. The readout noise increases with the read-out bandwidth, so slower read-out rates are used to achieve higher sensitivity. Dark currents of less than 0.02 electrons per pixel per second at –60 °C, and read-out noise levels of 10 electrons root-mean-square (rms) at 1 MHz, have been reported for scientific CCDs [117]. However, even at the lowest readout rates, the readout noise of standard CCDs is still too high for detection of single photons.

Electron multiplying CCD (EM CCD) (also known as LLL CCDs) utilize a multiplication register where photogenerated charges are amplified by impact-ionization in order to increase the signal level by two or three orders of magnitude above the read-out noise level, thus enabling single photon detection capability, [27],[118]-[121]. In an impact-ionization process, electrons accelerated by the electric field in a pn junction collide with valence-lattice electrons and generate additional free electron-hole pairs [115]. These pairs are also accelerated and generate additional pairs, so the ensuing avalanche current is much greater than the photocurrent without avalanche multiplication. However, impact-ionization is a random process, so the SNR decreases for large values of multiplication gain [120],[121].

Another important source of noise in EM CCDs is the clock-induced-charge (CIC) noise due to the generation of spurious charges during charge transfer. In conventional CCDs, these dark carriers are hidden within the read-out noise, but in an EM CCD, they are multiplied by the gain register and are indistinguishable from true photo-electrically induced electrons [27]. Since CIC is temperature independent, this noise source represents a fundamental limitation on the maximum attainable frame rate (typically several tens of MHz) of EM CCDs. So although they can detect single photons, EM CCDs cannot be used in applications requiring sub-nanosecond photon arrival-time information [27].

EM CCDs also require a specialized silicon fabrication process. In such a process, integration of on-chip ancillary circuits, such as frequency synthesizers, flash memory, microcontrollers, signal processors, and ADCs is generally not possible. This precludes cost-efficient mass-production of CCD-based imagers. It also leads to an inherent incompatibility with the requirements of low-power and miniaturized time-resolved single-photon imaging systems.

### Single-photon avalanche diode (SPAD)

An SPAD is the solid-state analogue to a Geiger-Muller detector, whereby the gain mechanism is avalanche breakdown of an avalanche photodiode (APD) rather than breakdown of a gas-filled diode [29],[90]-[92],[106]. When operated in the conventional linear mode at room temperature, APDs do not have enough internal gain to be used as single-photon detectors [122],[123]. However, when the APD is biased at an excess voltage *VEX* above the breakdown voltage *VBR*, the avalanche gain becomes infinite and the APD is said to operate in the Geiger-mode [108],[124]-[126].

In Geiger mode, the electric field in the pn junction’s depletion region becomes so high that both electrons and holes generated in the depletion region can undergo impact ionization, which results in avalanche breakdown of the device [127]-[129]. As depicted in Fig. 2-6(a), minority carriers travelling through the high-field depletion (multiplication) region can acquire enough kinetic energy to undergo scattering events with bound electrons in the valence band (VB). The VB electrons are promoted into the conduction band (CB), thus creating new electron-hole pairs of free charges which can further undergo avalanche multiplication, resulting in an exponential growth of the number of carriers.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑6: Principle of SPAD operation (a) Avalanche breakdown process in a reverse biased pn junction. (b) Load-line representation of SPAD operation [72].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 2‑12: Principle of SPAD operation (a) Avalanche breakdown process in a reverse biased pn junction. (b) Load-line representation of SPAD operation [72].

The gain of an SPAD is comparable to that of a standard PMT, but the timing resolution is typically much higher since the multiplication region is much thinner. Also, the higher the SPAD is biased above breakdown by *VEX*, the faster the turn-on transient of the avalanche current becomes, typically lasting tens of picoseconds [130],[131]. The timing jitter of an SPAD in response to a single-photon is thus very low, typically in the range of 10-100 ps for devices that are optimized for single-photon timing performance [132],[133].

Fig. 2-6(b) illustrates the operating principle of an SPAD. Initially, when the SPAD is biased above breakdown, the SPAD stays in an OFF state (1) for a short time until a photogenerated charge (electron or hole) enters the multiplication region and triggers an avalanche breakdown, bringing the device into its high-current ON state. Charges generated by means other than photon absorption can also trigger avalanches, representing the DCR of the SPAD. When triggered, the avalanche current reaches the peak value in a few picoseconds before the SPAD voltage has changes. Then, the bias point thus moves from point (1) to point (2). The resulting avalanche current passes through a quenching resistor providing a negative feedback, which causes a voltage drop across the SPAD due to the current increase (point (3)). At this point, a fast discriminator is used to sense the voltage drop and provide an output pulse that very precisely marks the photon arrival time if the first charge was generated by photon absorption [134]-[137]. The voltage drop across the quenching resistor causes the SPAD bias voltage to quickly drop to below the breakdown voltage, at which point the avalanche is no longer self-sustaining [125],[135]. When this occurs, the avalanche is fully quenched and the SPAD voltage begins recharging back to (*VBR* + *VEX*) through the quenching resistor connected to the supply voltage *VSPAD*.

The SPAD’s recovery time after each avalanche breakdown is determined by the quenching resistance and the SPAD’s capacitance. The main drawback of a passively quenched SPAD is the relatively long recovery time, which limits the maximum light intensity that the SPAD can measure before the detector saturates [138]. The slow voltage recharge causes another effect. When avalanches are triggered early during the reset transition, the excess bias is not fully recovered and the resulting output pulse may be too small to be detected. In order to overcome the drawbacks of passive quenching, an active quenching/reset (AQR) front-end circuit is commonly adopted [49]-[56],[125],[126],[132] -[137]. The onset of the voltage discharge is sensed as early as possible by the AQR circuit and the bias voltage is reduced below the breakdown voltage through the quenching circuitry so as to ensure complete termination of the avalanche current. The bias voltage is subsequently restored its original value by the reset circuit so the SPAD is ready to detect the next photon. The AQR circuit reduces the chance of having an avalanche occurring while the SPAD is in the recharge state, leading to improved performance and higher counting rates [138].

SPADs fabricated in custom silicon processes have reached levels of performance comparable to PMTs [132]-[135],[139]-[144]. The first generation of commercial SPAD modules were based on the ‘reach-through avalanche structure’, built in special ultra-pure high-resistivity silicon wafers with a dedicated technological process featuring thick depleted regions to ensure that the PDE was very high [108],[124]-[126] On the other hand, various features of the technology, such as high operating voltage (>100 V), high power dissipation (~10 W), and wavelength-dependent IRF in the range of 200–600 ps FWHM limited the prospects of multichannel development with this reach-through structure.

The current generation of SPAD devices fabricated in custom silicon planar epitaxial technology have a high photon-timing resolution (between 35–100 ps FWHM) and can also achieve high PDE (~75 %) in the UV and NIR range [124]-[126], [132]-[135],[139]-[144]. Active area diameters can range from 5 to 100 μm, and dark-counting rate down to tens of Hz are achievable when the SPAD is thermoelectrically cooled [141]. However, multichannel implementations are still limited to only a few channels and by high power dissipation (25 W) [126],[142]-[144]. Therefore, custom-silicon SPADs are not suitable for miniaturized, low-cost and versatile single-photon counting systems and there is no perspective of monolithic integration with ancillary circuits in order to obtain robust, fully-integrated single-photon imaging systems.

## CMOS SPADs

The commercially available detectors (e.g. PMTs, MCPs, custom-silicon SPAD) can provide the required single-photon sensitivity and picosecond-level time resolution. However, they are large and expensive devices that do not meet key requirements in the development of compact and economical detectors. For low-cost, compact devices, robust and inexpensive single-photon sensitive OEICs are needed. CMOS SPAD imaging arrays offer comparable performance at lower costs, lower power and smaller size compared to competing technologies. As a result, there are numerous applications being pursued for CSPAD single-photon imaging systems [61],[62].

In the following sub-sections, the main performance specifications of CSPADs will be summarized and reviewed, and the main research challenges in their development will be discussed.

### Dark Count Rate (DCR)

SPADs produce digital pulses upon the detection of individual photons and are therefore practically immune to read-out noise. Due to the absence of readout noise, when the illumination is very low, the sensitivity is only limited by the DCR. There are several carrier generation mechanisms responsible for the DCR and the relative contribution of each mechanism is dependent upon the technological factors related to the SPAD design and fabrication details in a CMOS process [145]-[177]. The main mechanism responsible for dark counts in SPADs at room temperature is the thermal generation of free-carriers within a diffusion length of the SPAD’s depletion region. Due to the relatively large bandgap of silicon, direct thermally-activated transitions of electrons from the valence to the conduction band are unlikely. However, according to the Shockley-Read-Hall (SRH) theory of recombination, defects that occur within the crystal will disrupt the perfect periodic potential function, thus creating discrete GR energy levels within the forbidden energy band [114]-[116]. These GR levels make it possible for electrons to become thermally excited to the conduction band. If this occurs in or near the avalanche multiplication region, a thermally-generated dark pulse may be triggered.

Since the SRH generation rate increases exponentially with temperature, its contribution may be reduced by using cooling methods such as thermoelectric Peltier elements or by forced air-cooling. The temperature dependence of DCR is therefore a very important indicator of an SPAD’s performance. Fig. 2-7 shows a DCR density figure-of-merit (FoM) (in units of Hz/(μm2√V)) plotted as a function of temperature for reported SPAD pixels in different CMOS technologies [61]. Typically, the DCR values are reported at different excess voltages. Therefore, the DCR density FoM was normalized by the electric field strength for a fair comparison. When the temperature is reduced, thermal generation is significantly reduced and tunneling becomes the dominant contributor to the DCR. Tunneling occurs when the electric field across a strongly reverse-biased p-n junction approaches 106 V/cm, resulting in a significant flow of electrons from the valence band to the conduction band [115],[129]. As the junction dimensions scale down, the doping concentrations increase, the depletion width decreases, and the junction doping profiles become more abrupt. Hence SPADs in DSM technology have DCR dominated by tunneling [145]-[147],[165]-[172]. As a result, DCR declines less rapidly at lower temperatures since the tunneling probability is weakly temperature dependent.

|  |
| --- |
|  |

Figure 2‑7: Summary of CMOS SPAD DCR temperature dependence.

|  |
| --- |
|  |

Figure 2‑14: Summary of CMOS SPAD DCR temperature dependence.

|  |
| --- |
|  |

Figure 2‑8: Summary of CMOS SPAD DCR excess bias dependence.

|  |
| --- |
|  |

Figure 2‑16: CMOS SPAD DCR Bias dependence summary.

An exponential dependence on excess bias is expected when tunneling is the dominant DCR mechanism. In Fig. 2-8, DCR is plotted as a function of excess voltage to investigate tunneling effects at room temperature of reported SPADs. The thermal-induced DCR tends to saturate at large excess voltages for the SPADs in HV technology [160]-[162] (according to its dependence on the avalanche triggering probability [131],[155]). On the other hand, the trend for DSM technology is a DCR increasing exponentially with excess voltage [165]-[172].

### Afterpulsing

In addition to thermal generation and tunneling, the DCR of SPADs includes statistically correlated avalanches that are due to the carrier trapping phenomenon. The impurities and defects in the space charge region tend to capture carriers that are generated during each avalanche breakdown. The traps de-excite exponentially in time, releasing carriers at random time intervals [178]-[180]. If the excess bias of the SPAD is fully restored before all the carriers in the trapping levels have been released, then secondary avalanches known as afterpulses are triggered. Afterpulsing becomes the dominant source of DCR when SPADs are cooled to reduce the thermal DCR, since the trapping lifetimes increase with temperature [155],[178]-[180].

To minimize afterpulsing, the bias should be kept below breakdown voltage for a sufficiently long time after quenching so that the trapped carriers can depopulate without triggering afterpulses [153],[138]. However, prolonging the dead-time to eliminate afterpulsing limits high-speed operation and introduces nonlinearity of the expected photon count rate with increasing incident light intensity [138]. The number of filled traps mainly depends upon the avalanche current density and its duration. Therefore, minimizing the charge flow during an avalanche by using optimizing SPAD front-end circuits is the most efficient way of reducing afterpulsing [138], [163] [181].

It is possible to gain a deeper insight into the impact of processing technology and front-end circuit design has on afterpulsing performance. In Table I, a summary of the most important parameters for SPAD front-end circuits highlights the beneficial effects integrated AQR circuits have on reducing the afterpulsing probability when a sufficient hold-off time is introduced. For 314 and 60 μm2 active area SPADs, the minimum required hold-off times were 100 and 40ns, respectively [163],[181]. These results represent state-of-the-art afterpulsing performance for CMOS SPAD. These SPADs were manufactured in advanced CMOS technology and they incorporate optimized front-end circuits for avalanche sensing and quenching with minimal afterpulsing.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CMOS  (nm) | Pixel  Type | # of FETs | SPAD area  (μm2) | FF  (%) | DT (ns) | AP (%) | REF |
| 90 | PQPR | 6 | 12.56 | n/a | 90 | 3.7 | [174] |
| 90 | PQPR | 5+CC | ~32 | n/a | 15 | 0.38 | [175] |
| 130 | PQPR | 4 | 81 | 6 | 27 | 4 | [86] |
| 130 | PQPR† | 5 | 110 | 10 | 30 | 0.1 | [86] |
| 130 | PQPR | 6 | 50.3 | n/a | 100 | 0.02 | [169] |
| 130 | AQAR | 18 | 50.3 | n/a | 5.4 | 1.28 | [138] |
| 130 | AQAR | >9 | 20.7 | 0.77 | 10 | 0.2 | [47] |
| 130 | AQAR | 9 | 50 | 10\*\* | 100 | 0.1 | [44] |
| 150 | PQPR | >6 | 78.5 | n/a | 30 | 1.3 | [165] |
| 180 | PQAR | 16 | 59.8 | n/a | 6 | ~0 | [163] |
| 350 | VLQC | 16 | 314 | 2.8 | 20 | 1.3 | [181] |
| 350 | AQAR | >8 | 400 | 1.5 | 500 | 4.5 | [50] |

Table I– Different reported SPAD pixels with low reported afterpulsing (AP) performance at room temperature.

\*Front-end circuit placed outside of pixel \*\*FF recovered by microlenses †time gated.

PQPR – passive quench, passive reset; AQAR – active quench, active reset

PQAR– passive quench, active reset; VLQC – variable-load quenching circuit

CC – coupling capacitor

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CMOS  (nm) | Pixel  Type | # of FETs | SPAD area  (μm2) | FF  (%) | DT (ns) | AP (%) | REF |
| 90 | PQPR | 6 | 12.56 | n/a | 90 | 3.7 | [174] |
| 90 | PQPR | 5+CC | ~32 | n/a | 15 | 0.38 | [175] |
| 130 | PQPR | 4 | 81 | 6 | 27 | 4 | [86] |
| 130 | PQPR† | 5 | 110 | 10 | 30 | 0.1 | [86] |
| 130 | PQPR | 6 | 50.3 | n/a | 100 | 0.02 | [169] |
| 130 | AQAR | 18 | 50.3 | n/a | 5.4 | 1.28 | [138] |
| 130 | AQAR | >9 | 20.7 | 0.77 | 10 | 0.2 | [47] |
| 130 | AQAR | 9 | 50 | 10\*\* | 100 | 0.1 | [44] |
| 150 | PQPR | >6 | 78.5 | n/a | 30 | 1.3 | [165] |
| 180 | PQAR | 16 | 59.8 | n/a | 6 | ~0 | [163] |
| 350 | VLQC | 16 | 314 | 2.8 | 20 | 1.3 | [181] |
| 350 | AQAR | >8 | 400 | 1.5 | 500 | 4.5 | [50] |

Table II– Different reported SPAD pixels with low reported afterpulsing (AP) performance at room temperature.

\*Front-end circuit placed outside of pixel \*\*FF recovered by microlenses †time gated.

PQPR – passive quench, passive reset; AQAR – active quench, active reset

PQAR– passive quench, active reset; VLQC – variable-load quenching circuit

CC – coupling capacitor

### Breakdown Voltage (BV)

Breakdown voltage is an important parameter which reveals important information about the mechanisms involved in junction breakdown. As the doping concentration increases when CMOS technology scales down, the junctions become shallower and the radii of curvature decreases, leading to premature edge breakdown (PEB) [146]-[148],[157],[182],[183]. DSM SPADs exhibit lower breakdown voltage and markedly higher DCR contribution from tunneling compared to those in HV CMOS with higher breakdown voltages, because of the higher doping concentrations and narrower depletion widths [166]-[177]. In Figs. 2-7 and Table II, the best performing devices in terms of DCR had breakdown voltages in the range of 23.1–27.5 V, while those with large DCR had breakdown voltages in the range of 9.4–11.4 V.

The temperature coefficient of breakdown voltage is another important parameter in assessing the SPAD’s performance. The breakdown voltage of SPADs increases with temperature because the phonon scattering increases at higher temperatures, which makes it more difficult for electrons or holes to achieve the threshold energy needed to provide an elementary act of impact ionization that leads to avalanche breakdown [115]. Higher breakdown voltage temperature coefficients as in [164],[172] indicate that the DCR is composed of mainly thermal generation. On the other hand, those with a lower temperature dependence [166],[177] have a higher tunneling contribution. Breakdown voltages and their temperature coefficients as well as DCR are summarized in Table II for SPADs reported in different technologies.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CMOS  (nm) | SPAD  Type | BV (V) | BV Coeff.  (mV/°C) | DCR/μm2  (Hz/μm2) | REF |
| 65 (dig.) | shallow – n+/p-well | 9.1 | 5 | 106/64 @ *VEX* = 0.4 V | [177] |
| 90 (CIS) | buried – p-well/DNW | 17.5 | 13.1 | 7.4/12.56 @ *VEX* = 0.8 V | [174] |
| 130 (dig.) | shallow – n+/p-well | 11.4 | 7.2 | 104/81 *VEX* = 1.3 V | [86] |
| 130 (dig.) | shallow – p+/n-well | 9.7 | 8 | 104/78.5 @ *VEX* = 1 V | [166] |
| 130 (CIS) | buried – p-well/DNW | 17.9 | 6.7 | 40/50.3 @ *VEX* = 0.8 V | [172] |
| 130 (CIS) | buried – p-well/DNW | 14.36 | 3.3 | 25/50.3 @ *VEX* = 0.8 V | [172] |
| 130 (CIS) | shallow – p+/n-well | 12.4 | 20 | 47/50.3 @ *VEX* = 0.8 V | [172] |
| 150 (dig.) | buried – p-well/DNW | 23.1 | - | 90/78.5 @ *VEX* = 1 V | [165] |
| 150 (dig.) | shallow – p+/n-well | 16.1 | - | 30/78.5 @ *VEX* = 1 V | [165] |
| 180 (CIS) | shallow – p+/n-well | 21 | 40 | 30/78.5 @ *VEX* = 1 V | [164] |

Table II – Summary of CSPAD breakdown voltages and DCR performance at room temperature for different technologies.

Dig. – standard digital process

CIS – CMOS Image Sensor process

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CMOS  (nm) | SPAD  Type | BV (V) | BV Coeff.  (mV/°C) | DCR/μm2  (Hz/μm2) | REF |
| 65 (dig.) | shallow – n+/p-well | 9.1 | 5 | 106/64 @ *VEX* = 0.4 V | [177] |
| 90 (CIS) | buried – p-well/DNW | 17.5 | 13.1 | 7.4/12.56 @ *VEX* = 0.8 V | [174] |
| 130 (dig.) | shallow – n+/p-well | 11.4 | 7.2 | 104/81 *VEX* = 1.3 V | [86] |
| 130 (dig.) | shallow – p+/n-well | 9.7 | 8 | 104/78.5 @ *VEX* = 1 V | [166] |
| 130 (CIS) | buried – p-well/DNW | 17.9 | 6.7 | 40/50.3 @ *VEX* = 0.8 V | [172] |
| 130 (CIS) | buried – p-well/DNW | 14.36 | 3.3 | 25/50.3 @ *VEX* = 0.8 V | [172] |
| 130 (CIS) | shallow – p+/n-well | 12.4 | 20 | 47/50.3 @ *VEX* = 0.8 V | [172] |
| 150 (dig.) | buried – p-well/DNW | 23.1 | - | 90/78.5 @ *VEX* = 1 V | [165] |
| 150 (dig.) | shallow – p+/n-well | 16.1 | - | 30/78.5 @ *VEX* = 1 V | [165] |
| 180 (CIS) | shallow – p+/n-well | 21 | 40 | 30/78.5 @ *VEX* = 1 V | [164] |

Table IV – Summary of CSPAD breakdown voltages and DCR performance at room temperature for different technologies.

Dig. – standard digital process

CIS – CMOS Image Sensor process

### Photon Detection Efficiency (PDE)

The PDE is a measure of the ratio of the number of detected photons to the number of incident photons. It is the product of the geometric fill-factor (the ratio of photo-sensitive area to total imaging or pixel area), absorption probability, and avalanche triggering probability [155],[184],[185]. CMOS technology parameters play an additional and significant role in determining PDE performance. In CMOS technology, impinging photons must pass through a thick layer of passivation that is put on top of the chip at the end of the chip fabrication process to protect it from external contaminants [186],[187]. Then, the light must pass through several dielectric layers with different refractive indices, possibly undergoing constructive and destructive interference in the insulating films directly above the active silicon surface [188]. CMOS Image Sensor (CIS) technology facilitates optimization of the dielectric above the active region to minimize reflections [189], but standard CMOS digital/RF technologies do not offer such options. Post-processing can be used to etch the top passivation layers, and this has shown to lead to a marked improvement in PDE performance [159].

|  |
| --- |
|  |

Figure 2‑9: Summary of CMOS SPAD PDE performance.

|  |
| --- |
|  |

Figure 2‑18: Summary of CMOS SPAD PDE performance.

Photons that reach the silicon surface must be able to strike the photosensitive area of the detector for avalanche breakdown to take place. However, only a fraction of the pixel area is photosensitive; associated electronic circuits occupy the remaining area. Pixels having smaller fill-factor have lower sensitivity, since a large portion of impinging photons do not strike the photosensitive area. Higher fill-factor can be attained by increasing the SPAD’s active area [50],[161], using smaller sized transistors that are feasible technology downscaling [171],[176],[177] , reducing the in-pixel transistor count [60], and using n-well sharing among SPADs [50],[51]. Micro-optical concentrators (micro-lenses) have also been successfully used to recover some of the lost sensitivity due to small FF [54],[60],[190],[191].

SPADs in standard CMOS technology can be made from shallow junctions or deeper junctions, depending on the technological options available [147],[165],[172],[173]-[175] Photons with longer wavelengths have a higher probability to be absorbed deeper in the substrate. Thus, deeper junctions are more sensitive to longer wavelengths as compared to shallower junctions closer to the semiconductor surface. SPADs fabricated in CMOS technologies with shallow active regions mostly detect photons absorbed near the silicon surface, resulting in a PDE response suitable for near-UV/blue incident light [162],[170]. However, SPADs with shallow-junctions have thinner active regions due to the higher doping concentrations used, thus their PDE is typically smaller than the deep junctions SPADs over the excess bias voltage range.

A deep-junction SPADs structure compatible with backside illumination was proposed in a 90 nm CMOS technology. Here the active junction was between the deep n-well and the low-resistivity p-substrate [175]. A shift towards red wavelengths of the peak detection efficiency was observed. This was because the active region was buried deep in the substrate where photons with longer wavelength are more likely to be absorbed. Fig. 2-9 illustrates the trend of PDEs of several different SPADs in CMOS technologies. As the absorption depth in silicon drastically increases for wavelengths beyond 1100 nm, then silicon SPADs are rendered transparent to light with λ > 1100 nm.

### Timing Resolution

Time-resolved applications such as FLIM, ToF PET, and Raman spectroscopy require time-stamping the arrival of photons with picosecond accuracy. However, SPADs have an inherent timing jitter due to the statistical distribution of delays between photon absorption and avalanche pulse detection by the pulse pick-up electronics [132]-[138]. The distribution is typically characterized by a sharp narrow Gaussian peak attributed to photons absorbed within the SPAD’s active region. The width of the peak depends on the lateral avalanche build-up time [130],[131]. An exponential tail component is also present in the timing distribution. This tail is caused by the diffusion of minority photo-generated carriers (electrons in p-layer and holes in n-layer) within each neutral layer of the SPAD towards the depletion region [87]. Since the diffusion time depends on the photon absorption depth, the diffusion tail is wavelength-dependent. Also, as the minority carriers undergo diffusion from the neutral layers toward the multiplication region, the diffusion properties will depend on the minority carrier type that initiated the Geiger pulse.

Typically, SPADs with shallow and thin active regions can achieve the best timing performance. However, this is at the expense of PDE since the photosensitive volume is smaller [162]. Junctions with lower electric field in the active region tend to have worse jitter, since the avalanche build up time is statistically more uncertain [192]. A summary of CMOS SPADs having FWHM time resolutions ranging from 36 – 230 ps is provided in [158],[192].

## Time-to-Digital Converters (TDC)

TDCs measure time intervals between two input signals and are fundamental building blocks of time-resolved single-photon imaging systems [193]-[195]. The circuitry found in TDCs can generally be classified either as digital or analog. Generally, analog methods allow better resolutions. However, digital methods are easier to implement in integrated circuits, consume less chip area, and are less sensitive to process, voltage and temperature (PVT) variations, making them the preferred choice for compact, multi-channel implementations such as TCSPC [43]-[47]-[49],[69]-[73],[196],[197].

The key performance requirements of TDCs designed for TCSPC applications include high time resolution and low non-linearity in order to accurately measure fluorescence decays. Wide dynamic range is required to record a wide range of lifetimes as well as short conversion time in order to minimize pile-up and increase acquisition speed. Robustness to PVT variations, multi-channel capability, low circuit area and low power consumption are other key requirements for the realization of high-performance TCSPC systems. In this section, a review of the main performance specifications of TDCs, and architectures used for state-of-the art TCSPC applications will be briefly reviewed.

### Key TDC Specifications

Time Resolution

Fundamentally, a time interval measurement involves the calculation of the elapsed time between a designated START phenomena (such as the time at which a laser pulse occurs), and a later STOP phenomena (such as the time at which the first fluorescence photon arrives at the detector). The TDC converts a START-STOP time interval at its input terminals into a digital value. The ideal TDC input–output characteristic, shown in Fig. 2-10(a) for a 6-bit TDC, represents a quantizer function that maps a continuous range of time intervals at the input onto discrete output values [198]. The least-significant-bit (LSB) of the digital word is the smallest time interval *TLSB* that can be measured by the TDC. The measured time interval is thus given *TM* = *nTLSB* where n is the digital word output by the TDC. An *N* bit TDC has 2*N* quantization steps and the maximum time interval that can be measured is given by *TLSB*×2*N*.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 2‑10: (a) Input-output characteristic of an ideal 6-bit TDC with *TLSB* = T*CLK*­/64 =­ 156.25 and T*CLK* = 1000 ps. (b) Associated quantization error values.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 2‑20: (a) Input-output characteristic of an ideal 6-bit TDC with *TLSB* = 156.25 and (b) associated quantization error.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 2‑21: (a) Input-output characteristic of an ideal 6-bit TDC with *TLSB* = T*CLK*­/64 =­ 156.25 and T*CLK* = 1000 ps. (b) Associated quantization error values.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 2‑22: (a) Input-output characteristic of an ideal 6-bit TDC with *TLSB* = 156.25 and (b) associated quantization error.

The quantization of a continuous time interval into a discrete value results in a quantization error that represents a fundamental source of uncertainty in TDC measurements [198]. The quantization error ε = *TM* – *TIN* for the ideal 6-bit TDC is illustrated in Fig. 2-10(b). Quantization error as a function of TDC input assumes the shape of a sawtooth wave with values between 0 and *TLSB* for an ideal TDC that is perfectly linear and free from timing jitter. The probability distribution function of quantization error over the input range is uniform since all error values between 0 and *TLSB* are equally likely to occur. The timing uncertainty *σTDC* of a single TDC measurement is given by the root-mean-square (rms) value of the quantization error distribution. For the ideal TDC, it is given by [194]

|  |  |
| --- | --- |
|  | (2-1) |

Eq. (2-1) represents the smallest possible uncertainty that a TDC can achieve solely due to quantization error. However, real TDC’s are subject to errors due to the non-linearity of the TDC characteristics and timing jitter. The non-linearity is a systematic deviation of the TDC characteristics from the ideal case due to the imperfections, such as device mismatch, of the electronic circuitry. The nonlinearity must be reduced as much as possible in order to realize a high-performance TDC.

Non-Linearity

Irrespective of their architecture, all practical TDCs are subjected to non-linearity causing the input-output characteristic to deviate from a perfect linear function [199]. Whereas an ideal TDC would have each time-bin with the exact same width, the time-bin widths of real TDC implementations can vary slightly from the ideal value due to imperfections in the TDC circuitry leading to non-linearity of the transfer characteristics and an increase of the quantization error. This is illustrated in Fig. 2-11(a) where the characteristic of a 6-bit TDC affected by non-linearity is shown.

The differential non-linearity (DNL) and integral non-linearity (INL) quantify the amount of variation in each time-bin. The INL is the deviation of the actual TDC characteristic from its ideal value, while the DNL is the deviation of each time-bin width from the ideal *TLSB* [200]. In other words, the DNL and INL are the respective microscopic and macroscopic deviation of the real TDC characteristic from the ideal case. Mathematically, DNL and INL are expressed as

|  |  |  |
| --- | --- | --- |
|  |  | (2-2) |

where *Sj* is the *jth* time-bin width of the TDC characteristic and *Tj* is the time difference between the ideal output time and the output TDC time for the *j*th time bin. Both are defined for each quantization step, however, the maximum or rms value is typically used to describe the total nonlinearity over all steps [201]. The rms DNL/INL reveals the effect of the nonlinearity on the TDC precision. The DNL/INL negatively influences the performance of TDCs because it introduces a systematic timing error that is additional to the quantization error as shown in Fig. 2-11(b). Here, the error function is no longer a sawtooth, and the quantization error is no longer *TLSB*/√3. The nonlinearity was computed using eq. (2-2) for the TDC characteristic shown in Fig. 2-11(a).

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 2‑11: (a) Input-output characteristic of a 6-bit TDC including non-linearity. (b) Plot of the associated quantization error.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 2‑24: a) Input-output characteristic including non-linearity of a 6-bit TDC. b) Plot of the associated quantization error.

Fig. 2-12(a) shows the obtained DNL/INL as a function of TDC code. In this case, the nonlinearity is such that the maximum INL < 1 LSB. When the linearity of the TDC transfer characteristic is degraded by the effects of circuit mismatch to the extent that the maximum INL> 1 LSB (as shown in Fig. 2-12(b)), then there is a possibility that there can be missing codes in the TDC characteristic (i.e. one or more of the possible 2*N* binary codes are never output). A missing code leads to higher quantization error and results in the reduction of the converter’s effective number of bits (ENOB) [201]. The probability mass distributions (PMF) of quantization error are illustrated in Fig. 2-13 for 6-bit TDCs with INL/DNL shown in Fig. 2-12. The distribution of quantization error is no longer rectangular as in the ideal case, but a Gaussian shape. This is due to the random timing errors caused by the non-linearity. As the non-linearity in the TDC characteristic increases, the quantization error distribution becomes much wider. Here, the standard deviation of the error distribution, *σTDC*, represents the timing uncertainty of the TDC due to both quantization and non-linearity. The effects of timing jitter will also affect the TDC’s timing uncertainty.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 2‑12: – Plot of DNL/INL for a 6-bit TDC with (a) max. INL < 1 LSB, and (b) max. INL > 1 LSB.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 2‑26: – Plot of DNL/INL for a 6-bit TDC with a) max. INL < 1 LSB, and (b) max. INL > 1 LSB.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 2‑13: Corresponding quantization error PMF of 6-bit TDC characteristics shown in Fig. 2-12. PMF of quantization error for max. INL < 1 LSB in case (a) is much closer to the ideal case than for max. INL > 1 LSB in case (b)

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 2‑28: Corresponding quantization error pmf of 6-bit TDC characteristics shown in Fig. 2-12. a) Quantization rms error for max. INL < 1 LSB case is much closer to the ideal case than for max. INL > 1 LSB (b)

Timing Jitter

The effect of timing jitter causes the measured TDC output to deviate from the true value whenever a constant START-STOP time interval is repeatedly measured. The corresponding standard deviation is called the single-shot precision and it is illustrated in Fig. 2-14. The single-shot precision is obtained by measuring a constant time interval repeatedly and taking the standard deviation of the distribution of the measurement results around the mean value [202]. Single-shot precision describes how reproducible a TDC measurement is in the presence of timing jitter and it reveals the effects of nonlinearity and quantization noise in a single parameter. The precision depends on the time interval to be measured, thus the maximum and rms values of the precision within a certain measurement range should be indicated when describing a TDC [203].

|  |
| --- |
|  |

Figure 2‑14: Representation of TDC output code histogram [202].

|  |
| --- |
|  |

Figure 2‑30: Representation of TDC output code histogram [202].

|  |
| --- |
|  |

Figure 2‑31: Representation of TDC output code histogram [202].

|  |
| --- |
|  |

Figure 2‑32: Representation of TDC output code histogram [202].

In practice, the single-shot precision is limited by the inherent jitter of the timing input signals and the imperfections of TDC circuits caused, for example, by power supply noise and device mismatch. The precision, *σTDC*, is defined in eq. (2-3) as

|  |  |
| --- | --- |
| . | (2-3) |

Eq. (2-3) is composed of contributions from rms quantization error *σq*, rms INL *σINL*, reference clock rms jitter *σREF*, and Voltage Controlled Delay Line (VCDL) rms jitter *σVCDL*. It is apparent that no significant improvement in precision can be achieved by improving the resolution if the nonlinearity dominates and does not scale down with the improved resolution [203]. The precision of the time interval measurement can be improved by averaging if the time interval to be measured is repetitive and the START-STOP signals are uncorrelated with the time base of the measurement [194]. However, averaging is a slow process as the improvement of the precision is proportional to √*N*, where *N* is the number of single-shot measurements.

There has been a considerable effort in developing circuit architectures and applying higher speed technologies to achieve a better resolution and single-shot precision for TDCs [193]. The choice of TDC architecture has a significant influence on the conversion time as well as the power consumption, circuit area and accuracy. In the following section, several important TDC architectures are presented with examples from published realizations.

### TDC Architecture Review

Time-to-Amplitude Converter (TAC)

Accurate time interval measurements required for TCSPC are routinely performed by Time-to-Amplitude Converters (TAC) [28],[144],[204],[205]. In a TAC, the time interval is converted to a voltage by using a switched current source to charge or discharge a capacitor. The START pulse switches the current on, the STOP pulse switches it off, and the final voltage at the capacitor represents the time between START and STOP. An ADC samples the final value to complete the conversion of the time interval into a digital format. This principle works with very high accuracy, and time differences of a few ps can be resolved [28],[144],[204],[204]. While TACs can perform very well with respect to resolution and dynamic range, their performance depends mainly on the ADC. High-performance ADCs can be implemented using discrete off-chip components. For multi-channel systems, however, this results in higher overall system cost, size and power consumption [144].

Delay Line Interpolation

Integrated TAC/ADC based approaches for time interval measurements are limited due to fabrication, operation and environmental impairments of analog circuits in a DSM CMOS technology. This is because analog circuits are more susceptible than digital circuits to device mismatch, ambient temperature variations, and external disturbances such as power supply noise. On the other hand, TDCs in DSM CMOS consisting of mostly digital circuits such as logic gates, counters, samplers, multiplexers, and decoders can operate at high-speed, are compact, robust and consume low power [194]. While the propagation delays of logic gates are sensitive to temperature and supply voltage, feedback techniques such as delay-locked loops (DLL) can be used to stabilize them against process, voltage and temperature (PVT) variations [202],[203],[206].

The simplest digital technique to quantize a time interval is to count the cycles of a reference clock within the measurement time interval with a digital counter [207]. This method has short conversion time, which enables a high measurement rate and is excellent for measuring long time intervals, with the range of time measurement being limited only by the number of bits of the counter logic. It can also be inexpensively implemented as a multi-channel arrangement in field-programmable gate-arrays (FPGA) [208]. However, the time resolution is limited by the clock frequency. The reference clock frequency and power consumption becomes unreasonably high for sub-nanosecond timing resolution.

High time resolution and low power consumption can be achieved simultaneously by means of interpolation techniques that digitize the position of the timing signal with respect to the reference clock period [194]. A chain of delay-adjustable digital gates (delay elements) divides the reference clock period into small, even-sized time intervals. The location of the timing signal within the reference clock cycle is resolved by recording the position of the timing signal within these delay elements. However, because of PVT variations, the delay elements require calibration of the delay elements with a stable time reference to achieve robust and accurate TDC performance.

An integrated DLL ensures that the delay elements can be self-calibrated using a stable external reference clock such as crystal oscillator [202],[203],[206],[209]-[211]. The delayed reference clock propagating through the voltage-controlled delay line (VCDL) is compared to the reference input by a phase detector [212]. If a delay different from one clock period is detected, then the closed loop will automatically correct it by changing the time constant of the delay cells via a charge pump and filter capacitor [213].

An initialization state machine prevents the DLL from false locking at startup and the closed control loop guarantees automatic calibration, drastically reducing the TDC sensitivity to PVT variations [214],[215]. However, the maximum attainable TDC resolution depends greatly on the CMOS technology and the architecture of the VCDL. To date, the best resolution achievable by a VCDL is ~300 ps in 0.35 μm CMOS technology [216], ~60 ps in 0.13 μm CMOS [217], and ~20 ps in 90 nm CMOS [218]. Achieving a time resolution of less than one digital gate delay over a wide measurement range is not practical with only one level of delay line interpolation, so multi-level interpolation is used.

Multi-level Interpolation

High-timing resolution, wide dynamic range and low power can be simultaneously achieved by means of multilevel interpolation techniques [202],[203]. In the two-level interpolation scheme, the first coarse interpolator, interpolates the timing signal within the reference clock cycle time with a coarse resolution. The second fine interpolator detects the location of the timing signal within the result of the first level with a higher fine resolution. The multilevel approach provides a wider dynamic range with a smaller number of delay elements and registers, on account of the fact that the second interpolation stage needs only to cover the delay of a single element in the first interpolation stage [216]. A smaller number of delay elements and registers reduces the size of the circuit, and the power consumption can be minimized because the second level operates only when the timing signal arrives. However, the coarse and fine interpolators must be properly synchronized in order to avoid the effects of metastability in the sampling flip-flop circuits [219].

Many circuit design techniques have been proposed in order to realize a fine interpolator that has sub-gate delay resolution [220],[221], most often by using the Vernier method [193],[195]. In a Vernier delay line (VDL) two delay VCDLs are used for the fine interpolation. The delay of one of the VCDL is slightly greater than the delay of the other. As the START and STOP signals propagate in their respective delay chains, the time difference between the START and the STOP pulse is decreased in each delay stage by the difference in the delays. In each stage, START and STOP signals are compared to determine which of the two input signals came first. The position in the delay line at which the STOP signal catches up with the START signal represents the measured time interval encoded in a thermometer code. Since each bit represents the difference between the delays of the VCDLs, picosecond-level resolution can be achieved. Resolutions down to 7 ps have been reported [222].

The accuracy of two-stage VCDL can be affected by error factors such as mismatch of the delay lines, the physical length of the delay line and by circuit noise [223]. Further, the number of delay stages grows exponentially with the number of bits, making the fine interpolation highly sensitive to delay jitter noise and mismatches especially when a high number of bits is required. Also, a VDL has an inherent burden of a long dead-time, allowing essentially only one STOP signal per measurement, and it is therefore not particularly suitable for high-speed applications.

Based on the considerations above, a coarse-fine interpolating TDC architecture was chosen in order to attain compact circuit area and low power consumption. The TDC should attain sub-nanosecond time resolution (~150 ps) and low maximum INL (< 0.5 LSB) to be able to measure the inter-arrival times of SPAD pulses accurately. Compact size is required for multi-channel realizations, as well as high sample-rate (~100 MHz) and low power consumption (~1 mW).

In Chapter 6, the design and measurement results of a prototype CMOS TDC chip that was fabricated using a 130 nm digital CMOS process will be described. The TDC prototype chip is targeted for integration with SPAD arrays for FLIM applications requiring high-speed, multi-channel TCSPC. The prototype IC was fabricated to characterize the TDC design for future utilization in multichannel TDC/SPAD SoCs. In the next two chapters, the design, modelling, fabrication, characterization and measurement of CMOS SPADs for FLIM applications is described.

# CMOS SPAD Design in standard 130 nm technology

This chapter begins with a discussion about the key technological considerations when implementing SPADs in standard DSM CMOS. This is followed by a review of SPAD structures previously reported in other CMOS technologies. Advantages and disadvantages of the different SPAD structures reported in the literature will be discussed. Then, the SPADs that were designed, fabricated and measured in this work will be presented. The key steps for successful implementation of SPAD pixels in a standard digital 130 nm CMOS technology will be described. Three different passively quenched free-running SPAD pixel structures were designed, simulated, measured and characterized. These were unbuffered SPADs, SPADs with a source-follower (SF) amplifier (SF-SPAD), and SPADs with a common-source (CS) amplifier (CS-SPAD). In addition, the output pulses of a passively-quenched, actively-recharged time-gated (TG) SPAD pixel were measured and characterized. Simulations of the pixels will be shown to match the measured results. The pixel structures introduced in this work can cover a wide range of biomedical imaging applications that utilize TCSPC and TGSPC.

## Standard deep-submicron CMOS Technology

The first SPAD implementations in CMOS technology were published almost 15 years ago [148]-[151]. Since that time, there has been remarkable progress in SPAD performance regarding their array size, photon detection efficiency (PDE), dark count rate (DCR), jitter and afterpulsing performance [145]-[177]. SPADs fabricated with custom CMOS image sensor (CIS) technology [166]-[169],[173]-[175], can achieve performance that surpasses those fabricated in standard digital/RF CMOS [53],[61],[62],[83]-[86],[147],[170],[176],[177]. On the other hand, the numbers of foundries that offer standard digital CMOS technology are more numerous, hence the fabrication costs using standard technology are much lower compared to CIS technology.

Modern CMOS processes are driven by logic and RF applications that always demand higher speed, smaller size and reduced power consumption [224]. As such, advances of the technology are not necessarily in line with SPAD requirements [227]-[229]. In fact, many of the process features that are standard for digital CMOS ICs, such as shallow-trench isolation (STI), shallow source/drain junctions, silicide and surface passivation introduce significant challenges in the realization of high performance SPADs [225]-[231].

The use of standard digital/RF technologies for single-photon detection is mostly beneficial in terms of the lower fabrication cost, increased parallelism and performance improvement of TDC circuitry rather than for the SPAD’s performance. As a result, many challenges and opportunities still exist in the development of high-performance SPADs in a standard DSM CMOS technology. The technological considerations of SPADs fabrication in 130 nm IBM CMOS technology are summarized next, followed by a review of reported SPADs structures in DSM CMOS.

### Technology features of standard DSM CMOS technology

The main attractive features of DSM technology are higher speeds, lower power, and increased integration density needed for digital circuits [224]-[229]. Meanwhile, the industry standard digital and RF CMOS fabrication technologies offered by the foundries have a precisely defined sequence of processing steps that cannot be modified by the user. This means that the SPAD performance cannot be improved by changing the processing parameters since this is not allowed in a standard CMOS process. Despite these challenges, SPADs can be successfully implemented in a standard digital DSM CMOS technology without any modifications to the processing steps provided that the available design masks and supported features are carefully and properly utilized during the design. Fig. 3-1 illustrates the cross-sections of a CMOS chip in a 130 nm IBM CMOS technology. In this figure, some of the main features of the manufacturing process are highlighted. The impact that these features have on the SPAD design is discussed next.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
|  | |
| (c) | |

Figure 3‑1: Cross-section views of a triple-well DSM CMOS technology: (a) Inter-metal dielectric stack [289] and (b) transistor structures [228]. (c) Three possible photodiode structures are available for SPADs: n+/p-substrate, deep n-well/p-substrate and p+/n-well. Arrows point in the direction of the electric field which causes the drift current due to the minority carriers.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
|  | |
| (c) | |

Figure 3‑2: Cross-section views of a triple-well DSM CMOS technology: a) Inter-metal dielectric stack [289] and b) transistor structures [228]. (c) Three possible photodiode structures are available for SPADs: n+/p-substrate, deep n-well/p-substrate and p+/n-well.

Triple Well

An optional high-energy ion implantation step is a standard feature of CMOS process generations beyond 250 nm used to make a deep n-well (DNW). This feature results in an enclosed p-well region formed when the DNW is contacted by a ring of n-well [129]. The isolated p-well (iso. p-well) is meant to shield nMOS transistors from the substrate and to reduce noise-coupling issues in mixed-signal circuits. By utilizing the DNW along with the n+/p+ source/drain implants and n-well/p-well diffusions, several different p/n junction structures can be used to form SPADs in a standard DSM CMOS process [146],[147],[165],[172],[177],[232]. The most commonly used structures are illustrated in Fig. 3-1(c), n+/p-sub, DNW/p-sub and p+/n-well.

The n+/p-substrate SPAD has the simplest structure. However, it is inherently limited since there is no electrical isolation between the SPAD and the substrate [176]. A high voltage has to be applied between the anode and cathode in order to exceed the SPAD’s breakdown voltage, but the substrate must be grounded for proper circuit operation. The high voltage must be applied to the cathode, but this is incompatible with the IC supply voltage limits. Using this structure for SPADs, an AC coupling capacitor is needed to bring the SPAD signal to CMOS voltage levels. A high voltage compatible two-metal layer metal-insulator-metal (MIM) can be placed inside the pixel for this purpose, but the penalty is a reduced fill-factor and increased pixel complexity. [147]

In a standard CMOS process, the lowest doped p-type material available is the p-substrate. As a result, the DNW/p-substrate type SPAD achieves higher breakdown voltage and suppression of tunneling [173],[175]. The depletion region is wider and deeper in the silicon, so it responds better to longer wavelength photons (green/red light) which penetrate deeper. However, the avalanche region is not isolated from the substrate, leading to similar voltage incompatibilities, as well as the same fill-factor and pixel pitch limitations as found with the n+/p-substrate SPAD.

The p+/n-well type SPAD utilizes the shallow source/drain implant of the standard digital process to achieve a shallow avalanche region with the additional benefit of substrate isolation [159]-[167]. No AC coupling is required for these pixels, since the HV can be applied directly to either the cathode or anode while safely grounding the substrate. In the so-called ‘positive drive’ configuration, a positive HV is applied to the n-well cathode with the p+ anode as the sense node connected to ground through a quenching element [161],[162],[163],[165]-[167]. Alternatively, in the ‘negative drive’ configuration, a negative voltage can be applied to the p+ anode with the positive CMOS voltage supply connected through a quenching element to the n-well cathode sense node [159],[169],[171]-[174].

In either positive or negative-drive cases described above, the n-well/p-substrate junction provides the isolation of the active region from the substrate. But when the negative drive configuration is used, the n-well cathode is the dynamic node and therefore the additional capacitance of the n-well/p-substrate parasitic diode contributes to the SPAD’s capacitance. This increases the amount of charge flowing through the SPAD’s avalanche region, implying higher afterpulsing probability as well as increased dead time due to the increased RC load. In addition, this configuration is susceptible to latch-up since it forms a vertical p+/n-well/p-sub bipolar transistor structure [233].

Latch-up occurs when the negative high voltage on the anode is increased too far beyond the breakdown voltage. This causes the cathode to discharge to a negative voltage in order to quench the avalanche. When this occurs, the parasitic n-well/p-substrate junction becomes forward biased. Now, holes from the substrate are emitted into the avalanche region, causing the SPAD to remain permanently in avalanche multiplication state. This limits the maximum excess voltage swing of the negative drive configuration of the p+/n-well SPAD.

In the positive drive configuration, the parasitic n-well/p-substrate diode no longer adversely contributes to the SPAD’s capacitance, and latch-up is avoided. This configuration permits the sharing of n-well regions by multiple SPADs within an array, leading to reduced pixel pitches and improved fill factor [234]. However, the range of excess bias is limited by the maximum voltage that can be applied to the n-well cathode without breakdown of the n-well/p-substrate parasitic junction. Also, since the avalanche region is formed using highly doped source/drain implants in an n-well, the tunneling-induced DCR is higher compared the same structures in HV CMOS

A structure more suitable for pixel arrays is the iso p-well/DNW SPAD, which benefits from the substrate isolation, DC coupling, n-well sharing and reduced inter-pixel spacing [160],[177],[217]. Higher breakdown voltage and suppressed tunneling effects were shown for these buried active region structures. The graded DNW doping profile additionally provides an implicit guard ring structure that enables miniaturized SPADs with pixel pitch less than 5 μm. This configuration makes the structure suitable for pixel miniaturization [160].

When comparing the n+/p-well junction to the p+/n-well junction, it is apparent that the parasitic capacitance of the device is lower in the former. This is due to the n-well having a larger capacitance to the substrate, while the n+ layer has a much smaller capacitance to the p-well. A lower parasitic capacitance at the sense node is more desirable for increased speed of operation, as well as a reduction of the avalanche charge [163]. These considerations lead to the implementation of a n+/p-well SPAD structure in 130 nm CMOS, in order to benefit from the substrate isolation and possible sharing of p-well regions by multiple SPADs. The implementation details of the n+/p-well SPAD structure are considered in detail in Section 3.2.1.

Shallow Trench Isolation (STI)

STI is needed in DSM technology since it prevents many of the problems associated with local oxidation of silicon (LOCOS). STI resulting in higher transistor densities and improved junction isolation [129],[166],[225],[228]. The STI trenches formed by reactive ion etching (RIE) of the exposed silicon are typically 450 nm deep and filled with oxide [226]. In a standard digital CMOS process, all source/drain implants are surrounded by STI. For the design of APDs in DSM technology, STI is used to eliminate the curvature effects that lead to PEB in p+/n-well or p-well/n+ junctions, since the edges of the shallow source/drain implant are better confined by the oxide trench.

APDs fabricated in the STI process described above show higher avalanche gain and better responsivity compared to those using diffused guard rings because the STI causes the planar area to achieve avalanche breakdown before the edges and corners do [188],[235]-[239]. However, higher leakage current results from the sidewall interface between the STI and n+/p+ regions, which is rich in defects at the SiO2-Si boundary [225]. The dark noise of APDs thus depends on the distance between the defective sidewall interface of the STI and the avalanche region [240],[241]. Similarly, in SPAD design the DCR increases significantly when STI is in direct contact with the avalanche region to form a guard ring [166],[167],[242].

To achieve the best possible DCR performance, the STI must be physically separated from the SPAD avalanche region in order to assure that the minority free carriers generated at the defective interface are recombined before potentially diffusing into the avalanche region [240],[241]. There are several design techniques that can be used at the layout level to force the physical separation of the STI interface from the SPAD multiplication region. Examples of such techniques that lead to a beneficial impact on the DCR are described in refs. [168]-[172],[243].

Silicide

Silicidation is the process of creating a surface layer of a metal on silicon in order to reduce the resistance increase associated with reduced feature sizes [244]-[250]. Because the silicidation step does not require a process mask, it is also called self-aligned silicide, or salicide. Silicides may be formed by the use of TiSi2 or CoSi2 [245]. Cobalt silicide (CoSi2) is utilized in 130 nm CMOS technology due to its low sheet resistance and high stability [244],[245]. However, silicidation of sources and drains becomes a problem in that the silicide can penetrate through the shallow junctions. This effect (called ‘silicide spiking’) has been shown to greatly affect the leakage current of n+-p junctions [245]-[250] and has serious implications for DSM SPAD design as the silicide leakage current flows not the junction periphery, but in the junction area. Fortunately, the reverse leakage current of silicided pn junctions is still within the limits considered acceptable for VLSI logic applications, provided that the supply voltages do not exceed their recommended limits.

SPAD voltages, on the other hand, must exceed the junction breakdown voltage. In this case, the higher leakage currents together with the field-assisted generation mechanisms are expected to contribute to the increased DCR. It was postulated that small regions of silicide penetration (i.e., so called silicide spikes) are responsible for an increased leakage current that flows across many localized defect points in the SPAD junction area [245]. The silicide penetrations may also be the cause of a stronger tunneling-current increase and activation energy decrease with increasing bias voltage, resulting in very high DCR levels for silicided SPADs. Further, since silicide is an opaque material for visible light, most of the incoming photons cannot pass through this layer [238].

During the fabrication process, all diffusions and polysilicon are silicided for low resistivity unless a silicide blocking mask is used. A silicide blocking mask, called oxide protect (OP), is available in the IBM 130 nm CMOS technology to fabricate on-chip resistors and input/output (I/O) transistors. When a shallow source/drain implant is utilized in the photodiode structure, optical windows for improved light transmission and reduced leakage can be made by selectively blocking the silicide formation with the OP layer [238],[249]-[250]. The PDE and DCR performance is expected to improve for non-silicided SPADs, compared to those that have silicide on the photosensitive area. Measurement results of SPADs with and without silicide are compared and analyzed in detail in sections 3.2.2, 3.3.3, 4.1.3, 5.1 and 5.2 to validate this hypothesis.

Passivation and Inter-Metal Dielectric (IMD) Layers

The final processing step is to add a protective glass passivation layer that protects the IC from mechanical abrasion and to provide a barrier to external contaminants [186],[187]. Final chip passivation is formed by a sequence of oxide, nitride, and polyimide film depositions. The nitride that serves as the ionic contamination barrier significantly attenuates the intensity of the incident UV light [252]. On the other hand, the thicker polyimide layer providing mechanical protection reflects light. Below the passivation, an inter-metal dielectric (IMD) stack is used to isolate the metal layers from each other and for planarization.

Light has to further travel through the thick IMD stack on top of the silicon substrate to reach the SPAD so the spectral response of the SPAD will be impacted [189], [249]-[250]. Since there are eight copper layers and one aluminum layer in the IBM 130 nm CMOS process, the distance between the top dielectric layer and the silicon surface can reach up to 10.5 μm. In addition, the thickness of the dielectric layers can vary by as much as 20%. The results of these oxide layers and their non-uniformities are a very irregular transmission characteristic as a function of wavelength [188].

In CIS technologies, special back-end processing steps are used to shrink the IMD thicknesses and to selectively etch passivation openings above the photosensitive pixel area to improve light transmission [249],[250]. When photosensitive regions are highlighted with the passivation opening layers in a CIS process, the foundry service translates this information into a physical etch of the passivation nitride layer resulting in improved sensitivity. Meanwhile, the standard low-cost CMOS does not include such features. Therefore, the PDE performance of standard CMOS SPADs is worse compared to those in CIS technologies.

In a standard process, the passivation is only etched away at the bonding pad positions to be able to make wiring connections from the chip to the package. In the 130 nm CMOS process, a design layer is available to remove the surface passivation. However, it is a wafer level option, meaning that whenever it applies, it has to cover the entire wafer. Because it is very expensive to purchase an entire wafer, fabricating SPADs through Metal Oxide Semiconductor Implementation Service (MOSIS) without a passivation layer using multi-project wafer (MPW) fabrication services is not a cost-effective solution. Polyimide can only be removed cost-effectively with a post-process, passivation etching step [159],[253].

### SPAD Guard-ring Structures in DSM CMOS

SPADs implemented in a DSM CMOS must be carefully designed in terms of their guard-ring structure. The edges of the SPAD must be protected as the standard geometries with sharp edges have higher electric fields there, resulting in premature edge breakdown (PEB) [182],[183]. When PEB occurs, the SPAD cannot be biased above its breakdown voltage uniformly across the entire photosensitive area and therefore, it cannot be used to detect single photons. A wide variety of SPAD guard-ring structures have been adopted to avoid PEB while conforming to the technological and design rule constraints of standard CMOS technologies [145]-[178]. All of these techniques have in common the reduction of the electric field at the edges of the device so as to maximize the probability that the avalanche is initiated within the planar multiplication region only.

The first SPADs implemented in HV CMOS had guard-ring structures that utilized deep diffused p-well implants representing the bulk of isolated nFET placed along the periphery of shallow p+ implantations, (representing source/drain regions of pFET) in an n-well [145],[148],[150],[151],[153]-[158]. STI guard ring structures were first introduced in 180 nm CMOS. The main goal of STI was to increase the fill-factor, reduce pixel size and reduce the spacing between pixels [242]. However, these SPADs with STI guard rings had very large DCR (~1 MHz). SPADs structures incorporating STI guard ring had improved DCR in 130nm CMOS technology [167]. These designs used low doped p-well passivation implants to reduce the surface leakage associated with the STI, a feature available only in CIS processes. However, DCR levels were still high (40–100 kHz) due to carrier tunneling in the high field regions [166].

To reduce the tunneling contribution to DCR, p-well/DNW SPAD structures were proposed with improved scalability, thanks to the use of virtual guard rings and buried junctions [147],[164],[168]-[172]. The guard rings were built by utilizing the retrograde doping profile of the DNW. The p-well implants were implicitly blocked at the active area borders. This resulted in a reduced doping concentration of the DNW towards the surface, which provided the guard ring formation. As a result of the retrograde doping characteristic at the periphery of the SPAD, PEB was eliminated. [168]. Another advantage of the virtual guard ring is that the SPAD active area could be scaled down to much smaller dimensions compared to the SPAD with diffused guard ring [171]. Smaller SPADs are favorable for lower DCR performance and better yield in DSM technology, since there is less probability of having a defect in the active area [156].

Smaller SPADs have lower jitter due to the reduction of lateral avalanche build-up time with area [130],[172]. On the other hand, small active areas imply a very low pixel fill factor, especially when in-pixel circuits such as AQR or TDC are included [43],[44],[47],[196]. In this work, large active areas of 81 μm2 and 100 μm2 were utilized. These are the highest reported SPAD sizes in standard digital DSM CMOS. These active areas were found to be a suitable compromise between acceptable pixel-fill factor, timing-resolution, afterpulsing and DCR performance, as will be described in detail in the following chapters.

## SPAD structures fabricated in 130 nm standard CMOS

The SPADs presented in this work were fabricated using a standard digital IBM 130 nm CMOS technology [226]-[229]. Unlike CIS technologies that offer specialized process modules that are tailored to obtain improved image sensor performance [173]-[175],[249]-[250], standard digital CMOS does not have any special processing steps optimized specifically for imaging. However, standard digital CMOS is attractive for SPAD implementation because it offers the lowest possible fabrication costs, in addition to all the other additional benefits of a digital/RF process, such as high speed and low power consumption [254].

In this work, SPADs pixels compatible for integration with high-performance digital circuits were designed, modeled, fabricated and characterized. The previously reported pixels in 130 nm CMOS were outperformed in terms of DCR and AP performance in comparison to other implementations in similar technologies [168]-[173]. However, this was partially due to the fact that silicided active regions were used in those test structures [53], [84]-[86].

In Chapters 4 and 5, it will be shown that SPADs without the silicide layer in the active region can attain higher levels of performance. In the following sections, the design, modeling and output pulse measurements of the designed SPAD pixel structures are presented. Then, the detailed characterization results of the SPADs pixels proposed in this section are presented in Chapters IV and V.

### Structural Characterization and Fabrication Details

A schematic cross-section to illustrate the various features of the SPAD structure is shown in Fig. 3-2. A local p-well is formed by the DNW. The ring of n-well provides lateral isolation and connects to the DNW. pFETs are not typically supported inside the isolated p-well, thus the formation of an n-well within the isolated p-well is forbidden by the normal design rules. Yet the design rules do not provide a specific guarantee that any design that passes or fails design rule checking (DRC) will operate correctly or incorrectly. It is possible to fabricate SPADs by carefully violating certain design rules. In this technology, when an n-well is implanted within a p-well that is isolated by a DNW, the n-well will not go deep enough to touch the DNW. By placing an n-well within the isolated p-well, a guard-ring structure can be created to suppress the PEB and enable the formation of n+/p-well SPADs in standard digital 130 nm IBM CMOS [53],[84],[86]

The diffusion and implantation characteristics of the n-well and DNW create the guard rings necessary for PEB prevention and for substrate isolation, respectively. The PEB guard ring is formed by the n-well diffusion placed along the periphery of the n+/p-well active region. The width of the n-well guard ring is 1.5 μm, which is about twice the minimum size allowed by the design rules. The lower doping of the n-well compared to n+ implant reduces the electric field at the edges so that a high and uniform electric field is encountered only within the planar region defined by the n+ area enclosed within the n-well guard ring. A rectangular shaped active area of 8.5×9.6 = 81 μm2 is formed by the n+ implantation area not covered by n-well. This area defines the n+/p-well avalanche region where the electric field exceeds the critical value for impact ionization and avalanche breakdown. The active region is isolated from the p-substrate by the DNW and the outer n-well ring. The outer n-well forms the outer guard ring and provides the lateral isolation. The spacing of the outer guard-ring is restricted by the DNW design rules that should not be violated so the outer n-well ring has a 2 μm width and dimensions of 21 × 23 μm2. Therefore, the fill-factor of the bare pixel (SPAD without any transistors) is 17%.

|  |  |
| --- | --- |
|  |  |
| (a) | (c) |
|  |  |
| (b) | (d) |

Figure 3‑2: (a) Layout view of SPAD test structure with important dimensions in microns. (b) Unbuffered SPAD pixels – Left pixel: Silicided n+ junction. Right pixel: Non-silicided n+ junction. (c) Cross-section view of SPAD and nMOS transistor. (d) The wire-bonded die resides in the cavity of a 68 pin-grid-array (PGA68) ceramic package.

|  |  |
| --- | --- |
|  |  |
| (a) | (c) |
|  |  |
| (b) | (d) |

Figure 3‑4: a) Layout view of SPAD test structure with important dimensions in microns. b) Unbuffered SPAD pixels – Left pixel: Silicided n+ junction. Right pixel: Non-silicided n+ junction. (c) Cross-section view of SPAD and nMOS transistor. (d) The wire-bonded die resides in the cavity of a 68 pin-grid-array (PGA68) ceramic package.

The outer n-well, which forms the outer guard ring and provides the lateral isolation, is electrically connected to substrate ground potential by an ohmic n+ contact. When a high-voltage (HV) is applied to the p+ ohmic contact at anode, the DNW becomes fully depleted and electrically isolates the SPAD avalanche region from the substrate. The DNW allows the substrate to be grounded for CMOS compatibility and enables the cathode to be directly coupled to CMOS gates without AC coupling capacitors. Additionally, the SPAD is unaffected by the minority carrier injection into the active region through the parasitic bipolar structure [255]. Further, the parasitic DNW/p-substrate diode does not increase the capacitance of the SPAD when sensing the avalanche at the n+ cathode. Since there is only the charging of the n+/p-well junction capacitance to consider, faster switching and reduced afterpulsing probability is possible because less charge flows through the junction when an avalanche is triggered. Sharing of p-well regions by multiple elements within an array is also possible in this configuration, which can lead to the realization of SPAD arrays with small pixel pitch and higher fill-factor. Using the p+/n-well junction also avoids having to bias the p-sub at a large negative voltage which is a reliability concern.

The side-walls and edges of the STI are separated from the depletion region by the extension of n+ over n-well. Since the STI is defined everywhere outside of the n+/p+ implant regions, the n+ overlap with the n-well defines the STI clearance from the avalanche region. When a large enough clearance is used, the minority free carriers generated at the defective STI/silicon interface mostly recombine before diffusing into the avalanche region [171]. In this implementation, the n+ extends into the n-well by 0.75 μm, which places the STI at a sufficient distance away from the active region. The DCR induced by the defective STI sidewalls and edges can thus be kept to a minimum [240],[241].

Another key factor in the SPAD design arises due to the silicide of the source/drain junctions in standard CMOS technology [244]-[248]. A silicide-blocking mask layer over the active multiplication region is required to maintain a transparent path for incoming photons [249]-[250]. Otherwise, the PDE will be poor due to the reflections from the silicide layer. The silicide blocking layer, OP, was used to keep n+ region outside the metal contact area non-silicided [238]. The silicide was maintained only on the periphery of the active area where a ring of contact vias was used to connect the n+ cathode to the metal interconnect. Two identical test structures were fabricated on the same chip, with and without silicide on the active area, and these are shown in Fig. 3-2(b). The impact of silicide on the SPAD’s performance in terms of DCR and PDE results are presented in Chapters 4 and 5.

To ensure the functionality of the prototype test chips, a proper design of the bond-pad to which the HV is applied is required. In a typical CMOS IC design, floating terminal pad design rules require that a connection be made between bond pads and silicon. The purpose of these rules is to provide a DC connection between the bondpad and substrate in order to sink the excess charge on the wafer substrate during wafer processing [129],[254].

Charge build-up on each metal layer during the fabrication process (plasma etching, sputtering or chemical-mechanical polishing (CMP)) can lead to the phenomenon referred to as antenna effect (since the charge is attracted to each metal layer like an antenna) [254]. Providing a DC connection to ground from a terminal pad ensures that charging induced damage during wafer processing is avoided. The DC connection is typically made by inserting a minimum size n+/p-sub diode (antenna diode) which discharges the metal during the processing sequence. Since the small area diode is reverse biased under normal circuit operation, it has a negligible effect on the IC performance.

If the antenna diodes are connected to a bond-pad supplying the negative HV required for the SPAD breakdown voltage, then the antenna diodes will become forward biased when the negative HV is applied resulting in a very large current consumption and IC malfunction. Therefore, antenna diodes should never be connected the bond-pads applied to HV. This design does not flag a DRC violation, since the p-well/DNW diode to which the HV bond-pad is connected to already provides a DC connection to ground and thereby eliminates the antenna effect.

In addition to antenna diodes, all the bond-pads of an IC require electro-static discharge (ESD) protections in order to prevent any harmful over-voltages that may damage the IC internal transistors [254]. The ESD diodes will become forward biased and will draw current from the substrate ground potential when the HV is applied to bias the SPAD above the breakdown voltage. Therefore, the ESD diodes must be removed from the bond-pads that are connected to HV as well. This may flag DRC violations, but the designer can obtain a waiver for these design rules, since the violations do not affect any other chips on the wafer.

On the other hand, the MIM capacitors are very susceptible to ESD. If a MIM capacitor be connected to the HV pad for decoupling purposes, then this would flag MIM DRC violations, requiring additional DRC waivers. This would be in addition to those required for the antenna diodes removal. For these reasons it is essential that proper handling procedures of the IC are followed. Proper handling will prevent ESD damage to the HV pad.

Due to the strict metal density rules in this technology, the SPAD must be surrounded by metal. This was managed by surrounding the SPAD active area with top-level dummy metal, which provides the added benefit of shielding the guard-ring area from photons. This prevents the absorption of photons in the guard ring region which can result in higher noise and higher afterpulsing probability. The top tier of thick metal layers provides the least parasitic capacitance and resistance for routing the cathode signal, but they require contact vias with very large areas according to the design rule manual. So only the top metal of lowest metal tier (metal 3) was used to route the cathode signal to a quenching resistor placed near the SPAD.

In this work, a 50 kΩ non-silicided high-resistivity polysilicon quenching resistor was used. This was because passive quenching is simple to implement, requires a minimum of power and pixel area, and is very robust [125],[155]. More details of the passively quenched SPAD pixels are given in Section 3.3

### I-V and Breakdown Voltage Measurements

In the n+/p-well/DNW/p-substrate structure, parasitic diodes are formed between the p-well and DNW and also between DNW and substrate [147]. These diodes must remain reverse biased below breakdown at all times in order for the SPAD to operate properly [234]. When the negative HV is applied to the anode in order to bias the SPAD above its breakdown voltage, only the n+/p-well junction should undergo avalanche breakdown. According to the IBM 130 nm CMOS design manual, the breakdown voltage of the DNW/p-substrate diode is approximately -9 V. Connecting both the n+ contacts for deep n-well and p+ contacts for p-substrate to ground potential, as in the SPAD structure shown in Fig. 3-2(c), ensures that the parasitic diode does not to undergo avalanche breakdown.

It should be noted that the DNW/p-well parasitic junction is susceptible to breakdown since the p-well anode is connected to the negative HV and the DNW cathode is connected to ground potential. Therefore, the breakdown voltage of the p-well/DNW had to be measured along with the breakdown voltage (BV) of the SPAD. This allows the determination of the maximum voltages that can be applied for the HV and IC supplies to bias the SPAD anode without exceeding the breakdown voltage of the parasitic diode.

The SPAD test structure, illustrated in Fig. 3-3, was fabricated in order to measure its breakdown voltage. I-V characteristics of the SPAD test structures were measured using an Agilent B1500A Semiconductor Parameter Analyzer (SPA). The source measurement units (SMU) of the SPA were connected to the various external leads, *VDD\_SPAD*, *VHV*, and *GND* on the circuit board housing the PGA68 package. To measure the BV, the voltage was swept between -7 and -14 V, using a 10 mV step size and 1 mA current compliance. When the SMU signal was applied to *VHV*, with *vC* common and *GND* and *VSPAD*floating, then the current only flows through the SPAD (blue in Fig. 3.3(b)). Similarly, when SMU signal is applied to *VHV* with *GND* common and *VC* and *VDD\_SPAD* floating, then the current only flows through the parasitic (red) diode. The measured I-V curves are shown for both the SPAD and parasitic diode in Fig. 3-3(b). The BV was obtained by taking the voltage at the peak value of the first derivative of the I-V curve [157]. The breakdown voltages for the DNW/p-well parasitic junction and SPAD junction were found to be approximately 9.5 V and 11.4 V, respectively.

Since 9.5 V is the maximum negative HV that can be applied to the p-well anode without parasitic diode avalanche breakdown, the SPAD voltage supply, *VDD\_SPAD*, with a minimum voltage of 2.4 V must be connected to the cathode in order to exceed the SPAD breakdown voltage of 11.4 V. However, the maximum supply voltage that can be applied to the cathode is limited by the transistor gate-oxide breakdown voltage.

In this IBM 130 nm CMOS technology, I/O transistors with a thicker 52 Å oxide, for which the maximum voltage supply is 3.6 V, are available [254]. These transistors can be directly connected to the SPAD because they are compatible with the voltage levels required to operate the SPAD. From voltage headroom considerations in Fig. 3-3(c), a 3.6 V supply voltage means that a maximum excess bias of 1.3 V can be applied to the SPAD with I/O transistors connected directly to the cathode.

|  |  |
| --- | --- |
|  |  |
| (b) |
|  |
| (a) | (c) |

Figure 3‑3: (a) Test structure used to evaluate SPAD breakdown voltages. (b) Measured breakdown voltages of SPAD and parasitic junction. (c) Illustration of the voltage headroom limits for proper SPAD operation.

|  |  |
| --- | --- |
|  |  |
| (b) |
|  |
| (a) | (c) |

Figure 3‑6: (a) Test structure used to evaluate SPAD breakdown voltages. (b) Measured breakdown voltages of SPAD and parasitic junction. (c) Illustration of the voltage headroom limits for proper SPAD operation.

For reliability reasons, it is essential to not exceed the maximum gate oxide potential of the front-end transistors, otherwise permanent oxide damage can occur. For this reason, if higher excess voltages are required, an additional 400 mV of excess bias can be obtained by raising HV to 9.4 V. This is the maximum limit, since the leakage current of the parasitic DNW/p-well diode increases when the HV bias is increased towards the parasitic junction breakdown voltage of 9.5 V, resulting in an increased DCR of the SPAD as shown in section 3.3. In this section for SPAD measurements, in which front-end transistors are included with the SPAD, the HV is kept at approximately 9 V, resulting in a maximum excess voltage of approximately 1.3 V.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 3‑4: (a) Measured breakdown voltage as a function of temperature for 7 randomly chosen SPADs. (b) I-V curves for SPAD3 between -40 °C and 60 °C.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 3‑8: (a) Measured breakdown voltage as a function of temperature for 7 randomly chosen SPADs. (b) I-V curves for SPAD3 between -40 °C and 60 °C.

I-V measurements were performed by varying the temperature of the SPAD between -30 to +40 °C in a sealed environment of a temperature chamber to assess temperature variations of the breakdown voltages. The temperature stability of the chamber was ±3 °C. There was standard deviation σ of 0.1 V in the BVs of 7 chips at room temperature. The average SPAD breakdown voltage was 11.4 V. A temperature coefficient of 7.2 mV/°C was extracted from the slopes of the linearly fitted data and is similar to the value of 7.14 mV/°C reported in [85].

These temperature coefficients are low compared to SPADs in DSM technologies that use p-well/DNW junctions (40 mV/°C in [164], and 20 mV/°C in [172]), which also had higher *VBRK*. This indicates that the tunneling mechanism contributes to the breakdown of the SPADs fabricated in this work. Since the SPADs are sensitive to temperature variations, a temperature control with fast response and low thermal resistance is essential. However, the breakdown temperature coefficient of these SPADs is rather low compared to other devices. Therefore, only small voltage adjustments to maintain a fixed excess voltage across all operating temperatures are required. In all reported temperature measurements hereafter (Section 3.3), the breakdown voltage variation is taken into account by adjusting the bias voltage *VDD\_SPAD*.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 3‑5: Comparison of breakdown voltages of non-silicided (SPAD1) and silicided (SPAD2) devices (a) Measured I-V curves at room temperature for 5 different chips. (b) Temperature variation of breakdown voltage for SPAD1 and SPAD2.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 3‑10: Comparison of breakdown voltages of non-silicided (SPAD1) and silicided (SPAD2) devices (a) Measured I-V curves at room temperature for 5 different chips. (b) Temperature variation of breakdown voltage for SPAD1 and SPAD2.

Breakdown voltages of the silicided and non-silicided SPAD test structure are compared in Fig 3-5. Fig. 3-5(a) shows the measured I-V curves at room temperature. The optimal avalanche characteristics are represented by a very sharp and sudden rise in current above the breakdown voltage, indicating full volumetric breakdown of the planar junction [157]. As can be seen from the figure, a sudden and rapid rise in the current occurs at a smaller voltage for the non-silicided SPAD compared to the silicided SPAD, giving a rough indication that the non-silicided SPAD is of higher quality than the silicided one. The non-silicided and silicided SPADs had breakdown voltages of approximately 10.5 V and 11.5 V, respectively. In Fig. 3-5(b), the variation of breakdown voltage with temperature is shown. Both SPADs have a similar temperature coefficient of breakdown voltage.

### SPAD pixel circuit modeling

Detailed and accurate SPAD models for circuit simulation are essential to correctly predict the static and dynamic SPAD behavior when designing the in-pixel circuits [256]-[260]. The SPAD model shown in Fig. 3-6 was utilized in this work to accurately design the SPAD pixels [256]. The temporal properties of SPAD behavior are accurately represented in the circuit model and both the avalanche build-up and self-quenching mechanisms are taken into account. The switches together with the inductors, represent relays that switch on whenever a threshold is crossed.

|  |
| --- |
|  |

Figure 3‑6: Circuit model used for simulating passively-quenched SPAD [256].

|  |
| --- |
|  |

Figure 3‑12: Circuit model used for simulating passively-quenched SPAD Cadence [256].

A pulse source connected to the ‘Photon’ terminal simulates an incoming photon by closing the switches STRIG and SSELF to discharge the pre-charged SPAD’s capacitance (*CSPAD*) through the SPAD resistor. This discharge results in a large and fast current spike through the HV supply (SPAD anode), causing switch *STRIG* to close immediately in response to the current spike, which in turn causes S1 to open. This eliminates the effect of the simulated photon pulse width on the duration of the simulated avalanche. The avalanche current duration is determined by the switch *SSELF*, which opens whenever the avalanche current drops below a predefined threshold. In the SPAD’s literature, this is typically set at 100 μA [135],[258]. When switch *SSELF* opens, the avalanche current is stopped completely, and the only current remaining is due to the passive recharge.

Time-domain simulations of the SPAD were performed using the Cadence Spectre circuit simulator. The circuit schematic used to simulate the passively quenched test structure is shown in Fig. 3-6. SPAD parameters were extracted from fabricated test structure. These parameters ensured that the model precisely reflects the static and dynamic avalanche breakdown characteristics. The SPAD circuit model was implemented with the following parameters: *RQ* = 50 kΩ, *CSPAD* = 250 fF, *Cload* = 10 pF, *RSPAD* = 600 Ω, *VHV* = -9 V, and *VDD\_SPAD* between 2.4 and 3.6 V.

The simulated and measured cathode voltage, anode current and VDD current during passive quenching and recharge are shown in Fig. 3-7(a). The left and right panels show the voltage and current pulses using microsecond and nanosecond time scales, respectively. The simulated and measured cathode voltages are in the top panels, showing good agreement during passive recharge. However, during quenching, there are some differences between the measured and simulated cathode voltage waveforms.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑7: (a) Simulated and measured cathode voltage, anode current and VDD current during passive quenching and recharge. (b) Simulation of cathode voltage and comparator output during avalanche re-triggering for different delays ΔT.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑14: (a) Simulated and measured SPAD waveforms during passive quenching and recharge. (b) Simulation of cathode voltage and comparator output during avalanche re-triggering for different delays ΔT.

The parasitic capacitance of the bond-pad, *Cload*, as well as the parasitic inductance of the packaging bond wire can create a resonance which causes distortions and small oscillations whenever a large current spike on the anode terminal occurs [254]. Since the simulated anode current spike, shown in the middle panels of Fig. 3-7(a), reaches a peak value of 6 mA in less than a nanosecond, then this spike can cause some small ripple on the cathode voltage through parasitic coupling. In addition, as the quenching time is due to the SPAD’s capacitance, *CSPAD*, and internal resistance, *RSPAD*, the uncertainty of these parameters will translate into uncertainty of the modeled quenching time.

Once the avalanche is triggered at *t* = 0, the cathode voltage, and hence the anode current, begin to decay, eventually falling below the latching current. The avalanche is fully quenched when this occurs and the SPAD begins recharging through the quenching resistor, *RQ*. The recharge current flowing through *RQ* is several orders of magnitude smaller than *RSPAD*, therefore the recharge time is in the microsecond range.

Simulations of the passively quenched structure were performed with *VEX* = 1.2 V and *VBRK* = 11.3 V. Initially, the cathode voltage is equal to *VDD\_SPAD* = 3.5 V since no current flows through the SPAD. When the avalanche is triggered at *t* = 0, the cathode voltage *VC* begins discharging according to

|  |  |
| --- | --- |
|  | (3-1) |

where *IBRK* is the breakdown current, *τQ* = *RSPAD*×(*CSPAD* + *CProbe*) is the quenching time constant and *RSPAD* is the resistance of the SPAD in avalanche breakdown. During the passive recharge phase, the voltage *VC* has an initial value roughly equal to (*VDD\_SPAD* ‒ *VEX*) and the cathode voltage is

|  |  |
| --- | --- |
|  | (3-2) |

where *VEX* = (*VSPAD* + |*VHV*|) ‒ *VBRK*, *τR* = *RQ*×(*CSPAD* + *CProbe*) is the recharge time constant and *t0* is the time it takes the passive quenching to exit Geiger mode, which is given by [84]

|  |  |
| --- | --- |
|  | (3-3) |

The excess bias has a negligible effect on the quenching time, because the avalanche process is very fast. The product of the quenching time with peak avalanche current gives an approximation to the total amount of avalanche charge, indicating the degree to which afterpulsing effects are expected to contribute to the SPAD’s performance. The total simulated avalanche charge with *Cload* = 10 pF was approximately 11.7 pC. Since *RQ* >> *RSPAD*, the recharge current is much smaller than the quenching current, and it has a much longer time constant. The recharge time constant was *τR* = 500 ns, and the time required for the SPAD to recover to 90% of the full excess voltage was 2.3·*τR* = 1.15 μs.

With passive quenching, the SPAD is not completely insensitive during the recharge process. It progressively recovers its excess voltage and is susceptible to re-triggering of the avalanche [135]. Fig. 3-7(b) shows the simulated SPAD cathode voltages and comparator outputs for different delays *ΔT* between a primary avalanche and re-triggered avalanche events. The re-triggered avalanches have smaller amplitude than the primary avalanche since *VEX* is very low immediately following the primary avalanche. This causes degradation of resolution in photon timing at high counting rates and leads to uncertainty of the passively quenched dead time.

The dead-time for passive quenching can be defined as the time required to quench the avalanche process plus the time to recharge the SPAD to 90% of *VDD\_SPAD*. When the threshold level is set in this manner, an extension of the output pulse width results when the avalanche is re-triggered, as shown in the bottom panel of Fig. 3-7(b). When a lower threshold value is used, rather than extending the output pulse width, the re-triggered avalanches instead cause re-trigging of the output pulse, leading to an increase in the measured count rate. The presence of any parasitic load capacitance in parallel with *CSPAD* increases the effective SPAD capacitance which affects the device behavior with respect to avalanche re-triggering. A larger capacitance results in a higher avalanche current and leads to longer recharge time. Therefore, to minimize the current flow associated with avalanche events, as well as to decrease the time constants governing the circuit’s response times, it is desirable to minimize *CSPAD*. The effects of SPAD capacitance will be discussed in detail in the next section when describing the design of the front-end circuits.

## Passively-Quenched Pixel Designs and Measurements

The SPAD’s pixel front-end circuits are very important since their design directly affects the SPAD’s performance [132]-[138]. They should be able to quickly sense the avalanche as well as to minimize the quenching time and the timing jitter of the resulting digital output pulse. DC coupling of in-pixel circuits is preferred over AC coupling since this eliminates a coupling capacitor that takes up valuable pixel area. Passive quenching (PQ) is used for all the SPAD pixels designs in this work because PQ satisfies the miniaturization requirements in terms of smaller pixel size and simpler in-pixel circuits, resulting in higher pixel fill-factor and lower SPAD capacitance. However, PQ does have some significant drawbacks.

First, the quenching resistor must be large enough to ensure that the avalanche process can be fully quenched [135]. Second, since the SPAD is susceptible to re-triggering during the excess voltage recovery transient, the SPAD output begins to saturate for high counting rates [138],[259]. This results in degradation of overall detector performance at very high count rates because the SPAD is unable to recover to the full excess voltage before the next incoming photon triggers an avalanche. For characterization purposes, a high counting rate and/or short dead-time is not a critical requirement, therefore passive quenching represents a simple and perfectly satisfactory mode of operation.

In the following sections, measurements of output pulses from the unbuffered SPAD test structures are presented. This is to assess the performance of different SPAD structures in the free-running mode and to characterize the statistical variation of the SPAD output pulses at different bias voltages and temperatures.

### Unbuffered SPAD pixels

The measurement set-up is shown in Fig. 3-8(a). The SPAD cathode voltage *vC* was connected directly to the IC output bond-pad (with load capacitance *CBP* ≈ 40 fF) and to switches on the PCB that connect the SPAD cathode either to an oscilloscope probe or the inverting input of an external comparator IC (with load capacitance *CIN* ≈ 1 pF) whose output connects to an oscilloscope. A MATLAB interface was used for setting the measurement parameters (bias voltage, measurement time, instrument control, etc.), and for data acquisition, post-processing and plotting. The measurements of waveform properties such as amplitude and pulse width (time between negative slope and positive slope crossings of a fixed voltage threshold) were collected in histograms by the oscilloscope.

Since no on-chip transistors are connected to the unbuffered SPAD, larger values of *VEX* could be measured because the gate oxide voltage limitations are not imposed. Increasing *VEX* is useful for studying count rate saturation effects and tunneling effects, as shown in Chapter IV. Also, larger *VEX* can be used to increase the PDE, as will be shown in Chapter IV. On the other hand, the large load capacitance (*CL* ≈ 1-10 pF) results in a very long recharge time and a large avalanche current for unbuffered SPAD. A larger avalanche current means more charge trapping and higher probability of afterpulsing.

Afterpulsing effects can be seen in Fig. 3-8(b) which shows an oscilloscope trace of the unbuffered SPAD cathode voltage for *VEX* = 1 V. Several afterpulses occur in quick succession at around *t* = 4 μs and *t* = 5 μs during the recovery time of the previous avalanche. The afterpulses have smaller peak-to-peak amplitude since *VEX* is not fully restored to its final value when they are triggered. As a result, there exists a large variation in the amplitude of the avalanche pulses.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑8: (a) Unbuffered SPAD test structure. External switches S1 and S2 enable selection between two different load capacitances for the SPAD chip. (b) Typical cathode waveform and illustration of comparator outputs for two different thresholds.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑16: (a) Unbuffered SPAD test structure. External switches S1 and S2 enable selection between two different load capacitances for the SPAD chip. (b) Typical cathode waveform and illustration of comparator outputs for two different thresholds.

The long recharge time constant of the unbuffered SPAD test structure leads to another important effect, namely saturation of the SPAD at higher counting rates. When high intensity light is incident on the SPAD, a recharge time on the order of microseconds means that most of the avalanches will occur before the SPAD has had time to fully reset, resulting in the saturation of the measured count rate. Nevertheless, provided that the light intensity is low enough or the SPAD is kept in the dark, then the SPAD will be far from saturation. In this case, the amplitude and duration of the pulses can give important information and insight into the SPAD’s behavior and performance, which can then be used to verify the SPAD circuit model.

Opening the switch *S1* and closing switch *S2* in Fig. 3-8(a) disconnects the oscilloscope probe to reduce the capacitive loading and connects the cathode to a comparator IC with a much lower input capacitance (1 pF). The high-speed comparator IC has an externally controlled threshold voltage *VTH*. Output pulses are produced for each avalanche discharge upon crossing of the threshold by the cathode voltage. Although the cathode voltage can no longer be observed directly by the oscilloscope, and thus information of the pulse amplitudes and pulse shape is no longer available, the reduced input capacitance of the comparator IC results in shorter output pulses and less charge flowing through the SPAD, which can reduce the afterpulsing effects [163],[181].

Fig. 3-8(b) shows the simulated comparator output for two different threshold voltages. Different comparator threshold voltages can be used to examine the effects of afterpulsing. When *VTH1* is used, afterpulses can be detected by measuring variations in the output pulse width. When *VTH2* is used, then the afterpulses can be counted directly.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑9: (a) Measured SPAD pulse amplitude histograms for as a function of excess voltage (b) Measured average pulse amplitudes as function of *VEX* of eight different SPAD chips. Saturation effects are prominent for SPAD5 and SPAD8.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑18: (a) Measured SPAD pulse amplitude histograms for as a function of excess voltage (b) Measured average pulse amplitudes as function of *VEX* of eight different SPAD chips. Saturation effects are prominent for SPAD5 and SPAD8.

Fig. 3-9(a) shows the measured avalanche voltage pulse amplitude histograms at room temperature. The amplitude of each avalanche pulse is based on the total amount of charge generated during the avalanche. Since avalanche multiplication is a Poisson process, the amplitude variance is equal to the amplitude mean, resulting in a wider output pulse amplitude distribution as *VEX* increases.

The average avalanche pulse amplitude increases almost linearly with excess voltage as shown in Fig. 3-9(b) for eight different measured SPAD chips. At very high *VEX*, the amplitudes begin to saturate. Two pixels, SPAD5 and SPAD8, attained saturation of the pulse amplitude for relatively lower values of *VEX* compared to the other pixels. This behavior is an indication of a defect occurring in the active region leading to a greater DCR contribution from tunneling effects. In these defective SPADs, the cathode voltage is prevented from completely recovering to the full excess voltage before another avalanche occurs, because the average IAT of dark pulses is comparable to the recharge time constant of the SPAD as *VEX* increases. Defective pixels can be identified in this manner by measuring the linearity of average pulse amplitudes over the excess voltage range.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑10: Measured pulse width histograms of unbuffered SPAD pixels for (a) *CSPAD* = 1 and 9.5 pF and (b) *VEX* = 1.5 and 2.5 V.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑20: Measured pulse width histograms of unbuffered SPAD pixels for (a) *CSPAD* = 1 and 9.5 pF and (b) *VEX* = 1.5 and 2.5 V.

The threshold of the comparator was set to according to *VTH* = *VDD\_SPAD* ‒ η*VEX* where η = 0.3. Thus for *VEX* = 1.5 V, the SPAD and comparator voltages were *VDD\_SPAD* = 3.9 V and *VTH* = 3.68 V respectively. When a lower comparator threshold is set (η = 0.5), then the afterpulses are more likely to trigger output pulses from the comparator, resulting in a higher DCR. As *VEX* increases, a larger percentage of avalanches occur during the passive recharge time resulting in a larger variance for the output pulse width. When the pulse width histograms are examined in Fig. 3-10, several smaller peaks following the primary peak are evident. This indicates that the SPAD can be re-triggered several times during the recharge period.

When the SPAD is re-triggered during its recharge time, the cathode voltage does not cross the comparator threshold and trigger a falling edge on the output pulse at the same time as it does when there is no avalanche re-triggering. So, when the comparator threshold is near *VDD\_SPAD*, instead of registering a dark count, the afterpulses cause prolongation of the output pulse width. Fig 3.10(a) shows that the pulse width distribution has a single dominant peak for *CSPAD* = 10 pF, and several other smaller peaks each with similar amplitude. The dominant peak corresponds to the case when a complete recharge occurs without re-triggering, and this is the most probable case for the dark counts. However, the smaller peaks following the dominant peak indicate that the avalanche is re-triggering during the passive recharge.

The histograms in Fig. 3-10 show that when *CSPAD* ≥ 1 pF, secondary avalanches can be re-triggered many times following a primary avalanche, even at room temperature. The rate of re-triggering is expected to decrease as SPAD capacitance is reduced. The output pulse width distributions are shown for two excess biases with *CSPAD* = 1 pF in Fig. 3-10(b). In this case, the average pulse width is reduced on account of the smaller *RC* recharge time constant.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑11: Mean and standard deviation of pulse widths as a function of *VEX*. (a) Measured results of eight chips for *CSPAD* = 1 pF. (b) Measured mean and standard deviation of pulse widths of SPAD1 as a function of *VEX* for two different *CSPAD* values.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑22: Mean and standard deviation of pulse widths as a function of *VEX*. (a) Measured results of eight chips for *CSPAD* = 1 pF. (b) Measured mean and standard deviation of pulse widths of SPAD1 as a function of *VEX* for two different *CSPAD* values.

Reducing *CSPAD* speeds up the SPAD quenching and recharge time, but the rate of decay of the trapped carrier population is unaffected by the reduced capacitance. This means that for *CSPAD* = 10 pF, most of the trapped charges were released while the SPAD excess bias was still very low, which leads to a lower re-triggering rate. For *CSPAD* = 1 pF, most of the trapped charges were released while the SPAD was almost completely recharged, leading to a larger re-triggering rate, hence the appearance in the histograms of a double peak immediately following the dominant peak.

In Fig. 3-11, the plots of the average output pulse width and the standard deviation are shown for different measured SPADs as a function of *VEX* for *CSPAD* = 1 pF. Since higher *VEX* results in higher frequency of tunneling events, pulses with shorter inter-arrival times become more likely and the DCR increases. Thus, a larger portion of the avalanches occur during the recharge phase of a previous avalanche, which leads to large output pulse width variance. The pixels *SPAD5* and *SPAD8* are evidently defective, since the pulse width variance increases very rapidly with *VEX*. All the other pixels had a similar variation of pulse width variance, indicating a high degree of DCR uniformity. The DCR and afterpulsing performance of these pixels will be examined in greater detail in Chapter IV.

### Free-running (FR) Source-follower SPAD (SF-SPAD) pixel

In CIS, the standard 3-transistor (3T) or 4-transistor (4T) active pixel sensors (APS) employ an in-pixel source-follower (SF) amplifier to transfer the signal voltage at the photodiode to the output. Since the output signal of an APS pixel is an analog voltage, the unity gain SF is ideal for bringing the signal to on-chip or external amplifiers. In SPAD design, a SF amplifier can be used to buffer the SPAD cathode from the large output bond-pad and oscilloscope probe capacitance.

An SPAD with in-pixel SF amplifier pictured in Fig. 3-12(a) was fabricated in order to study the dynamic and statistical properties of avalanche pulses for an excess voltage range of 0.2 – 1.2 V over the temperature range between -40 – 30 °C. According to the equations for the gain and output resistance of a nFET source follower with current source load,

|  |  |  |
| --- | --- | --- |
|  |  | (3-4) |

the transconductance *gm1* of *M1* and output resistance *ro2* of current source transistor *M2* should be very large to achieve a near-unity gain. The SF amplifier also needs very high slew rate in order to discharge a very large output capacitance during each avalanche. The slew rate of the SF stage is given by

|  |  |
| --- | --- |
|  | (3-5) |

indicating that a large bias current is required to quickly discharge the load capacitor. If 1 V is discharged across a 10 pF capacitor within 1 ns, then the resulting current spike is

|  |  |
| --- | --- |
| . | (3-6) |

Therefore, a 40 mA current source *IL* was used to bias a very large nFET transistor *M1*, providing *gm1 =* 110 mS, while *ro2* was 3.3 kΩ, resulting in *AV* ≈ 1 with *CL* = 10 pF. The circuit was simulated over the entire voltage and temperature operating range, achieving good performance stability in terms of amplifier characteristics such as gain and temperature insensitivity.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑12: (a) SPAD front-end with nMOS source follower and current-source load. (b) simulated source-follower gain and bias current variation versus temperature for different process corners.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑24: (a) SPAD front-end with nMOS source follower and current-source load. (b) simulated source-follower gain and bias current variation versus temperature for different process corners.

The simulation results in Fig. 3-12(b) shows that the gain of the SF is nearly unity within the SPAD operating voltage range and the bias current has low temperature sensitivity. The simulated input capacitance of the SF was 0.4 pF, which is 20x smaller compared to the load capacitance of the unbuffered SPAD. This corresponds to a SPAD recharge time constant of *τR* = 20 ns. Thus, the SF effectively isolates the SPAD’s cathode from the large load capacitance and can drive an oscilloscope directly for pulse shape analysis.

Fig. 3-13(a) shows a simulated and measured output pulse for the case when the avalanche is re-triggered before the cathode voltage has fully recharged from a previous avalanche (right panel), and when an avalanche is not re-triggered (left panel). Ideally, the SF output should be an exact replica of the input. However, in practice, the bond wire inductance and probe capacitance resonance causes ripples in the power supply rails of the IC whenever an avalanche current spike occurs. Also, the mismatched impedance condition when connecting to a high-impedance oscilloscope probe results in signal reflections, causing ripple that can couple to the SPAD and be observed as an oscillation of the measured output signal. Nevertheless, the measured and simulated results show good agreement, with PW1 = 25 ns and PW2 = 30 ns. In spite of the amplitude ripple at the output, secondary avalanches occurring during the recharge time could be detected within approximately 15 ns of the primary avalanche.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑13: (a) Simulated and measured waveforms for SPAD/SF pixel. PW2 > PW1 due to an afterpulse occurring during the recharge time. (b) Measured pulse amplitude distributions for different temperatures.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑26: (a) Simulated and measured waveforms for SPAD/SF pixel. PW2 > PW1 due to an afterpulse occurring during the recharge time. (b) Measured pulse amplitude distributions for different temperatures.

Fig. 3-13(a) shows that an afterpulse occurring within 15 ns can be differentiated from the output ripple based on the amplitude of the secondary spike, which attains the same minimum value as the primary pulse. On the other hand, the afterpulses are much better resolved by their amplitude as their delay with respect to the primary pulse increases, since the amplitude ripples decay with time. The distribution of avalanche pulse amplitudes was obtained to examine afterpulsing effects of the SF-SPAD, which could attain much shorter times scales compared to the unbuffered SPAD test structures.

Fig.3-13(b) illustrates the pulse amplitude distributions obtained at -30 and 40 °C with *VEX* = 0.8 V with the SF-SPAD. At -30 °C, there is a constant tail in the amplitude distribution which is due to the avalanches that are re-triggered during the recharge time of the primary avalanche. Since the trap lifetimes are increased at lower temperatures, then the possibility of trapped charges being released during the recharge and the re-triggering of the avalanche is increased. This in turn leads to more avalanches having reduced amplitude since they occur when *VEX* is recharging. As the temperature is raised, the tail in the distribution disappears since the trapping lifetimes decrease exponentially. Thus, most of the trapped charges are released while the SPAD is quenched, which in turn reduces the probability of avalanche re-triggering during the recharge time.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 3‑14: Average pulse amplitudes of SPAD-SF pixel as a function of (a) excess voltage, and (b) temperature.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 3‑28: Average pulse amplitudes of SPAD-SF pixel as a function of (a) excess voltage, and (b) temperature.

Fig. 3-14 shows the measured variation of the SF-SPAD output pulse amplitude and width as a function of voltage and temperature. The temperature coefficient of the breakdown voltage of 7 mV/°C was taken into account when applying the bias to the SPAD so that *VEX* is nearly constant over all temperatures. As a result, the average amplitude of the output pulses should be the same at each temperature. However, due to temperature drift in the chamber (±3 °C) and uncertainty in the breakdown voltage and temperature coefficient of the measured chip, the average pulse amplitudes deviate slightly from the ideal values as the temperature is varied. For the measured chip in Fig. 3-14, the average amplitudes appear to be increasing slightly with temperature, indicating that the breakdown voltage temperature coefficient for this particular SPAD is slightly greater than 7 mV/°C.

Fig. 3-15(a) shows the measured average output pulse amplitude as a function of temperature for the SF-SPAD pixel. Although the average pulse width has been reduced to 14 ns, which is more than an order of magnitude lower that for the unbuffered SPAD pixel, the mean and variance of the output pulse width increases rapidly as temperature is reduced. This increase is due to the longer decay times of trapped charges, leading to trapped charges being released during the SPAD recharge which prolongs the eventual re-charge. Fig. 3-15(b) shows that the pulse width mean and standard deviation increases more rapidly with *VEX* at lower temperatures. This is similar to the behavior of the unbuffered SPAD at room temperature, which showed an increase in pulse width standard deviation with *VEX* due to the long recharge time. These results indicate that the afterpulsing probability is expected to increase with the number of avalanche charges, since more charges are trapped and then subsequently released.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑15: Average pulse widths of SPAD/SF pixel as a function of (a) excess voltage and (b) temperature.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑30: Average pulse widths of SPAD/SF pixel as a function of (a) excess voltage and (b) temperature.

Although the SF-SPAD was effective in reducing the dead-time by effectively shielding the cathode from the large capacitive load, the power consumption of the pixel was too high to be implemented in pixel arrays. Therefore, in the next section, an in-pixel common-source (CS) amplifier together with a CMOS inverter output buffer sized to drive large capacitive loads was utilized for lower power consumption.

### Free-running (FR) Common-Source SPAD (CS-SPAD) pixel

It was shown that for passively quenched SPAD pixels, the average width of the output pulses defines the nominal SPAD dead-time. The variance in pulse width represents the degree to which afterpulsing contributes to the DCR. The first pixel studied used unbuffered SPAD test structures which were characterized by a long recharge time constant, leading to a high rate of re-triggering during the recharge and output pulse width variations. Although the SF-SPAD pixel was successfully used to isolate the SPAD from a large load capacitance, leading to a significant reduction in the average dead-time, and a corresponding decrease of the re-triggering rate, the SF-SPAD required a very large nMOS transistor and large current in order to quickly discharge the load capacitance, leading to a very low pixel fill factor and high power consumption.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑16: (a) Schematic of CS-SPAD pixel using common-source amplifier with current source load. (b) Simulated and measured waveforms with relative positions of V*TH1* and V*TH2* labeled.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑32: (a) Schematic of SPAD pixel using common-source amplifier with current source load. (b) Simulated and measured waveforms with relative positions of V*TH1* and V*TH2* labeled.

A dual-threshold, passively-quenched CS-SPAD front-end circuit was developed with emphasis on miniaturization and power consumption and was targeted for a shorter dead-time for higher counting rate operation. The CS-SPAD pixel shown in Fig. 3-16(a) uses a comparator circuit with a threshold of V*TH1* = (*VDD* - 400) mV to sense avalanche pulses on the cathode. A fixed comparator threshold is set near *VDD\_SPAD*. As a result, afterpulses that occur during the passive recharge phase extend the cathode voltage recharge duration and increase the variance of the digital output pulse width rather than re-trigger new output pulses. The comparator output *vD* is input to a CMOS inverter that has a threshold of V*TH2* = 800 mV. The output buffer consists of a cascade of CMOS inverters to drive a large capacitance load of ~ 10 pF.

Fig. 3-17(a) shows the simulation results and a measured output pulse of the proposed dual-threshold front-end circuit. The rising edge of the output pulse rising is triggered when the cathode voltage *vC* falls below V*TH1*, and the falling edge is triggered when the drain voltage *vD* falls below V*TH2*, thus providing a full-swing digital pulse at the output. Avalanches that occur before *vD* crosses V*TH2* are not converted to output pulses, but are responsible for prolonging the output pulse width because the recharging restarts with each afterpulse. A dead-time of ~30 ns was chosen as a suitable compromise between the requirements for low afterpulsing (dead-time greater than longest trap lifetime), and a high counting rate.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑17: (a) Left: Simulated I/O characteristic of CS front-end shows that relative positions of V*TH1* and V*TH2* are unchanged at different temperatures Right: Measured and simulated output pulse widths at different excess voltages. (b) Simulated and measured SPAD output pulses that illustrate the effects of afterpulsing on the output pulse width (PW). PW1 > PW2 due to an afterpulse occurring during the recharge time.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑34: (a) Left: Simulated I/O characteristic of CS front-end shows that relative positions of V*TH1* and V*TH2* are unchanged at different temperatures Right: Measured and simulated output pulse widths at different excess voltages. (b) Simulated and measured SPAD output pulses that illustrate the effects of afterpulsing on the output pulse width (PW). PW1 > PW2 due to an afterpulse occurring during the recharge time.

The I/O voltage characteristics of the front-end circuit were simulated at room temperature and T = -30 °C, and the results are shown on the left of Fig. 3-17(a). Although the threshold values are insensitive to changes in temperature, simulation and measurement results in Fig. 3-17(b) indicate that the pulse width is sensitive to the SPAD supply voltage, *VDD\_SPAD*, since *VEX* = *VDD\_SPAD* + *VHV* – *VBRK*. As the threshold is fixed by the value of *VDD* = 3.6 V, the time between threshold crossings increases as *VDD\_SPAD* is reduced, resulting in wider output pulses.

Fig. 3-17(b) shows the CS-SPAD output pulses at room temperature with *VEX* = 1.2 V. The case of an extended output pulse occurring when the avalanche is re-triggered during the recharging time is shown, as well as the case of a successful recharge without re-triggering during recharge. An avalanche occurs at *t* = 50 ns, followed by another avalanche that occurs during the recharge time of the SPAD. The simulation results show that a second avalanche occurring at *t* = 75 ns is detected by the output circuit. However, the measurements show that this second pulse cannot be detected by the output circuit, since the falling edge of the first pulse and the rising edge of the second pulse merge together to form an extended pulse with PW1 > 50 ns. A successful recharge without an avalanche re-triggering occurs at t = 200 ns, where a single pulse occurs with PW2 ≈ 30 ns.

Afterpulsing become much more prominent at lower temperatures, as evidenced by the increased variation of the pulse width. The distribution of output pulse widths was measured T = –30 °C and at room temperature to compare the effects of afterpulsing in the CS-SPAD. The measurements were performed in the dark and with background illumination from a halogen lamp. The distribution of output pulse widths is plotted on a semi-log scale in Fig. 3-18(a). There are several peaks in the histogram at –30 °C in addition to an exponential decay portion corresponding to afterpulses. The valleys after the peaks correspond to the reduced avalanche triggering probability immediately following an avalanche, as was shown in Fig. 3-10 for the unbuffered SPAD at room temperature. The histogram data, shown on a linear scale, in the insets indicates that two afterpulses can typically occur during the passive recharge for the CS-SPAD at low temperatures. However, these afterpulses do not contribute to the measured afterpulses since their effect is only to prolong the output pulse width.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑18: (a) Measured pulse width distribution in the dark and with light at -30 °C and room temperature at *VEX* = 3.6 V. (b) Measured pulse width (with standard deviation error bars) as a function of temperature at two different excess voltages.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑36: (a) Measured pulse width distribution in the dark and with light at -30 °C and room temperature at VEX = 3.6 V. (b) Measured pulse width (with standard deviation error bars) as a function of temperature at two different excess voltages.

At room temperature, there is a single dominant peak in the histogram and a much shorter decay portion. This indicates that at elevated temperatures, the majority of the trapped charges are released during the initial portion of SPAD recharge phase. The release of the trapped carriers while the avalanche triggering probability is very low indicates low afterpulsing probability, as will be shown in section 4.2.

Fig. 3-18(b) shows the measured average output pulse widths of the CS-SPAD pixel as a function of temperature. The variation in average pulse width is due to afterpulses occurring during the cathode recharge period, but before the output falling edge has been triggered. Avalanches that occur before *vD* crosses *TH2* do not trigger output pulses, but are instead responsible for prolonging the output pulse width, since they restart the recharge process. Detailed DCR and afterpulsing measurements for the CS-SPAD circuit are presented in Section III.

### Time-Gated SPAD (TG-SPAD) Pixel

The SPAD pixels described in the previous sections operated in the free-running mode where the bias applied to the SPAD is kept constant. Once the SPAD is quenched following an avalanche, it immediately begins the passive recharge phase which lasts until the excess bias is fully restored to its original value. During this time, there is a chance that another avalanche can occur. The most probable case of avalanche re-triggering is due to the release of charges that were trapped during a previous avalanche. When the de-trapping time time-constants are comparable to, or shorter than, the recharge time constant, then most of the trapped carriers are released during the excess bias recovery time, leading to variations of the digital output pulse width, as was shown for the CS-SPAD pixel. It was also shown for the SF-SPAD pixel that the reduction of the recharge time constant and the SPAD’s capacitance can greatly reduce the variation in output pulse width due to afterpulsing. However, afterpulsing effects are still prominent at low temperatures since the most of the trapped charges are released after the short dead-time.

A better way to prevent the afterpulses from occurring is to prevent the SPAD from recharging immediately after an avalanche by keeping it biased below breakdown. Then the excess bias is quickly restored once all the trapped charges from the previous avalanche have been released. However, this requires the use of AQR circuits to quickly restore the excess bias after a fixed and well controlled period of time [49]-[56]. The inclusion of the extra circuitry in the pixel reduces the fill-factor and requires precise timing circuit to control both quenching and the rest. To solve this problem, PQAR was adopted whereby a pFET biased in the off-state provides the quenching resistance and also provides the active recharge when biased in the on-state. The pixel area can thus be reduced by re-using the same circuit element for both quenching and reset operations.

The time-gating (TG) SPAD used in this work to reduce afterpulsing effects is illustrated in Fig. 3-19(a). The in-pixel time-gating circuitry turns on the SPAD only within a very narrow gate window and only the photons that arrive within this time are detected. As a result, the power consumption of TG SPAD pixels may be much lower than for free-running pixels. However, since the SPAD is turned on for only a brief period of time, the PDE will be lower. The TG SPAD pixel requires only 5 transistors. All required timing circuits for external control and signal read-out are generated on-chip. The three generated control signals, Recharge (*VR*), Quench (*VQ*) and Gate (*VG*) are synchronized to a single external clock signal, which is the only external signal needed to operate in TG mode.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑19: (a) Schematic and layout of TG SPAD pixel. (b) Simulated and measured waveforms. In the first time gate, a photon is detected, resulting in an output pulse being produced. In the second time gate, no photon is detected and therefore no output pulse is produced.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 3‑38: (a) Schematic and layout of TG SPAD pixel. (b) Simulated and measured waveforms.

The simulation results of Fig. 3-19(b) show that upon arrival of the external trigger at *t* = 0, a short pulse *VR* charges the cathode *vC* up to *VDD\_SPAD*. The on-chip delay generator provides a fixed delay of ~800 ps between the falling edges of *VQ* and *VG*, (same as pulse width of *VR*) during which time the cathode is charged through the pMOS transistor. If an avalanche occurs during *TON*, the cathode will discharge, pulling up the drain voltage *vD* and a driver circuit produces output pulses *vOUT* with fixed duration. The rising edge of *vOUT* precisely marks the photon arrival time within the time-gate. The very large resistance of the pFET when *VR* = *VSPAD* ensures very fast quenching which limits the avalanche current. It also prevents the SPAD from recharging during the gate-on time following an avalanche, which, leads to lower afterpulsing probability. If no avalanches occur within the gate window, then no output pulses are produced and the cathode is fully discharged when the quenching signal *VQ* is applied to the nFET to remove the excess bias voltage at the end of the gating interval.

The hold-off time is determined by the clock frequency, where *THO* = [*TG* – (*TON* + *TRCH*)], where *TG* is the gating period, *TON* is the gate-on time, and *TRCH* is the recharge time which can be neglected when approximating the hold of time since the recharge time is typically less than one nanosecond. The hold-off time is imposed to allow the trapped charge population to sufficiently decay so that the possibility of afterpulsing is reduced. The afterpulsing performance of the time-gating circuit is presented in Chapter 4.

# Dark Count Rate and Afterpulsing Performance of Free-running and Time-gated SPADs

Compared to other single-photon detectors (PMTs and iCCDs), SPADs have higher DCR per unit of detector area (DCR/μm2) resulting from semiconductor impurities and defects localized in the avalanche multiplication region. As a result, CMOS SPADs require very small photosensitive areas in order to achieve acceptable DCR levels because the defect density distribution is typically non-uniform [156],[157]. However, smaller active areas result in limited pixel fill-factor, especially considering the inclusion of SPAD guard-rings and in-pixel circuitry such as AQR for TGSPC, and TDC for TCSPC [43]-[48]. On the other hand, large active areas are useful for improving the collection efficiency in fluorescence decay experiments as well as easing alignment tolerances for fiber-optic coupling [88],[160]-[162]. Therefore, SPADs with active areas of 81 and 100 µm2 were chosen in this work to study the performance of larger SPADs and to achieve a higher fill-factor. However, the larger SPAD active area implies higher DCR and afterpulsing compared to the miniaturized ones. Although it is possible that the average DCR can be subtracted when measuring photon counts, the statistical variation in the DCR cannot. This quantity has to be measured accurately in order to assess the quality of the fabricated SPAD.

In this section, the DCR and afterpulsing performance of the fabricated SPAD pixel structures are measured and characterized. The DCR will be shown to depend on fabrication process and SPAD structure, in addition to front-end circuit design, the dead-time, excess voltage, and temperature. The effect of afterpulsing on free-running and time gated SPADs will be analyzed and compared in detail in section 4.2.

## Dark Noise: Dark Count Rate (DCR)

### Dark Noise Mechanisms

There are three main sources of dark counts in SPADs [155]: 1) thermal generation in the depletion region due to bulk traps; 2) thermal generation at the SPAD surface due to the interface states; and 3) tunneling. These dark counts are uncorrelated with respect to each other (i.e. the probability of dark count occurrence in a given time interval does not depend on occurrences in previous intervals) and are considered as the primary dark counts. The secondary dark counts, also known as afterpulses, on the other hand, are correlated in time and can strongly enhance the total measured DCR. The primary dark counts mechanisms are shown in Fig. 4-1. Afterpulses and are treated in Section 4.2.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑1: (a) Illustration of DCR mechanisms [90],[155]. (b) DCR as a function of temperature for a commercially available SPAD [260].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑2: (a) Illustration of DCR mechanisms [90],[155]. (b) DCR as a function of temperature for a commercially available SPAD [260].

Considering the carrier generation mechanisms in n+/p-well junction in reverse bias, firstly there is the thermally generated minority carriers which diffuse from the neutral regions into the avalanche multiplication region and induce the avalanche process (case 1 in Fig. 4-1(a)). When the SPAD is illuminated with red and near infra-red (NIR) light, carriers are photogenerated in the undepleted regions and diffuse to the multiplication region to trigger avalanches due to a higher penetration depth in silicon for longer wavelengths [175]. However, this carrier generation mechanism is negligible in the DCR, since carrier generation in the depletion region is the dominant process.

The GR processes (case 2 in Fig. 4-1(a)) described by SRH theory involve the capture and release of electrons and holes by trapping energy levels *Et* located in the forbidden band [115]. The trapping levels that trap and release carriers are due to lattice defects and impurities in silicon. These defects and impurities are produced during junction and silicide formation, which are main causes for pn junction reverse leakage in DSM processes [244]-[250]. According to SRH statistics, under reverse bias, the *pn* product deviates from its equilibrium value, *ni*2, so both *n* and *p* are very low, (*n* << *ni* and *p* << *ni*) in the depletion region. The carrier generation rate is given by

|  |  |
| --- | --- |
| , | (4-1) |

where *ni* is the intrinsic carrier concentration, *Et* and *EF* are the trap energy level and Fermi energy levels, respectively, *k* is Boltzmann’s constant, and *T* absolute is temperature. *τn* and *τp* are the electron and hole lifetimes, and Γ*n* and Γ*p* are the field enhancement factors for electrons and holes, respectively [115],[179].

In very high electric fields, the rate of traps capturing electrons is increased by the tunneling of the electrons from the valence band into the trap energy level and the emission rate is also increased by the tunneling of the electron from the trap level into the conduction band (cases 3 and 4 in Fig. 4-1(a), respectively). The field-enhancement factors in eq. (4-1) are to account for the trap-assisted tunneling reduced lifetimes of electrons and holes [155]. Electrons can also tunnel directly from the valence band to the conduction band (case 5 in Fig. 4-1(a)), and this mechanism dominates the reverse characteristic of heavily doped junctions [261].

The dominant source of dark counts in SPADs can be identified by studying the temperature dependence of the DCR. For SRH dominated dark counts, an order of magnitude decrease in dark counts for every 27 °C drop in temperature is expected [149]. When SPADs are cooled to suppress thermal generation, the tunneling-related dark counts will dominate at low temperatures since the tunneling mechanisms are less temperature dependent than the thermal generation mechanisms [165],[168],[169],[171],[172].

Fig. 4-1(b) shows the typical dependence of DCR on temperature for a 50 μm commercial SPAD, where the relative contributions of tunneling and SHR generation to the DCR have been identified [141],[260]. At around -10 °C, the contributions of the thermal generation and tunneling mechanisms become similar, and further cooling is no longer beneficial in reducing the DCR. While lowering the temperature helps eliminate the dark counts, at the same it increases the afterpulsing probability for SPADs that have short dead-time. To reduce the tunneling and afterpulsing contributions, proper doping profiles with very low levels of contaminations are required [168].

Customized processing used in CIS technologies allow for selective tailoring the doping profile [187],[230],[231]. However, these are not available in a standard low-cost CMOS technology intended for RF, analog and mixed signal applications [224]-[229]. Additionally, whereas custom silicon technologies feature efficient gettering processes to minimize the concentration of GR centers that are responsible for the primary DCR [152], the gettering operation in standard CMOS is much less efficient in the removal of defects and impurities in the SPAD active areas [156],[157]. As a result, SPAD structures in standard CMOS are limited to small active areas to maintain the DCR at reasonable levels, since larger active areas are more likely to have defects responsible for SRH generation [171].

The total DCR can be obtained by integrating over the depletion region boundary the product of eq. (4-1) and the avalanche triggering probability. The triggering probability increases with the excess bias voltage and depends on the ionization coefficients for the electrons and holes. It is approximated here by

|  |  |
| --- | --- |
| , | (4-2) |

where η is an empirical parameter that sets the exponential slope [258]. Hence, DCR increases with *VEX* not only because of the field-assisted enhancement of the emission rate from GR centers and the trap-assisted tunneling represented by Γ, but also because of the increase of the avalanche triggering probability, although at a much slower rate.

Overall, the total DCR depends on the temperature, the electric field profile in the avalanche multiplication region, the trap concentration, and on the energy levels of the different traps in the forbidden band. The closer the trap level is to the intrinsic Fermi energy level *EF*, the higher the probability of GR between electrons from the conduction band and holes from the valence band [115],[182].

### Characterization Methods

Pulse Counting Distribution

The primary dark counts of an SPAD are distributed according to Poisson statistics and possess the following attributes:

1. the number counts from non-overlapping time intervals are mutually independent (i.e. memory-less) and
2. for sufficiently small time intervals, the probability of a count is proportional to the duration of the time interval, and the probability of more than one count in this interval is negligible [262].

For a Poisson process, the probability of having *n* counts occurring within an interval of Δ*t* is given by its probability mass function (PMF)

|  |  |
| --- | --- |
| , | (4-3) |

where λΔ*t* is the average number of counts in Δ*t* and λ is the average number of counts per second.

The number of SPAD dark counts in a fixed time interval was measured by a LeCroy Waverunner 625Zi high-speed sampling oscilloscope. Fig. 4-2(a) shows a typical output voltage waveform of a CS-SPAD pixel. The oscilloscope was set to measure a Δ*t* = 500 ms interval and the sample rate was 1 GS/s. A fixed threshold was set at 1 V in order to reject the electronic noise centered on 0 V. The noise is due to the avalanches of neighboring SPAD pixels which couple through the substrate and power supply capacitance into the output driver circuits. Since the amplitude of the noise pulses was much smaller than the SPAD pulses, only the SPAD pulses above the threshold were counted. 20,000 acquisitions were taken by the oscilloscope and corresponding histogram of the number of counts during Δ*t* is shown in Fig. 4-2(b). Since λΔ*t* >> 1, the Poisson distribution can be accurately approximated by a Gaussian distribution [262]. As the number of acquisitions increases, the histogram becomes a more accurate estimation of the Poisson PMF. The rate parameter λ can be estimated by fitting the data to eq. (4-3).

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑2: (a) Output pulses measured by the oscilloscope for an SPAD pixel with SF front-end (b) Corresponding histogram of pulse counts in a 5 ms interval.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑4: (a) Output pulses measured by the oscilloscope for an SPAD pixel with SF front-end (b) Corresponding histogram of pulse counts in a 5 ms interval.

Inter-Arrival Time (IAT) Distribution

An alternative description of a Poisson process states that the first-order inter-arrival times (IAT) are independent identically distributed (IID) random variables [262]. This analysis provides the necessary temporal information to distinguish between primary dark counts and correlated afterpulses in SPADs by calculating the deviation of the measured IAT distribution from the ideal Poisson distribution [263],[264]. In the case of zero afterpulsing, the IAT probability density function (PDF) of primary dark counts is a single-exponential decay

|  |  |
| --- | --- |
| , | (4-4) |

with rate constant λ representing the mean arrival rate of dark pulses. The mean IAT and variance of the IATs are given by eqns. (4-5) and (4-6), respectively.

|  |  |
| --- | --- |
|  | (4-5) |
|  | (4-6) |

The ratio of the standard deviation to the mean represents the coefficient of variation (CV) of the statistical data,

|  |  |
| --- | --- |
| . | (4-7) |

The CV characterizes the variability in the IATs and distinguishing the IATs as a Poisson process when CV = 1 [265]. Systematic errors of the IAT measurement such as afterpulsing and counting losses due to the SPAD’s dead-time alter the statistical distribution of the dark-counts. A CV > 1 indicates that the IATs do not follow a pure exponential decay due to the presence of correlated afterpulses in the dark counts. This is because afterpulses are most likely to occur within a very short time after a primary pulse, causing a reduction of the average IAT and a corresponding increase of the variance. A CV < 1 indicates a non-Poissonian distribution, whereby dark counts occurring with very short delay times are suppressed due to the recovery time of the SPAD where *VEX* and thus the triggering probability recovers to its initial value following an avalanche. Counting losses result in distortion of the IATs characterized by an increased average IAT and less variance in the IAT statistics. These two limiting cases dictate the behavior of the SPAD and are a direct consequence of non-idealities arising from the dead-time and the afterpulsing phenomenon. The effects of these imperfections on the DCR of SPADs can be accurately accounted for by analyzing the IAT distributions.

To achieve accurate results for the DCR characterization, high resolution and wide dynamic range are required simultaneously for IAT measurements. Fig. 4-3(a) shows the cathode voltage waveform obtained by directly measuring the output pulses of unbuffered SPAD test structure. The IATs of the dark counts were obtained by calculating the elapsed time between each pair of successive pulses. The exponential distribution of IATs stipulates that short delays between successive dark counts are more likely than long delays. Therefore, it was necessary to use a sufficiently long enough measurement interval to collect enough counts for the tail of the exponential distribution so that accurate curve fitting could be performed. Further, the sample rate had to be high enough to accurately measure the arrival times of individual SPAD pulses. Samples rate of 1-5 GS/s and acquisition intervals between 0.2-5 ms could be achieved with the high-speed oscilloscope with short acquisition times. Approximately 50k-500k IATs were taken at each temperature and bias point in order to obtain accurate IAT statistics. The IAT histograms were accumulated by the built-in statistics functions of the oscilloscope. The histogram data was exported into MATLAB where curve smoothing and curve fitting operations were performed.

|  |  |  |
| --- | --- | --- |
|  | |  |
| (a) | (b) | |

Figure 4‑3: (a) Calculation of avalanche inter-arrival times for unbuffered SPAD pixel. (b) Resulting IAT histogram showing raw histogram data, data after smoothing and resulting exponential fit.

|  |  |  |
| --- | --- | --- |
|  | |  |
| (a) | (b) | |

Figure 4‑6: (a) Calculation of avalanche inter-arrival times for unbuffered SPAD pixel. (b) Resulting IAT histogram showing raw histogram data, data after smoothing and resulting exponential fit.

A typical IAT histogram is shown in Fig. 4-3(b) for an unbuffered SPAD test structure at room temperature. Here, the statistical distribution of the measured IATs has the characteristic exponential behavior of independent Poissonian events that is observed over three decades.

A CV < 1 indicates that the dark counts are more affected by counting losses by the slow recovery of *VEX* rather than by afterpulsing effects. To further illustrate this point, Fig. 4-4(a) shows the same IAT distribution on a log-log plot. In this plot, the portions of the IATs at shorter time scales that deviate from the exponential decay are apparent. The dead-time, *TDT*, was extracted from the data as the first non-zero data point in the IAT histogram. For the unbuffered SPAD, *TDT* was 2 μs. In Fig. 4-4(b), an IAT distribution is shown for a CS-SPAD pixel at room temperature. A dead-time of 60 ns and a CV > 1 was obtained from the IAT distribution, indicating the occurrence of afterpulses in the dark counts.

The afterpulsing effects are evident in the IAT histogram by the increase of IAT probability at shorter IATs. The primary DCR was obtained by fitting the IAT decay to a single exponential function, regardless of counting losses or afterpulses, The reciprocal of the IAT decay time constant yielded the primary DCR used for characterization at different temperatures and voltages.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑4: IAT distributions at *VEX* = 1.25 V displayed on a log-log scale for (a) unbuffered SPAD pixel and (b) CS-SPAD pixel.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑8: IAT distributions at *VEX* = 1.25 V displayed on a log-log scale for (a) unbuffered SPAD pixel and (b) SPAD-CS pixel.

### Experimental DCR results for free-running and time-gated pixels

The primary dark counts of the SPAD pixels fabricated in this work were characterized as a function of excess voltage and temperature. This allowed the determination of relative performance among pixel designs, and to find optimal operating conditions for SPADs pixels fabricated in standard CMOS. It was shown in Section 4.1.1 that as the excess bias is increased, the probability of avalanche breakdown and tunneling increases, resulting in higher DCR. However, it is often desirable to operate with higher excess voltage, since the PDE and timing resolution both increase with *VEX* (as will be shown in Section 4.4).

This section shows the experimental results of the DCR characterization of the designed SPAD pixels in terms of *VEX* and temperature. The statistical models given in the previous section were used to fit theoretical curves to the measured data from which the dark count statistics were extracted. Excess voltage and operating temperatures were selected that would give optimum SPAD performance in terms of the sensitivity, noise and timing resolution for utilization in fluorescence lifetime measurements presented in the next chapter.

Unbuffered SPAD pixels

The DCR of the unbuffered silicided SPAD pixels with 81 μm2 active area was measured at room temperature for excess voltages ranging between 0.25 to 2.5 V. The distributions that were obtained by measuring the number of dark counts in a 0.5 ms interval are shown in Fig. 4-5(a). The dark average count values (plotted in Fig. 4-5(a) inset with standard deviation error bars) were found at each excess voltage by fitting the histogram data to a Poisson model (eq. 4-3). The distribution of dark counts has contributions from both the thermal and tunneling dark counts as well as from the afterpulses. However, because the timing information between dark counts is lost, the non-Poissonian dark counts (afterpulses) cannot be distinguished from the Poissionian ones (thermal generation and tunneling), so the primary DCR and the afterpulses cannot be distinguished from each other based on the pulse counting data.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑5: (a) Measured distribution of dark counts for unbuffered SPAD during a 0.5 ms interval. Inset: Average count values (with standard deviation error bars) as a function of excess voltage. (b) Distribution of IATs for the same SPAD. Inset: Calculated CV values as a function of excess voltage.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑10: (a) Measured distribution of dark counts for unbuffered SPAD during a 0.5 ms interval. Inset: Average count values (with standard deviation error bars) as a function of excess voltage. (b) Distribution of IATs for the same SPAD. Inset: Calculated CV values as a function of excess voltage.

To extract the primary DCR information, the IATs of the individual dark pulses were measured and then fitted with an exponential function. The results for the unbuffered SPAD are shown in Fig. 4-5(b) on a log-log scale. For small delays, the IAT distribution deviates from the ideal exponential behavior because of the reduced avalanche triggering probability immediately following the previous avalanche. As a result of the long excess voltage recovery time, dark carriers that are generated within this time interval have a reduced chance of triggering an avalanche, resulting in the measured DCR underestimating the true DCR. The inset of Fig. 4-5(b) illustrates this trend with a plot of the measured CV as a function of excess bias, showing a decreasing CV with increase of excess voltage. The elevated excess bias increases the electric field as well as the depletion width, both of which increase the DCR. This reduces the average IAT of dark counts, resulting in a larger percentage of generated carriers failing to trigger an avalanche since they occur while the SPAD is still in the recovery state.

The dead-time counting losses can be corrected by using a non-paralyzable dead-time model (eq. 4-8), where the SPAD is completely insensitive during the deadtime *TDT* and the arrival of new pulses does not restart the dead time [138]. For the non-paralyzable dead-time model, if *m* is the measured count rate, then the corrected count rate *n* is expressed as

|  |  |
| --- | --- |
| , | (4-8) |

where *TDT* is the time at which first non-zero value occurs in the IAT histogram [259].

The measured and corrected DCRs obtained from the mean values of the pulse count histogram are shown in Fig. 4-6(a). The difference in the measured DCR from the true DCR increases at higher excess voltages as counting losses start becoming more apparent because of the electric field enhancement of the DCR. Also shown is the DCR obtained from exponential fitting of the IAT data. The exponential-fit DCR agrees very well with the DCR obtained from eq. (4-8), suggesting that the extraction of the DCR from IAT data is effective in correcting the dead-time counting losses.

For SPADs fabricated in a CMOS process, fluctuations of the SPAD quality over the wafer can occur. Thus, the DCR statistics are of paramount importance in assessing the quality of the fabrication process [165],[266]-[268]. DCR characterization has been performed for several different SPAD pixels from the same fabrication run to assess the variation in performance. A set of chips from the lot were tested and the mean values obtained were analyzed and compared. The measured DCR values at room temperature from 8 different chips of the silicided, unbuffered SPAD test structures are shown in Fig. 4-6(a). As the DCR of SPADs is extremely dependent on the cleanliness of the fabrication, such a disparity between the pixels is not surprising [266]-[268]. The saturation effects are most apparent for the defective SPADs (5 and 8) because these have very high DCR. As the DCR increases, avalanche breakdown is more likely to occur during the long passive recharge from a previous avalanche, resulting in saturation of the count rate. It is suspected that these noisy pixels have a defect near the active region. These defective pixels were also identified by analyzing the avalanche pulse amplitude and pulsewidth distributions in Section 3.3.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑6: (a) Measured and corrected DCR as a function of excess voltage extracted from dark count histograms (0.5 ms time interval) and from exponential fitting of IAT histograms. (b) Measured DCR of eight different chips for the unbuffered SPAD test structure.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑12: (a) Measured and corrected DCR as a function of excess voltage extracted from dark count histograms (0.5 ms time interval) and from exponential fitting of IAT histograms. (b) Measured DCR of eight different chips for the unbuffered SPAD test structure.

CS-SPAD pixels

SPADs with a shorter dead-time are required for DCR characterization to minimize the count rate saturation effects. However, shorter dead-time leads to an increase of the afterpulsing probability, which must be taken into account in the measurements. For the CS-SPAD pixel, afterpulsing is expected to become more prominent because the dead-time is two orders of magnitude shorter compared to the unbuffered SPAD (see Fig. 4-4 for comparison of the two dead-times).

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑7: (a) Measured distribution of dark counts for CS-SPAD pixel during a 0.5 ms interval. Inset: Average count values (with standard deviation error bars) as a function of excess voltage. (b) Distribution of IATs for the SPAD. Inset: Calculated CV values as a function of excess voltage.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑14: (a) Measured distribution of dark counts for CS-SPAD pixel during a 0.5 ms interval. Inset: Average count values (with standard deviation error bars) as a function of excess voltage. (b) Distribution of IATs for the SPAD. Inset: Calculated CV values as a function of excess voltage.

Fig. 4-7(a) shows the measured distribution of pulse counts of a CS-SPAD pixel in a 5 ms interval for *VEX* between 0.85 and 1.2 V; the inset shows the average counts as a function of *VEX*. Again it is not clear from this data what percentage of the dark counts are afterpulses. Fig. 4-7(b) shows the corresponding IAT distribution of a SPAD-CS pixel with CV in the inset. The apparent peaks in the distribution at are indicative of afterpulsing. Afterpulses effects become more apparent as excess voltage increases, hence CV > 1. The evaluation of the DCR in the presence of afterpulsing was done by fitting the IAT histogram data to an exponential function representing the Poissonian contribution from thermal generation and tunneling, with the reciprocal of the time constant representing the primary DCR component.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑8: (a) Measured total DCR of SPAD-CS chips as a function of *VEX*. (b) Corresponding CV as a function of *VEX*.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑16: (a) Measured total DCR of SPAD-CS chips as a function of *VEX*. (b) Corresponding CV as a function of *VEX*.

The primary DCR of 11 CS-SPAD pixels are plotted in Fig. 4-8 showing the degree of variability of DCR and CV amongst the different pixels. The increase of DCR with the rise of excess voltage is mainly due to the trap-assisted tunneling and band-to-band tunneling, while the increase in CV with excess voltage is mainly due to afterpulsing. The best pixel from the lot (SPAD1) achieved a maximum total DCR of approximately 50 kHz (617 Hz/µm2), while the worst pixel (SPAD10) had a maximum total DCR of approximately 300 kHz (3.7 kHz/µm2) at *VEX* = 1.2 V. Although the measured statistics has a limited accuracy due to the small number of measured samples, the quality of SPAD pixels can be assessed by the different ranges DCR and CV values. The pixels with DCR between 5-10 kHz were of the best quality out of this lot, with a corresponding CV ≈ 1 over the excess voltage range indicative of negligible afterpulsing effects. Pixels with prohibitively high DCR falling in the range of 50-500 kHz had CV > 1 where the afterpulsing contributes significantly to the DCR.

Since the SPADs are designed for photon-starved applications, only the pixels with lowest noise are to be considered. Cooling further reduces the thermal DCR contribution and enables the SPAD to be applied effectively for photon starved applications. However, the afterpulsing phenomenon does not favor low temperature operation because the detrapping time of the carriers increases exponentially with a reduction in temperature. Also, the tunneling effects are rather insensitive to temperature, so the reduction of DCR with deep cooling is not as effective beyond a certain low temperature. The choice of operating point is chiefly dictated by the combined temperature dependences of thermal generation, afterpulsing probability and tunneling. A good cut-off point is the lowest temperature at which the DCR is no longer decaying as rapidly as it does when thermal generation dominates (as shown in Fig. 4-1(b)). Below the cutoff temperature, the DCR will not change drastically, since tunneling becomes the dominant DCR mechanism. In order to determine the cut-off point, the DCR was measured as a function of temperature. The SPAD temperature was varied between –30 to +40 °C in a temperature-controlled chamber at different bias voltages and the data was collected for post-processing. The breakdown voltage variation was taken into account by adjusting the SPAD bias (*VDD\_SPAD* – *VHV*) accordingly at each temperature.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑9: Temperature measurements of DCR for SPAD-SF pixel. (a) DCR as a function of temperature for *VEX* = 0.4 V – 1.6 V. The solid lines represent the total measured DCR while the dashed lines represent the primary DCR component obtained from exponential fitting of the IAT histograms. (b) Corresponding Arrhenius plot and extracted activation energies as a function of excess voltage.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑18: Temperature measurements of DCR for SPAD-SF pixel. (a) DCR as a function of temperature for *VEX* = 0.4 V – 1.6 V. The solid lines represent the total measured DCR while the dashed lines represent the primary DCR component obtained from exponential fitting of the IAT histograms. (b) Corresponding Arrhenius plot and extracted activation energies as a function of excess voltage.

The temperature characterization results of a silicided SF-SPAD pixel are shown in Fig. 4-9. Fig. 4-9(a) shows the total measured DCR versus temperature at different *VEX*. The DCR exhibits a weaker temperature dependence when the SPAD is cooled, suggesting that the DCR has a non-negligible tunneling contribution. This was expected due to relatively higher doping levels of the n+ and p-well layers in this CMOS technology. The increasing afterpulsing contribution is responsible for the deviation between the primary and measured DCRs at lower temperatures and higher excess voltages.

Arrhenius plots, shown in Fig. 4-9(b) for different excess voltages, were used assess the activation energies *EA* [267]. The DCR was expressed in terms of the Arrhenius equation as

|  |  |
| --- | --- |
| , | (4-9) |

where *k*B is Boltzmann’s constant, and *T* is the absolute temperature in Kelvin [269]. The temperature dependence of DCR, as revealed by the slopes of Arrhenius plots, provides useful insight into the mechanisms for the dark current and DCR [261],[267],[270]-[272]. The magnitude of the thermal activation energy is of interest as an indication of the primary defect type leading to the measured DCR as well as for making quantitative assessments of the change of DCR with temperature.

If *EA* is similar to the band gap energy, then diffusion is the major factor in the dark leakage current. If the activation energy is near the mid-gap, the dark counts take place by SRH GR mechanisms. When *EA* is smaller than mid-gap, then the dark counts are described by the field-assisted generation mechanisms such as Frenkel-Poole barrier lowering and/or tunneling-generation [261],[273]. The relative contribution of each mechanism will depend on the temperature region of operation.

In Fig. 4-9(b), the dashed lines represent the slopes of the tangents of the DCR curves in the temperature ranges between 0 – 20 and -30 – 0 °C. The slopes cross over at around 0 °C, representing the point at the tunneling becomes the main mechanism. The extracted activation energies are plotted as a function of excess voltage on the right of Fig. 4-9(b). Since activation energy *EA1* is closer to mid-gap than *EA2*, the GR mechanism is more dominant in the 0 – 20 °C temperature range, and it is relatively independent of excess voltage as expected. On the other hand, the lower *EA2* corresponds to the more dominant tunneling mechanism in the -30 – 0 °C temperature range and has an activation energy that varies linearly with excess bias.

The variation of activation energies between different SPAD pixels were also examined for the CS-SPAD pixels. In Fig. 4-10(a), best (SPAD1) and worst (SPAD2) cases of DCR performance are plotted as a function of temperature along with mean and standard deviation (as error bars) of the DCR for seven measured pixels. Primary DCR components were extracted and are shown as dashed lines in the figure. An Arrhenius-type relationship was fitted to the data and the plot is shown in the left of Fig. 4-9(b).

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑10: Temperature measurements of DCR for SPAD-CS pixel. (a) DCR as a function of temperature at *VEX* = 1.3 V for seven measured chips. SPAD1 and SPAD2 represent best-case and worse-case DCR measurements. Average DCR of seven pixels and the standard deviation error bars are also shown. (b) Corresponding Arrhenius plot and extracted activation energies.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑20: Temperature measurements of DCR for SPAD-CS pixel. (a) DCR as a function of temperature at VEX = 1.3 V for seven measured chips. SPAD1 and SPAD2 represent best-case and worse-case DCR measurements. Average DCR of seven pixels and the standard deviation error bars are also shown. (b) Corresponding Arrhenius plot and extracted activation energies.

From the Arrhenius plots, the slopes give activation energies of 0.26 and 0.2 eV between 0 °C and 30 °C for SPAD1 and SPAD2, respectively. At lower temperatures, there is a leveling off in the DCR, which is consistent with increased tunneling contribution. The field-assisted mechanisms of the SPAD were also confirmed by the observed exponential dependence of DCR on excess voltage in Fig. 4-8(a). Also shown in Fig. 4-10(b) (right) are the extracted activation energies of the seven measured SPAD. The activation energies indicate that there are two shallow trapping levels. Additional trapping levels may operate as generation centers, so the measured activation energy may actually represent the average of several different levels. Nevertheless, the low *EA* values obtained are consistent with the influence of the electric field on *EA* [267].

Non-silicided SPAD

A digital 130 nm CMOS technology uses the cobalt silicide process to reduce the sheet and contact resistance of the ultra-shallow (~200 nm) source drain implantations in DSM transistors [226]-[229],[244],[245]. For SPAD design, avoiding the introduction of impurities and damage in the active region by blocking the silicidation is a crucial step in achieving lower leakage current and hence lower DCR [249]-[251]. Of all the SPAD pixels that were studied in this work, the non-silicided SPADs had the lowest DCR across the entire range of *VEX*. Fig. 4-11 compares the total and primary dark DCR between non-silicided and silicided SPAD test structures (shown Fig. 3-1(b)) for *VEX* between 0.1 and 2.5 V. The DCR of the non-silicided SPADs ranged from 9 Hz/μm2 at *VEX* = 0.4 V, up to 281 Hz/μm2 at *VEX* = 1.4 V. This corresponds to a 19× and 3× improvement over the silicided SPAD at the same excess voltages. However, the high sensitivity of the DCR on excess voltage for both SPAD structures indicates that tunneling remains the dominant DCR mechanism. Therefore, the SPADs should be operated at a relatively low excess voltage to reduce tunneling DCR.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑11: Measured DCR of silicided (red) and non-silicided (blue) SPADs (a) DCR as a function of *VEX*. Afterpulses are subtracted from Primary DCR. Exponential dependence on *VEX* is indicative of tunneling effects. (b) IAT distributions for *VEX* = 0.5 V and 1.2 V. Afterpulsing effects seen as deviations between total and primary DCR for lower IATs.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑22: Measured DCR of silicided (red) and non-silicided (blue) SPADs (a) DCR as a function of *VEX*. Afterpulses are subtracted from Primary DCR. Exponential dependence on *VEX* is indicative of tunneling effects. (b) IAT distributions for *VEX* = 0.5 V and 1.2 V. Afterpulsing effects seen as deviations between total and primary DCR for lower IATs..

In Fig. 4-11(b) the IAT distributions of the two test structures are plotted at *VEX* = 0.5 V (left) and 1.2 V (right). Afterpulsing is much more prominent for the non-silicided SPADs based on the IAT deviations from the ideal exponential distribution. The silicidation formation in the n+-p-well junctions induces significant amounts of additional trapping energy levels, which has been shown to greatly affect the leakage current of n+-p junctions [244]-[250]. Small regions of silicide penetration (i.e., so called silicide spikes) are responsible for the resulting leakage current that flows across many localized defect points in the junction area [246]. The silicide penetrations may also be the cause of a stronger tunneling current increase and activation energy decrease with increasing bias voltage, resulting in very high DCR levels. The better quality of the non-silicided SPAD has allowed operation at higher *VEX* without suffering a penalty in increased DCR. This has resulted in better PDE, and higher sensitivity and wider dynamic range performance, as will be shown in sections 5.1.1 and 5.1.2.

## Afterpulsing Characteristics of Free-running and Time-gated SPADs

Afterpulsing in SPADs is caused by carriers that are trapped by previous avalanches and then released at a later time [178]-[180]. If a carrier is released after the excess voltage has been fully restored from a previous avalanche, then that released carrier may trigger an avalanche which is indistinguishable from a photon detection. Therefore, afterpulses appear as delayed, secondary pulses that are correlated to primary dark or photon initiated avalanches. The afterpulsing probability (AP) defines the probability that an afterpulse occurs due to a primary avalanche. Important parameters for AP are impurity concentration and carrier lifetime.

Afterpulsing is a very fast process, with time constants on the order of few tens to few hundreds nanoseconds [178],[179]. Therefore, AP can be expected to increase when the dead-time is reduced. Unlike thermal generation, afterpulsing cannot be reduced by cooling. With SPAD miniaturization, faster quenching and recharge times are feasible. However, the recharge must still be delayed until all the carriers trapped from previous avalanches have been released. As a result, when the SPAD’s operating temperature is lowered to minimize the thermal contribution, longer dead-times are necessary to minimize the afterpulsing effects which ultimately limits the dynamic range performance. Afterpulsing can be greatly minimized with front-end circuit designs that limit the amount of trapped charge and/or introduce a hold-off time before the excess voltage is fully restored [134]-[138],[181]. However, imposition of a longer dead-time is problematic since it limits the performance, particularly in applications of multi-photon timing, [275], photon correlation [91],[274]-[278], and time-gating [49]-[58]. In the following section sections, the afterpulsing phenomenon will be discussed and the afterpulsing characterization of free-running and time-gated SPADs will be presented.

### Afterpulsing mechanisms

Trapping centers are due to defects in the semiconductor lattice which cause an energy level within the forbidden energy band. The energy levels that are located near the middle of the band-gap (mid-gap) have similar probabilities of capturing electrons and holes. Thus they operate as efficient GR centers that are responsible for the primary DCR [155].

Traps located at intermediate energy levels between mid-gap and band edge, called deep levels, may also exist. These deep-levels act as minority carrier traps, where the probability of capturing only one carrier type is much higher. Also, the probability that the trapped carrier will be reemitted is much higher than the probability that it will recombine. During an avalanche breakdown, the deep-level traps are filled by carriers and subsequently released. The released carriers are responsible for triggering delayed, secondary avalanches. Hence, in the presence of afterpulsing, the measured average number of avalanches will exceed the expectation from Poisson statistics.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑12: (a) Schematic representation of trapping and subsequent release of an electron by a deep level. (b) The probability density of afterpulse generation in a silicon SPAD operating at room temperature. Increasing the hold-off times reduces the afterpulsing probability [153].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑24: (a) Schematic representation of trapping and subsequent release of an electron by a deep level. (b) The probability density of afterpulse generation in a silicon SPAD operating at room temperature. Increasing the hold-off times reduces the afterpulsing probability [153].

The afterpulsing process is illustrated in Fig. 4-12(a). In this case, an electron trap captures an electron from the conduction band at time *t* and then releases the carrier at some later time *t* + Δ*t*. The released carrier may re-trigger avalanche breakdown if Δ*t* is longer than the SPAD recharge time constant *τR*. On the other hand, if Δ*t* < *τR*, then the afterpulsing may be partially suppressed because carriers are released while the SPAD bias is below the nominal value. When this occurs, the released charges are less likely to cause afterpulses. Since the release time is statistical, the emission probability per unit time is defined for each trapping level by a characteristic time constant [178]-[180]. There may be many different trapping levels in the active region so the measured afterpulsing probability is characterized by a multi-exponential decay as is shown in Fig. 4-12(b) for a commercial SPAD at room temperature. The figure also demonstrates the dramatic impact hold-off time has on AP [61].

Many different types of traps that capture free carriers may be present, and the time constants that describe the emission rate of the filled traps may differ by many orders of magnitude [178]-[180]. Further, the electron traps are expected to contribute the most to afterpulsing, since holes in silicon have much lower ionization coefficient than electrons. Because electrons and holes are only generated inside the high field region, only electron and hole traps located inside the high-field region play a role in afterpulsing. AP is related to trap parameters such as capture cross section and trap concentration [155]. The probability that one random electron generated during an avalanche will be captured by a trap and will initiate an afterpulse when it is released is approximately

|  |  |
| --- | --- |
| . | (4-10) |

where *Nt* is the electron trap concentration, *σt* is the electron trap cross section, *We* is the effective depletion width and *C* is the SPAD’s capacitance [179].

The presence of contamination and damage during the processing is the most critical factor that influences *Nt* and *σt*. However, these are determined by foundry and not by the SPAD design. *PAP* mostly depends on excess voltage because *We* depends on *VEX*. A straight-forward way to minimize the afterpulsing probability is thus the reduction of *VEX*. However, this may not be desirable in terms of timing resolution and PDE performance, so afterpulsing effects are in practice limited by reducing the quenching time and reducing the SPAD’s capacitance. As a result, the features of the SPAD front-end circuit play a critical role in the afterpulsing performance [134]-[138],[153],[163],[181]. Temperature variations also play an important role [86],[279]. At high temperatures, the trapped carriers are released more rapidly and are less likely to cause afterpulses. This is in contrast with the temperature dependence of the primary DCR, which gets increases at higher temperatures.

The total DCR can be expressed in terms of the afterpulsing probability and the primary DCR as

|  |  |
| --- | --- |
|  | (4-11) |

where *DCRPR* is the primary DCR associated with thermal generation and tunneling, and *DCRT* is the total DCR with afterpulsing included [152]. The afterpulsing effect creates a positive feedback loop that may become self-perpetuating because each afterpulse can generate new afterpulses. Therefore, the afterpulsing probability should be reduced as much as possible to minimize the total DCR.

### Afterpulsing Characterization

Accurate and reliable afterpulsing characterization techniques are needed to obtain the AP of SPAD pixels. Methods that use autocorrelation to evaluate AP require auto-correlation computation circuits to get an accurate evaluation of afterpulsing probability [28],[278]. A more commonly used characterization methods involve measurement of a histogram of avalanche IATs using TDC circuits [180],[264],[280]-[282]. This analysis provides the necessary temporal information to distinguish between primary dark counts and correlated afterpulses. As was shown in section 4.1.2, the IAT statistics of SPAD dark pulses is represented theoretically by Poisson statistics. In the case of zero afterpulsing, the IAT PDF is a single-exponential function, with 1/λ representing the arrival rate of dark pulses. Afterpulsing causes short IATs to be more likely, since the trap lifetimes are on the order or hundreds of nanoseconds and since the detrapped carriers can immediately retrigger avalanches. This makes the IAT PDF a multi-exponential function, with a slow exponential decay representing the (uncorrelated) Poisson process, and a fast exponential decay representing the fast afterpulsing process.

The AP can be obtained by measuring the probability density of the occurrence of an afterpulse following a primary pulse as a function of time, and fitting the resulting distribution with a multi-exponential function,

|  |  |
| --- | --- |
|  | (4-12) |

where *λ0* is the primary DCR, *Ai* and *λi* are the exponential pre-factors and detrapping rates, respectively, *N* is the number of trapping levels considered and *CT* is the total number of counts measured. Fitting an exponential to the uncorrelated noise and then finding the fraction of events above the fit curve yields the afterpulsing probability. The IAT histogram was fitted by eq. (4-12), and the AP was calculated according to

|  |  |
| --- | --- |
|  | (4-13) |

where the difference in areas was taken between the multiple exponential (representing the afterpulses) and the single exponential (representing the primary dark counts), relative to the total area in eq. (4-12) to yield the fraction of dark pulses due to afterpulsing.

High time resolution and wide dynamic range are required for the IAT measurements, since the primary time constant *τ0* = 1/*λ0* is typically much longer than the afterpulsing time constants *τi* = 1/*λi*. In this work, afterpulsing probabilities were evaluated by recording the IATs using the built-in statistics functions of a 20 GS/s oscilloscope and using the histogram curve-fit method described above. Two histograms were assigned to the oscilloscope channels to record the IATs using short (from 0.2 to 1 ns) and long (from 2 to 200 ns) time-bins to extract the primary Poisson and afterpulsing portions of the DCR, respectively. The histograms each had 5,000 time-bins and at least 100,000 counts were accumulated. This characterization method allowed for fast and accurate AP evaluation.

An evaluation of the accuracy of the afterpulsing characterization method was performed by measuring the AP of a commercially available SPAD from MicroPhoton Devices (MPD PDM) [283]. This SPAD has a 50 µm active area and its temperature is controlled by an integrated Peltier cooler [141]. The SPAD is quoted for DCR < 100 Hz and AP between 1 and 3 % [283]. Fig. 4-13 shows the measured IAT distributions. Two main contributions were observed: the correlated events, representing the effective afterpulse distribution, and the uncorrelated background representing the thermal DCR. In this case, only two exponentials were required to obtain a good fit to the data. The measured results show good agreement with the quoted values, confirming the accuracy of the experimental set-up for afterpulsing characterization.

|  |
| --- |
|  |

Figure 4‑13: Measured IAT distribution and exponential fitting results for MPD SPAD.

|  |
| --- |
|  |

Figure 4‑26: Measured IAT distribution and exponential fitting results for MPD SPAD.

### Experimental results for free-running and time-gated pixels

The AP was determined from measurements performed for free running SPADs using the unbuffered and CS-SPAD test structures. Since AP is a function of the number of change carriers in each avalanche pulse, large SPAD capacitances can cause high probability of afterpulsing. Although the capacitance for the unbuffered SPAD pixel is high (~10 pF), the afterpulses are not so apparent in the IAT distributions at room temperature (Fig. 4-5(b)). This is because at these temperatures, the carrier trapping lifetimes are rather short (tens of nanoseconds), so a significant fraction of afterpulses are ‘lost’ in the long dead time (~1 µs).

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑14: (a) Measured and fitted IAT distributions for unbuffered SPAD pixel at T = -30 °C. (b) Measured afterpulsing probability and count rate as a function of excess voltage. A halogen lamp was used to provide background illumination to increase the count rate and reduce the measurement time as a result.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑28: (a) Measured and fitted IAT distributions for unbuffered SPAD pixel at T = -30 °C. (b) Measured afterpulsing probability and count rate as a function of excess voltage. A halogen lamp was used to provide background illumination to increase the count rate and reduce the measurement time as a result.

When the pulse width histogram of this SPAD at room temperature is examined (Fig. 3-10), there are several peaks following the primary peak, indicating that most of the afterpulses occur while the SPAD is recharging from a previous avalanche [255]. Cooling the SPAD to -30 °C reduces the dark counts, however the AP becomes more dominant as shown in Fig. 4-14(a) for *VEX* = 1.25 V and 2.5 V. At low temperatures, the probability that trapped carriers are released past the dead time increases and the afterpusling effects become more visible in the IAT histogram.

Fig. 4-4(b) shows a strong increase of AP and count rate as *VEX* is increased for the unbuffered SPAD pixel. Because the unbuffered SPAD is connected to the load capacitor directly, the number of carriers flowing through the SPAD during an avalanche is quite high. Therefore, the avalanche charge flowing through the SPAD is expected to be decreased dramatically when a passively quenched front-end circuit is integrated inside the pixel. However, this does not necessarily correspond to a reduction of the afterpulsing probability since the dead-time is also reduced.

Fig. 4-15 shows the fitted IAT distributions at –30 °C for two measured CS-SPAD pixels, SPAD1 and SPAD2. The histogram time origin was shifted by the period of time required to reach the first count in the histogram, which signifies a dead-time of approximately 40 ns at this temperature. At –30 °C, the background DCR of SPAD1 was small enough to identify three afterpulsing components (with time constants of 48 ns, 328 ns, and 1.26 μs) to produce a good fit. Pixels with higher background DCR (SPAD2) required only two components with time constants of 54.8 ns and 885 ns to achieve a good fit at -30 °C temperature.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑15: Measured and fitted IAT histograms of two FR pixels: (a) SPAD1 and (b) SPAD2 at –30 °C. All data are for *VEX* = 1.3 V.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑30: Measured and fitted IAT histograms of two FR pixels: a) SPAD1 and b) SPAD2 at –30 °C. All data are for *VEX* = 1.3 V.

Since the lifetimes strongly depend on the process technology as well as material quality, these parameters have to be determined for different SPADs. The fitted decay curves of the IAT distributions for best and worst performing SPADs (SPAD1 and SPAD2) are shown in Fig. 4-16(a) at different temperatures, along with the corresponding CV of the IAT data. Almost 90% of the afterpulses are encountered in the first microsecond after the avalanche event. Due to the short dead-time, the deviation of the IATs from the Poisson distribution was considerable for SPAD1. SPAD2 had much higher DCR so this deviation was less apparent. The difference in AP behavior between SPAD1 and SPAD2 within the first hundred nanoseconds was attributed to the large difference in DCR between the pixels. As the temperature was elevated for SPAD2, the afterpulsing inducing traps that have longer lifetimes became indistinguishable from the DCR background counts, and only the shortest afterpulsing decay constant was clearly identifiable. Above room temperature, the number of carriers that were released after the SPAD has been fully recharged drops considerably, so the AP becomes almost negligible (~2%).

The total AP of the two pixels are shown in Fig. 4-16(b) together with the minimum hold-off time required for minimal (1%) AP. Although the AP of SPAD2 was smaller than SPAD1 and had less dependence on temperature, the two pixels had very similar decays in the minimum hold-off time as a function of temperature. As shown in Fig. 4-16(a) for IAT distributions for free-running SPAD’s operating at room temperature, a few hundred nanoseconds hold-off can reduce AP by orders of magnitude, since it covers most of the release transient and practically eliminates afterpulsing. However, a 40 °C reduction in operating temperature to reduce the DCR requires at least an order of magnitude increase in the hold-off time to achieve AP < 1 %.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑16: (a) IAT probability distributions at different temperatures for SPAD1 (Left) and SPAD2 (Right) at *VEX* = 1.3 V. CV of histogram data is shown in the insets. (b) Calculated afterpulsing probabilities (left) calculated hold-off time required for 1% afterpulsing probability (right) for FR SPAD pixels as function of temperature at *VEX* = 1.3 V. Errors in the calculations arise from small uncertainties in temperature, breakdown voltage, and from variations in quality-of-fit to afterpulsing data.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑32: (a) IAT probability distributions at different temperatures for SPAD1 (Left) and SPAD2 (Right) at *VEX* = 1.3 V. CV of histogram data is shown in the insets. b) Calculated afterpulsing probabilities (left) calculated hold-off time required for 1% afterpulsing probability (right) for FR SPAD pixels as function of temperature at *VEX* = 1.3 V. Errors in the calculations arise from small uncertainties in temperature, breakdown voltage, and from variations in quality-of-fit to afterpulsing data.

Due to the large DCR of the free-running SPAD pixels, SPAD were designed to operate in a time-gated mode to reduce the probability of detecting a dark count down to 10-6 per gate with a gating window of ~3 ns [53]. AP characterization was performed at –30 and 27 °C with *VEX* = 1.3 and 1.5 V for varying hold-off times. The time required to observe afterpulsing effects in the dark at low temperatures with sufficient accuracy for the TG SPAD was very long due to their low DCR (~2 kHz). This was compounded by the very short gating time, which meant that the time required to obtain a sufficient number of counts was prohibitively long. The photon-induced AP was therefore measured by illuminating the TG SPAD with an attenuated halogen light source in order to increase the per gate counting probability (~10-3) and thereby considerably reduce the measurement time.

In Fig. 4-17(a), the TG SPAD’s avalanche IAT distributions at –30 °C are shown for two different hold off times. On the right, the first 25 time-gates following an avalanche are shown for a 100 MHz gating frequency (corresponding to 7 ns hold-off time). There is an absence of counts in the histogram during the hold-off periods, since the histogram time resolution is 200 ps (equivalent to 15 histogram bins per time gate). Two afterpulsing components were identified with time constants of 9.7 ns and 29 ns. On the left, the IAT distributions are shown at –30 °C for a gating frequency of 6.25 MHz (corresponding to a 157 ns hold-off time). It can be seen that this hold-off time was sufficient to eliminate afterpulsing effects. For the TG SPADs at room temperature, afterpulsing effects were eliminated by using a ~40 ns hold-off time, which was considerably lower than the minimum hold-off time required the FR SPAD pixels.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑17: (a) Measured distributions of afterpulsing in time-gated mode for illuminated SPAD at –30 °C with *VEX* = 1.3 V. 7 ns hold-off time results in 34.5 % afterpulsing probability (left). No detectable afterpulses for 157 ns hold-off time (right). The inset shows the superposition of the first 31 time-gates following an avalanche. The measured full-width at half-maximum (FWHM) of the gate width is 2.9 ns. (b) Measured afterpulsing probability versus hold-off time for TG SPAD. (Left) Fitted results at –30 °C show temporal behavior between exponential and power law. (Right) At room temperature the behavior is exponential.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 4‑34: a) Measured distributions of afterpulsing in time-gated mode for illuminated SPAD at –30 °C with *VEX* = 1.3 V. 7 ns hold-off time results in 34.5 % afterpulsing probability (left). No detectable afterpulses for 157 ns hold-off time (right). The inset shows the superposition of the first 31 time-gates following an avalanche. The measured full-width at half-maximum (FWHM) of the gate width is 2.9 ns. (b) Measured afterpulsing probability versus hold-off time for TG SPAD. (Left) Fitted results at –30 °C show temporal behavior between exponential and power law. (Right) At room temperature the behavior is exponential.

The differences in AP performance of the FR and TG SPAD pixels was attributed to the different front-end circuits. While the FR SPAD had a ~45 ns recharge time, the TG SPAD had very fast recharge (~800 ps) with a minimum hold-off time of 7 ns. Due to the shorter dead-time, the TG SPAD could detect afterpulsing decays with shorter time constants. Further, although the total avalanche charge obtained from simulations for the FR and TG SPAD pixels were very similar (~260 fC and ~290 fC at *VEX* = 1 V, respectively). Thus, the afterpulsing effects could be avoided in the time-gated mode by introducing a sufficiently long hold-off time that allows for complete de-trapping of the avalanche charge before the SPAD bias is restored. The time-gated mode was also more effective in suppressing higher-order afterpulsing effects (afterpulses of afterpulses), because the SPAD was turned off immediately after each avalanche. Therefore, long sequences of afterpulses were less likely to be encountered.

Fig. 4-17(b) shows the calculated AP of the TG SPAD at –30 °C and 27 °C using different hold-off times. As described in [281], the AP as a function of hold-off time for silicon SPADs falls somewhere between a power law and an exponential behavior. This is interpreted as broadening of the detrapping rate distribution as the temperature is lowered. When the temperature is elevated, the temporal decay of the detrapping process becomes exponential, corresponding to a narrowing of the distribution of detrapping rates. Regardless of the distributions of the traps, selecting a suitable hold-off time can practically eliminate afterpulsing effects operating the SPAD in the time-gated mode.

# CMOS SPAD Optical Characterization And Fluorescence Lifetime Measurement Results

The main benefit of using SPADs is the ability to detect extremely weak optical signals, which is critical for FLIM applications. However, the capability to detect low-level light is affected by the DCR and AP performance. Extensive SPAD characterization presented in the previous chapter has shown that the DCR and AP are extremely sensitive to the excess voltage, temperature, dead-time and fabrication quality. Therefore, these factors also affect the SPAD’s optical performance.

In Section 5.1, characterization results of the fabricated SPAD pixels in terms of dynamic range (DR), photon detection efficiency (PDE), and photon timing jitter are presented. Time-correlated single-photon counting (TCSPC) and continuous-wave (CW) and experiments were performed to assess the SPAD’s capability in detecting fast fluorescence decays in a laboratory setting. Results from these measurements are shown in Section 5.2 for the SPADs fabricated in this work. The key limitations and benefits of using SPADs fabricated in a low-cost standard digital CMOS technology for fluorescence lifetime measurements are highlighted and their performance is compared to a commercially available single-photon detector.

## Optical Characterization Results

### Dynamic Range

The ratio between the maximum and minimum detectable optical power by the SPAD defines the dynamic range. Whereas the lowest detectable light power is determined by the DCR and AP, the highest detectable light power is determined by the dead-time. The high-end of the dynamic range is very important for SPADs because the output can easily saturate when exposed to bright light (such as background light). Since the counting losses depend on the dead-time, the AP performance is important at the upper end of the dynamic range [101],[284].

The dynamic range is limited by the photon flux at which the SPAD count rate begins to deviate from a linear function [285]. Reducing the dead-time is the most straightforward way to achieve higher dynamic range, but because of the carrier trapping phenomenon in the active region, the detector dead-time cannot become too low, otherwise afterpulsing will dominate and the DCR will increase to reduce the lower end of the dynamic range. Indeed, the avalanche must be quenched and subsequently recharge only when the trapped carriers have been removed from the active region, resulting in a dead-time that can range from few tens of nanoseconds up to several microseconds, depending on the operating temperature. As a result, the maximum counting rate achievable is severely limited when low AP is required.

Measurement Set-up

In order to characterize the dynamic range performance of the SPAD, its response at different incident light intensities was measured. Fig. 5-1(a) shows the instruments that were used in the optical characterization. All optical instruments and SPADs under test were mounted on a vibration-free optical table in an isolated dark chamber. A xenon lamp generated a broadband light which was subsequently collimated by lenses in order to obtain a uniform beam. The desired wavelength was selected with optical bandpass filters (BPF) with 10 nm bandwidth.

Fig. 5-1(b) shows the measured spectrum of the xenon lamp as well the measured spectra obtained with 400, 500, and 600 nm BPF. Each BPF was verified to have 10 nm optical bandwidth as shown in Fig. 5-1(b). The light beam was attenuated through neutral density (ND) filters to vary the photon count rate at the detector. The optical intensity (power per unit area) was evaluated by means of a calibrated silicon photodiode (SiPD) placed at a known distance (50 mm) from the lamp where the beam was uniform over a diameter of several centimeters. The range of light intensities obtained by attenuation of the light beam was measured with the calibrated SiPD and the results are shown on the right of Fig. 5-1(b). The optical power meter was capable of measuring light sources from 50 nW to 50 mW at wavelengths between 400 nm to 1100 nm. The SiPD was then removed and replaced with SPADs that were to be characterized. The measurements were carried out through a custom-made MATLAB program. This program was used to control the oscilloscope and SPAD bias voltage supplies and for data collection and analysis.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑1: (a) Instrumentation for optical characterization of SPADs. (b) Right: Measured optical spectrum of xenon lamp and filtered light. Left: Measured optical powers for λ between 520 – 580 nm.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑2: a) Instrumentation used for optical characterization of SPADs. b) Right: Measured optical spectrum of xenon lamp and filtered light. Left: Measured optical powers for λ between 520 – 580 nm.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑3: a) Instrumentation for optical characterization of SPADs. b) Right: Measured optical spectrum of xenon lamp and filtered light. Left: Measured optical powers for λ between 520 – 580 nm.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑4: a) Instrumentation used for optical characterization of SPADs. b) Right: Measured optical spectrum of xenon lamp and filtered light. Left: Measured optical powers for λ between 520 – 580 nm.

The optical powers measured by the calibrated SiPD were converted to photon flux striking the SPAD active area according to

|  |  |
| --- | --- |
|  | (5-1) |

where *PSiPD* is the light power measured by the calibrated SiPD, *λ* is the wavelength, *h* is Planck’s constant, *c* is the speed of light, *ASPAD* is the active area of the SPAD, and *ASiPD* is the photosensitive area of the SiPD [184]. The measured count rate (CR) of the SPAD, as a result of the incident photon flux, *Φ'SPAD*, was obtained by fitting the measured IAT histogram using an exponential decay model as described in Chapter 3. From this analysis, the true counting rate of the SPAD, *ΦSPAD*, corresponding to the measured photon counting rate of the SPAD, was found according to

|  |  |
| --- | --- |
|  | (5-2) |

where the first term is the correction for the dead-time losses of the SPAD (section 4.1.1). The DCR was subtracted from the measurements to give the actual number of counts due to the incident photons.

To order to validate the results of the optical characterization set-up and verity its accuracy, comparative tests were carried out with a commercially available SPAD module from MicroPhoton Devices (MPD), a world-recognized and widely adopted manufacturer of standard instruments for single-photon counting measurements. The measured dynamic range of the MPD SPAD is shown in Fig. 5-2. In Fig. 5-2(a), the measured input photon flux striking the detector, *ΦIN*, as well as the measured SPAD counting rate, *Φ'SPAD*, is plotted as a function of optical attenuation level. The effects of the detector dead-time are apparent by the deviation from the linear function at the upper end of the dynamic range, while the DCR effects appear at the lower end.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑2: (a) The measured input photon flux striking the detector along with the uncorrected SPAD counting rate is plotted as a function of the optical attenuation (b) The corrected SPAD counting rate as a function of optical intensity is shown together with a linear fit. Background DCR is indicated by the dashed line.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑6: a) The measured input photon flux striking the detector along with the uncorrected SPAD counting rate is plotted as a function of the optical attenuation b) The corrected SPAD counting rate as a function of optical intensity is shown along with a linear fit.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑7: a) The measured input photon flux striking the detector along with the uncorrected SPAD counting rate is plotted as a function of the optical attenuation b) The corrected SPAD counting rate as a function of optical intensity is shown together with a linear fit. Background DCR is indicated by the dashed line.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑8: a) The measured input photon flux striking the detector along with the uncorrected SPAD counting rate is plotted as a function of the optical attenuation b) The corrected SPAD counting rate as a function of optical intensity is shown along with a linear fit.

Fig. 5-2(b) shows the same data as a function of the optical intensity corrected for the counting losses and after background DCR subtraction. *ΦSPAD* deviates from the ideal case at lower light intensities due to stray background light. This is because the subtraction of DCR in eq. (5-2) has a greater impact in the photon starved regime of the dynamic range. The measured counts showed good agreement with the linear fit over nearly six orders of magnitude in accord with the MPD SPAD specifications. These results served to validate the measurement-set up used subsequently for CMOS SPAD dynamic range characterization.

CS-SPAD Pixels

The measured count rates as a function of the optical attenuation are shown in Fig. 5-3(a) for the silicided CS-SPAD pixel at different excess voltages. The minimum light intensity from which signal counts can be distinguished above the background noise (i.e. the level where the SNR is approximately unity) is determined by the DCR level. Because CS-SPAD pixels have a much higher background DCR compared with the MPD SPAD, the response of the count rate flattens at the lower end of the dynamic range. Since the dead time of the CS-SPAD is around 30 ns, the maximum counting rate is approximately 33.3 MHz. However, the optical intensity in this set-up could not be set high enough to reach this counting rate. This is because the thin layer of silicide on top of the photosensitive area reflects a large fraction of the incident photon flux. As a result, the measured count rate was well below the incident photon flux *ΦIN*.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑3: (a) The measured input photon flux striking the detector is plotted along with the uncorrected SPAD counting rate at different excess voltages for decreasing levels of optical attenuation. (b) The corrected SPAD counting rate with background DCR subtracted is plotted as a function of optical intensity. Measurement taken at λ = 560 nm.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑10: a) The measured input photon flux striking the detector is plotted along with the uncorrected SPAD counting rate at different excess voltages for decreasing levels of optical attenuation b) The corrected SPAD counting rate with background DCR subtracted is plotted as a function of optical intensity. Measurement taken at λ = 560 nm.

Fig. 5-3(b) shows that the measured count rate linearly depends on the incident photon flux over a range of two decades indicating a dynamic range of at least 20 dB. Optical lenses can be used to focus more light on the SPAD to allow for a significant increase in the maximum count rate that can be measured.

Non-silicided SPAD test structures

The results in Chapter 4 revealed that non-silicided SPADs have significantly lower DCR compared to silicided ones. The absence of silicide - a well-known source of reverse leakage current in reverse biased pn junctions - on the photosensitive region results in a significant performance improvement in image sensor design [249]-[251]. As a result of the lower DCR, weaker light intensities could be detected with the non-silicided SPAD pixels. The measured dynamic range of such a pixel is shown in Fig. 5-4.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑4: Measured dynamic range of non-silicided SPAD. (a) Plot of the measured input photon flux striking the detector along with the uncorrected SPAD counting rate as a function of the optical attenuation (b) Plot of corrected SPAD counting rate as a function of optical intensity on the SPAD. Measurement taken at λ = 560 nm.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑12: Measured dynamic range of non-silicided SPAD test structure a) Plot of the measured input photon flux striking the detector along with the uncorrected SPAD counting rate as a function of the optical attenuation b) Plot of corrected SPAD counting rate as a function of optical intensity on the SPAD. Measurement taken at λ = 560 nm.

Compared with Fig. 5-3, the non-silicided SPAD shows better response to low-level light. This is because at the lowest excess voltage setting of *VEX* = 0.4 V, the DCR was as low as 500 Hz. This DCR is comparable to that of the MPD SPAD, and as such, the optical response at the lower end of the dynamic range was similar. In contrast, the silicided SPAD pixels were practically unresponsive to the low-intensity light. On the other hand, the non-silicided SPADs had worse performance at the higher end of dynamic range, since these devices were fabricated using an unbuffered test structure and this resulted in a long dead time of approximately 0.75 μs. Integrating the non-silicided pixel with the CS front-end circuit is expected to reduce the dead-time, thereby increasing the dynamic range of the pixel by up to 20 dB. For the silicided SPAD pixels, the light saturation level was considered to be reached when the optical intensity striking the detector increased to the point where the time interval between photon arrivals was comparable to the dead time (at approximately 1 MHz).

### Photon Detection Efficiency (PDE)

Photon detection efficiency (PDE) is defined as the probability that the detector generates a digital output pulse corresponding to an incident single photon. When biased below the breakdown voltage, QE is normally used to indicate the percentage of photons incident on a photodiode’s active area that produce an electron‐hole pairs. Although photons can be converted into carriers with high QE (which can be as high as 95% at 560 nm), none of the carriers can trigger a self-sustained avalanche so effect the PDE is considered to be zero.

When biased above the breakdown voltage, the PDE depends not only on QE, but also on the probability that an electron or hole will generate an avalanche, and this depends greatly on the excess voltage. In practice, the PDE is determined as the percentage of photons that actually trigger an avalanche compared to the total number of photons incident on the SPAD active area,

|  |  |
| --- | --- |
|  | (5-3) |

where *PAV* is the avalanche triggering probability (eq. (4-2)), QE is the quantum efficiency, and FF is the pixel fill-factor, defined as the ratio of the diode’s photosensitive area to the total area for a pixel. ΦIN and ΦSPAD are the incident photon count rate incident and photon counting rate of the SPAD, defined in eqns. (5-1) and (5-2), respectively. These were the quantities that were precisely measured by the optical set-up in Fig. 5(a).

PDE can be increased most straight-forwardly by increasing *VEX*. As excess bias increases, the internal electric field becomes stronger, leading to the increase of the avalanche trigger probability. However, the potential drawbacks include higher DCR and considerable afterpulsing effects due to the increased number of avalanche charges flowing through the avalanche junction. Therefore, the excess bias must be carefully chosen to achieve best overall performance.

Measurement set-up

The crucial point of the PDE characterization is to count only detections of incoming photons [263],[287]. The experimental procedure based on measurement of IAT statistics was used in order to only consider real photon detections and to neglect the afterpulses. The PDE measurements were carried out by selecting the desired wavelength with BPFs and evaluating photon flux (power per unit area) at the SPAD’s position by means of a calibrated SiPD. The power meter was then removed and replaced with the SPAD to be characterized. In order to verify the accuracy of the measurement set-up used for SPAD PDE evaluation, the PDE of the MPD SPAD was measured and compared to the quoted values.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑5: (a) Measured PDE of the MPD SPAD compared to the quoted value. (b) Left: Measured IAT histogram used to evaluate PDE of MPD SPAD at λ = 400 nm. Effects of afterpulsing are eliminated from the measurement by taking the primary counting rate as 1/τ2. Right: CV of measured IAT histograms used in PDE evaluation as a function of wavelength.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑14: a) Measured PDE of the MPD SPAD compared to the quoted value. b) Left: Measured IAT histogram used to evaluate PDE of MPD SPAD at λ = 400 nm. Effects of afterpulsing are eliminated from the measurement by taking the primary counting rate as 1/τ2. Right: CV of measured IAT histograms used in PDE evaluation as a function of wavelength.

Fig. 5-5(a) illustrates the measured PDE of the MPD SPAD with a very good agreement with quoted PDE values [283]. A representative IAT histogram of the MPD SPAD counts was fitted with a double exponential function as shown in Fig. 5-5(b). The accuracy of the measured PDE was thus guaranteed by only considering the primary count rate and discarding the afterpulsing component. Fig. 5-5(b) shows a plot of the calculated CV at each measured wavelength indicating the presence of afterpulsing effects in the measured IATs. Carefully accounting for the DCR and afterpulses is essential in order to quantify the true PDE of SPADs.

Another important consideration for accurate PDE evaluation involves the determination of the minimum incident photon rate to ensure that the measured SPAD’s count rate sufficiently exceeds the DCR. Ideally, the intensity of the incident light should be high enough so that the measured detector’s count rate greatly exceeds the DCR. Otherwise, the subtraction of the DCR in eq. (5-2) introduces errors when determining the actual count rate, as was the case shown in Fig. 5-3(a) for the silicided CS-SPAD pixel. On the other hand, if the optical intensity is too high, then the measured SPAD count rate enters saturation, as was the case shown in Fig. 5-4(b). Fortunately for the silicided pixels, the SPAD was free from saturation effects at the high end of the dynamic range because the silicide reflects a large portion of the incident light.

Fig. 5-6(a) illustrates the PDE overestimation for the silicided pixels, where the PDE values obtained at different levels of optical attenuation for wavelengths between 520 and 560 nm are plotted. At the highest attenuation levels, the PDE is overestimated by 100×. This occurs because the counting rate remains constant, even as the incident optical intensity decreases below the DCR. So as optical intensity is reduced and the denominator in eq. (5-3) decreases, the numerator remains constant leading to an apparent increase in the PDE. The PDE reaches its true value at higher levels of optical intensity, since the DCR is well below the measured counting rate. This is illustrated in Fig. 5-15(b) which shows the measured pulse count histograms in a 1 ms interval. With optical attenuation between 30 and 55 dB, the SPAD count rate distribution is very similar to the DCR distribution, whereas for 25 dB and below, the photon counts become apparent over the background DCR. As a result, the PDE values in this range were not overestimated when subtracting the DCR to obtain the true photon counting rate of the detector.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑6: (a) Measured PDE at different wavelengths as a function of optical attenuation. The PDE is overestimated at the higher optical attenuations. b) Measured pulse count histograms with 1 ms integration time at 580 nm.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑16: a) The measured input photon flux striking the detector along with the uncorrected SPAD counting rate is plotted as a function of the optical attenuation. b) The corrected SPAD counting rate.

Silicided SPAD

Based on the considerations above, the PDE of the silicided SPAD pixels was evaluated in the set-up using an optical attenuation less than 30 dB. The measurement results are shown in Fig. 5-7 for three different excess voltages. A histogram of CV values that were obtained at each wavelength and bias point is also shown in the inset. CV values near-unity indicate that afterpulsing effects were rather negligible for the SPAD at room temperature. The PDE curves exhibit maxima and minima at 500 and 540 nm respectively because the stack consisting of air/dielectric layers/silicon forms a Fabry-Perot resonator [188],[189]. Because the thickness of each dielectric layer can vary by as much as 20%, the optical transmission coefficient can vary strongly for different process corner cases and from chip to chip (Fig. 5-17(b) [289]. As a result, the process variations cause relative positions of the maxima and minima of the SPAD PDE characteristic to vary significantly between different chips. Post-process modifications are needed to remove the dielectric stacks and to add an anti-reflection coating (ARC) to improve the PDE [189].

|  |
| --- |
|  |
| (a) |
|  |
| (b) |

Figure 5‑7: (a) Measured PDE of silicided CS-SPAD pixel at three different excess voltages. The measurements were performed with an optical intensity that ensured accurate PDE calculations. Inset shows a histogram of CV values obtained from the IAT distributions measured at each wavelength and bias point. (b) Transmittance of light passing through the stack of dielectric layers in 130 nm IBM CMOS technology with ±20% process variation in thickness of layers [289].

|  |
| --- |
|  |
| (a) |
|  |
| (b) |

Figure 5‑18: (a) Measured PDE of silicided CS-SPAD pixel at three different excess voltages. The measurements were performed with an optical intensity that ensured accurate PDE calculations. Inset shows a histogram of CV values obtained from the IAT distributions measured at each wavelength and bias point. (b) Transmittance of light passing through the stack of dielectric layers in 130 nm IBM CMOS technology with ±20% process variation in thickness of layers [289].

The lower PDE at short wavelengths compared to the larger PDE at around 500 nm can be explained with the following reasoning. First, the transmittance of light through the silicon nitride passivation layer ranges between 5-50% in the 300-500 nm wavelength range, until it reaches nearly 100% at 580 nm [290]. Second, because the short wavelength light is mostly absorbed near the silicon surface (i.e. in the n+ layer above the multiplication region), the illumination results in pure hole injection into the SPAD’s avalanche region. In silicon, holes have a lower ionization coefficient than electrons so a hole-initiated avalanche breakdown has a lower probability than an electron-initiated one, resulting in a lower PDE at shorter wavelengths.

At longer wavelengths, electrons photogenerated deeper in the p-well are more likely (compared to holes) to drift into the avalanche region and initiate an avalanche. This is because the p-substrate and the DNW are electrically shorted, so the holes photogenerated in the DNW will either drift across the DNW/p-well junction towards the p+ contact in the p-well, or towards the p+ contact in the substrate. In either case, they do not enter the active region and trigger avalanches due to the potential barrier of the DNW/p-well. Also, the p-well is thicker than a n-well in a DSM technology, resulting in increased sensitivity of n+/p-well SPAD over a wider spectral range compared to p+/n-well structures [232],[288]. As such, the peak sensitivity of the measured SPADs occurs in the green to red wavelength range where most FLIM setups operate. However, the PDE performance is still below expectations due to the reduced optical transparency introduced by the thin silicide layer on the active region surface. Therefore, the silicidation of the n+ region was blocked, resulting in improved PDE performance.

Non-silicided SPAD

The measured PDE of the non-silicided SPAD test structure is shown in Fig. 5-8 for excess voltages between 0.4 and 1.2 V. This PDE also has the characteristic similar peaks and valleys due to multiple reflections within the IMD and passivation layers covering the active area of the photodiode. However, because of the lack of a silicide layer on the surface, the PDE is approximately 6× to 8× larger for the same values of *VEX*. The improved PDE, together with the reduced DCR of these pixels, opens up the capability of using these devices for high-sensitivity fluorescence lifetime analysis, as is shown in Section 5.2.

|  |
| --- |
|  |

Figure 5‑8: Measured PDE of non-silicided SPAD test structures pixel at five different excess voltages.

|  |
| --- |
|  |

Figure 5‑20: (a) Measured PDE of non-silicided SPAD test structures pixel at five different excess voltages.

Fig. 5-9(a) shows the comparison of PDE as a function of *VEX* for different illumination wavelengths. The PDE increases with *VEX* due to the increased avalanche triggering probability. For lower *VEX*, the rate of increase of breakdown probability is higher than that at lower values of *VEX*. This is because the triggering probability asymptotically approaches unity and eventually reaches saturation for high *VEX.* In contrast, the DCR keeps rising exponentially as *VEX* increases. Therefore, the plot of DCR versus PDE is approximately linear on a semi-log scale.

Fig. 5-9(b) shows that in order to reach the best PDE with an acceptable DCR, the optimal wavelength should be near 600 nm for the non-silicided n+/p-well SPAD structures. In contrast to the previously reported p+/n-well SPAD structures in similar DSM technologies, the SPADs in this work showed peak PDE response for blue/green light (470-500 nm). [165]-[167],[170]-[172]. The optimal excess bias lies between 0.4 V and 0.8 V for the SPADs in this work. Further increasing *VEX* has less of a benefit on the PDE, and only serves to increase the DCR, which lowers the dynamic range. Cooling and time-gating is expected to significantly reduce the DCR and AP, thereby allowing higher PDE and dynamic range values, opening up the possibility of further improvement of CMOS SPAD’s performance.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑9: (a) The measured PDE as a function of excess voltage for non-silicided SPAD pixel at three wavelengths. (b) The corresponding plot of DCR versus PDE for the same pixel.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑22: a) The measured PDE as a function of excess voltage for non-silicided SPAD pixel at three wavelengths. b) The corresponding plot of DCR versus PDE for the same pixel.

Time Gated SPAD (silicided)

The time-gating feature of SPADs was exploited in Chapter 4 to drastically reduce the effect of DCR and afterpulsing. However, by acquiring photons only in short and well-defined time slots while rejecting all the others, the PDE of the TG-SPAD is correspondingly reduced. As a result, in continuous wave (CW) operation, the effective PDE in the TG mode is lower compared to the PDE of the FR mode by a factor of the duty cycle (*TON*/*TG*, where *TG* is the gating period and *TON* is the gating on time). A fair comparison between FR and TG mode requires that the PDE inside the gate-on time is the same as the PDE of the FR SPADs, so the measured PDE of the TG SPAD should be normalized by the duty cycle.

Fig. 5-10(a) shows the measured and normalized PDE values of the TG-SPAD pixel in comparison to the FR-SPAD. The PDE of the TG SPAD is very similar to the PDE of the FR SPAD, other than the different location of peaks/valleys in the PDE response. The variation in wavelengths at which the PDE minima and maxima occur for the different measured chips is due to constructive/destructive interference of the incident light, which is affected by the variation of IMD thickness. The fabricated SPAD chips exhibited the most variation in optical sensitivity for green light, since the PDE variations were the most pronounced in this wavelength range.

A plot of the measured PDE performance of the FR and TG is shown in Fig. 5-10(b). The non-silicided SPAD showed a considerable improvement over the silicided structures. However, the PDE is still much lower than the MPD SPAD. Post-processing steps to remove the passivation and IMD layers above SPAD active area, as well as deposition of optimized antireflection coatings on the chip surface, is expected to improve the PDE of standard CMOS SPADs [189].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑10: (a) Measured PDE performance of TG-SPAD pixel. The measured PDE (green) was normalized (red) by the duty cycle and compared with free-running PDE performance (blue). Both pixels were silicided. (b) Comparison of PDE performance between silicided and non-silicided SPAD pixels as well as MPD SPAD. The y-axis is shown on a log scale.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑24: a) Measured PDE performance of TG-SPAD pixel. The measured PDE (green) was normalized (red) by the duty cycle and compared with free-running PDE performance (blue). Both pixels were silicided. b) Comparison of PDE performance between silicided and non-silicided SPAD pixels as well as MPD SPAD. The y-axis is shown on a log scale.

### Timing Resolution

The timing resolution is defined as the timing jitter between the true photon arrival time and the time instant when the output pulse is recorded. It is a critical parameter for SPADs, since in practical applications jitter limits the overall time resolution of the TCSPC system. The timing resolution is described by the full-width at half-maximum (FWHM) of the resulting statistical distribution of the delays between the photon arrival time and the time of the leading edge of the SPAD output pulse. When photons are absorbed, the photo-generated carriers may either undergo immediate avalanche multiplication, or may diffuse before undergoing avalanche multiplication. Although the statistics of these processes are different, the final result can be described as the superposition of a Gaussian response and an exponential tail, with the Gaussian component representing the variation in the avalanche build-up time and the exponential component representing the diffusive processes [87],[130],[139],[152].

Timing responses of the CS-SPAD were characterized by the TCSPC method at two different wavelengths (λ = 470 nm and 520 nm) using high repetition rate (up to 80 MHz) picosecond pulsed (FWHM = 70 ps and 110 ps) diode lasers (Picoquant P-C-470 and P-C-510). Fig. 5-11(a) shows the experimental set-up. The laser beam was aligned in free-space such that it was directly incident on the chip and thus illuminated the entire photosensitive area of the SPAD. The laser intensity was adjusted by placing ND filters in the path of the beam so that the detector operated in a photon starved mode. The laser driver (PDL 800-B) provided a low-jitter electrical signal synchronized with the laser pulse which was used as the START signal in the TCSPC set-up. The STOP signal was provided by the output pulses of the CMOS SPADs. The oscilloscope was programmed to digitize the time intervals between START and STOP pulses and to evaluate the corresponding IAT histogram. The histogram of time differences between the laser pulse and the SPAD output pulse yielded the overall timing uncertainty of the TCSPC system. Removal of the jitter components generated by the measurement setup (mostly dominated by the laser’s 70 ps jitter) yielded the SPAD’s jitter.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑11: (a) The experimental set-up used to measure the timing response of SPADs. (b) Measured timing response of MPD SPAD (left) had good agreement with the manufacturer’s specifications. The measured timing jitter of CS-SPAD pixel (right) showed excellent performance in comparison.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑26: a) The experimental set-up used to measuring the timing response of SPADs. b) Measured timing response of MPD SPAD (left) had good agreement with the manufacturer’s specifications. The measured timing jitter of CS-SPAD pixel (right) showed excellent performance in comparison.

To validate the measurement setup, the timing response of the MPD SPAD was measured with the results shown in Fig. 5-11(b), left. The measured FWHM is 86.4 ps for excitation at 470 nm. Taking the pulse width of 70 ps of the excitation laser into account, the resulting 50.6 ps timing resolution of the MPD SPAD was in good agreement with the manufacturer’s specifications. The measured FWHM results of the CS-SPAD pixels (Fig. 5-11(b), right) are comparable to the MPD SPAD, whereas the full-width at one-hundredth of maximum (FW(M/100)) value was better for CS-SPAD pixels. This could be on account of the variations of laser pulse shape when higher power settings on the laser driver were used to characterize the silicided SPADs.

The CS-SPAD pixels were optimized for high timing resolution performance. This was achieved through the design of the front-end circuit used for sensing the SPAD avalanche and generating digital output pulses. The timing jitter of the output pulses was minimized by utilizing a sensing threshold ~300 mV below the SPAD cathode supply voltage and utilizing the full logic-swing CMOS circuits in the output-driver circuit.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑12: Measured timing jitter distributions of CS-SPAD pixels for two excess voltages, *VEX* = 1 V and 1.2 V. The laser wavelengths are (a) λ = 470 nm and (b) λ = 510 nm.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑28: Measured timing jitter distributions of CS-SPAD pixels for two excess voltages, *VEX* = 1 V and 1.2 V. The laser wavelengths are a) λ = 470 nm and b) λ = 510 nm.

Fig. 5-12 presents the timing jitter measurement of four CS-SPAD pixels for the same design at two excess voltages and two different wavelengths (λ = 470 nm and λ = 510 nm). The measured histograms were fitted by a Gaussian model from which the FWHM was obtained. Each laser has its own specific fixed delay from the laser trigger to the laser output, so the mean time intervals are also shifted accordingly. Also, the timing jitter includes the laser and detector jitter. FWHM was measured for SPAD1 to be 78 ps at λ = 470 nm and 132 ps at 510 nm for 1 V of excess bias. The increased timing jitter for λ = 510 nm occurs because the laser has a FWHM of 110 ps compared to 70 ps at 470 nm. In addition, because the avalanche region of the SPAD is very shallow, photons with shorter wavelengths are absorbed closer to the surface and therefore generated avalanches with less timing uncertainty.

Compared to the SPAD structures with active region built into the deeper substrate, these devices show better time resolution. They are also free from the long exponential tail in the timing response that typically results from diffusion of minority carriers generated deep beneath the SPAD reaching the multiplication region [173]. Thanks to the high doping of the n+ layer and the depleted regions of reverse-biased p-well-/DNW and DNW/p-substrate junctions, the thickness of the neutral region beneath the SPAD was reduced, hence the effects of photo-generated carrier diffusion were avoided [160]. On the other hand, the thin avalanche multiplication region that results in very good timing performance also causes the dominance of band-to-band tunneling which leads to the high DCR (especially for the silicided junctions) that limited the optical sensitivity of the SPAD pixels.

The CS-SPAD showed the best timing performance amongst the different pixels because these were designed with an integrated front-end circuit to convert the cathode voltage discharge to digital output pulses with very low timing jitter. The non-silicided SPAD test structure utilized an external comparator circuit that contributed additional jitter to the measured response shown in Fig. 5-13(a). Fig. 5-13(b) shows that the jitter could be reduced by increasing the excess bias, but at the cost of increased DCR. In the next section, fluorescence lifetime measurements are demonstrated with both silicided and non-silicided SPADs.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑13: (a) Measured timing jitter distributions of unbuffered SPAD test structure for four excess voltages and (b) corresponding FWHM values as a function of excess voltage at λ = 510 nm.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑30: a) Measured timing jitter distributions of unbuffered SPAD test structure for four excess voltages and b) corresponding FWHM values as a function of excess voltage at λ = 510 nm.

## Fluorescence Lifetime Measurements

Fluorescence lifetime analysis is an extremely powerful tool because of the sensitivity of a sample’s fluorescence lifetime properties to its environment. In this section, experimental results of time-resolved fluorescence lifetime measurements are presented to assess the timing and sensitivity performance of the fabricated CMOS SPAD detectors relative to a commercially available detector (MPD PDM). Here, two different fluorescence lifetime experiments were performed using two different experimental set-ups.

In the first set-up, TCSPC was utilized to obtain the fluorescence lifetime of Rhodamine 6G (R6G) using silicided and non-silicided SPAD pixels and were compared to reference measurements provided by the MPD SPAD. The measured results demonstrate the capabilities of standard low-cost CMOS technology for resolving fluorescence decays on a nanosecond time scale. To perform fluorescence lifetime analysis on samples using such a short time scale, a pico-second pulsed laser was utilized to excite the sample.

In the second set-up, time resolved fluorescence measurements of ruby crystal were performed. Because of the millisecond-range fluorescence lifetime of ruby, the lifetime measurements could be performed with a low-cost CW laser in order to demonstrate the capabilities of standard CMOS SPADs for detecting 690 nm light. This wavelength is important for biological imaging, in particular for NIR diffuse optical tomography (DOT) experiments. NIR light is also transmitted deeper into thick tissues, so it is very suitable for sub-surface biomedical imaging [291].

### Rhodamine 6G Lifetime

Rhodamine 6G (R6G) is one of the most frequently used dyes for applications in dye lasers and it is also widely used as a fluorescence tracer. Its fluorescence lifetimes are well known in a variety of different solvents [292]-[297]. In methanol and ethanol, the absorption of RG6 peaks are at 530 nm and 526 nm, respectively [297],[298]. Thus, it is ideally suited for excitation by lasers operating at 532 nm. The resulting emission spectrum of R6G varies from about 510 nm to around 710 nm, with the peak at around 550 nm, depending on the solvent and the dye concentration.

Ethanol and methanol were chosen as the solvents, since R6G fluorescence is emitted at a peak wavelength of 573 and 568 nm respectively [297], matching well with the maximum PDE of the n+/p-well shallow junction SPADs employed in this work. When a very short pulsed excitation light irradiates the R6G specimen, the molecules enter an excited state and emit fluorescence light with rapidly decaying intensity. The time *t* at which the fluorescence intensity becomes 1/*e* of the initial intensity *A0* is defined as the fluorescence lifetime *τ*. For R6G, the fluorescence decay times are between 3 and 5 nanoseconds, depending on the solvent and solution concentration [292]-[296]. At higher concentrations, the fluorescence intensity and lifetime is influenced by self-absorption and reemission (the so-called ‘inner filter’ effects) that depend on the geometrical arrangement used. These phenomena can affect the measurement results greatly [1],[293],[296].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑14: (a) Experimental set-up used to measure florescence lifetime of R6G. (b) A picture of the laboratory set-up.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑32: a) Experimental set-up used to measure florescence lifetime of R6G. b) A picture of the laboratory set-up.

A TCSPC set-up was used to obtain the fluorescence decays of RG6 using the fabricated CMOS SPADs and was compared to those obtained by the MPD SPAD used as the control device. The experimental arrangement is depicted in Fig. 5-14. A 532 nm solid-state pulsed laser (Passat Compiler 355) with 7 ps FWHM pulse width and maximum repetition rate of 200 Hz was used as an excitation source. A pulse generator (Berkley Nucleonics Model 745T) was programmed to trigger the laser at 100 Hz with a timing jitter of < 5 ps. The laser beam passed through a beam-splitter to illuminate the sample. A perpendicular geometry was used whereby the emitted fluorescence was detected along an axis at right angles to the excitation beam to limit the amount of scattered excitation light reaching the detector [1]-[3]. A long-pass filter with 550 nm cut-on wavelength was used as an excitation filter to further suppress the 532 nm excitation light incident on the SPAD detector.

The laser beam illuminated a SiPD detector to produce voltage pulses synchronous to the laser light pulses which were used to trigger the oscilloscope. The histogram of delay time between the rising edges of the SiPD triggering signal and the SPAD output was obtained from the statistics functions in the oscilloscope. Since the laser pulses were so short (7 ps FWHM), they could be regarded as near-ideal δ pulses. Re-convolution of the source IRF with an exponential decay model to obtain the measured response is thus unnecessary in this case to obtain a good estimate for the fluorescence lifetime. Rather, the fluorescence lifetime was estimated directly by fitting the data with a best-fit mono-exponential decay function. This was achieved by applying the least square data fitting method in MATLAB.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑15: (a) Measured IRF from MPD SPAD. (b) Measured fluorescence decays of R6G in ethanol (left) and methanol (right) obtained with MPD SPAD. Variations in fluorescence lifetime were apparent for different concentrations (10-4 to 10-6 M) of R6G solutions in ethanol and methanol.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑34: a) Measured IRF from MPD SPAD. b) Measured fluorescence decays of R6G in ethanol (left) and methanol (right) obtained with MPD SPAD. Variations in fluorescence lifetime were apparent for different concentrations (10-4 to 10-6 M) of R6G solutions in ethanol and methanol.

A series of aqueous R6G solutions in methanol and ethanol were prepared spanning a concentration range from 10-3 to 10-7 M [1]. They were prepared by dilution of a small volume of a concentrated standard solution. Fluorescent solutions were chosen as the target samples for SPAD characterization since they are easy to prepare and easier to align in the optical measurement set-up compared to fluorescent dyes embedded in a material. For each lifetime measurement, the R6G solutions were filled into a quartz glass cuvette and illuminated by the laser. To obtain the IRF, the fluorescent samples were replaced in the set-up by a scattering solution consisting of a laboratory standard scattering solution of colloidal silica particles mixed with distilled water at approximately the same concentration as the fluorescent solution. Fig. 5-15(a) shows the measured timing response of the optical set-up with the scattering solution (absence of fluorophore). The measured FWHM is 13.6 ps for the MPD SPAD detector and Passat Compiler laser.

TCSPC is subject to many measurement artifacts and care needs to be taken to minimize stray light reaching the detector. Fig. 5-15(b) shows the recorded fluorescence decays measured by the MPD SPAD for R6G solution in ethanol (10-5 to 10-6 M) (left) and methanol (10-4 to 10-5 M) (right). The fluorescence decays measured by the MPD SPAD show a secondary peak in the distribution which was likely caused by reflected laser light in the measurement set-up that was detected by the highly sensitive MPD SPAD. Sources of this stray light were mainly due to laser light reflections from the metallic posts and sample holders. Nevertheless, the reflected light did not significantly affect the extraction of the fluorescence lifetime when using only the tail-end of the measured exponential decay data.

The fluorescence lifetime of the solution was extracted by fitting the tail-end of the histogram data with a best-fit mono-exponential decay function. In all the measurements, the best-fit decay function characterized by a single characteristic lifetime τ. The measured lifetimes for R6G in ethanol at 10-5 and 10-6 M concentrations were 3.55 and 3.32 ns versus the reported value of 3.99 ns at 10-6 M concentration [296]. Meanwhile, the measured lifetime values were 8.86 and 4.56 ns for R6G in 10-4 and 10-5 M methanol solutions, respectively. This measured lifetimes were longer than the reported value of 4.13 ns [296], [298] because the measured lifetimes were not corrected for the effects of self-absorption and re-emission that can cause the measured lifetimes to increase with higher sample concentration [299]. In addition, variations in the precise placement of the cuvette within its holder were additional sources of the experimental error because the fluorescence quantum yield is strongly dependent on the distance the fluorescence signal travels inside the dye solution [293]. However, the measured lifetime values of R6G in methanol (10-4 M) did correspond to the value of 8.2 ns reported in [293].

In this experimental set-up, it was found that the MPD SPAD was sensitive to the lowest R6G sample concentration of approximately 10-6 M. Since the emission spectra of R6G shifts to longer wavelengths at higher concentrations [1], less fluorescence signal was detected by the SPAD because of its decreasing PDE with wavelength. At 10-6 M R6G concentration in ethanol and methanol, the measured count rates were far below the DCR of the CMOS SPADs. Therefore, more concentrated solutions (10-5 M and 10-4 M) were used for characterization of the CMOS SPADs.

The MPD SPAD used for reference measurements was replaced by the fabricated SPAD pixels at approximately the same location in the set-up in order to study their comparative performance. Measured IRFs of the non-silicided and silicided SPAD pixels are shown in Fig. 5-16(a) and Fig. 5-17(a), respectively, obtained with the scattering solution at two different excess voltages. The IRF of the non-silicided test structure pixel depended strongly on the excess voltage because an external comparator IC was used to generate the digital output pulses from the passively quenched SPAD. Limitations on the maximum threshold value relative to the SPAD bias voltage of the comparator IC resulted in more jitter at lower excess bias. This effect was not seen in the IRF of the silicided pixels because an integrated CS front-end circuit with fixed threshold was used to generate the digital signal from the passively quenched SPAD. In addition, the CS-SPAD had less excess voltage range, leading to less variation in the measured timing jitter.

The measured IRF histograms of both silicided and non-silicided SPADs showed a primary dominant peak and several smaller secondary peaks. These peaks are attributed to afterpulses as well as to the reflected light in the optical set-up which experiences different delays on its path to striking the SPAD. These secondary peaks were also present when measuring the fluorescence lifetimes with the MPD SPAD. It was found that the amplitudes of the secondary peaks were sensitive to small changes in the position of the scattering solution sample relative to the position SPAD detector in the measurement set-up. In all the measurements, the scattering and fluorescence solutions were positioned so as to minimize the light reflections and thus the amplitude of the secondary peaks.

The measured fluorescence decays obtained with a non-silicided SPAD pixel are shown in Fig. 5-16(b). This particular SPAD was chosen out of the fabricated lot for its low DCR (~500 Hz at 0.4 *VEX*), making it sensitive enough to obtain the fluorescence lifetimes of R6G solutions down to 10-5 M concentration. Below this concentration, the fluorescence decay could not be detected due to the dominance of the background DCR over the fluorescence photon counts. Fig. 5-16(b) left panel shows that the fitted decay curves remain approximately parallel for both excess voltages as they all represent the same sample lifetime of R6G in ethanol (10-5 M). The measured lifetimes for *VEX* = 0.4 V and 1 V were 3.92 ns and 3.75 ns, respectively. These were similar to the 3.55 ns lifetime obtained by the MPD SPAD for R6G in ethanol (10-5 M). The R6G solution in methanol (10-4 M) showed greater variation between the extracted lifetimes at both excess voltages due to the lower signal counts. This occurs because at higher concentrations the self-absorption affects the peak intensity and lifetime of the fluorescence signal [1],[299].

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑16: (a) Measured IRFs of the non-silicided SPAD pixels obtained with the scattering solution at two different excess voltages. (b) Measured fluorescence decay data of R6G solutions in ethanol (left) and methanol (right) at 10-5 M and 10-4 M concentrations, respectively. Best fit exponential decays are also shown with their corresponding lifetimes.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑36: a) Measured IRFs of the non-silicided SPAD pixels obtained with the scattering solution at two different excess voltages. (b) Measured fluorescence decay data of R6G solutions in ethanol (left) and methanol (right) at 10-5 M and 10-4 M concentrations, respectively. Best fit exponential decays are also shown with their corresponding lifetimes.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑17: (a) Measured IRFs of the silicided SPAD pixels obtained with the scattering solution at two different excess voltages. (b) Measured fluorescence decay data of R6G solutions in ethanol (left) and methanol (right) at 10-5 M and 10-4 M concentrations, respectively. Best fit exponential decays are also shown with their corresponding lifetimes.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑38: a) Measured IRFs of the silicided SPAD pixels obtained with the scattering solution at two different excess voltages. (b) Measured fluorescence decay data of R6G solutions in ethanol (left) and methanol (right) at 10-5 M and 10-4 M concentrations, respectively. Best fit exponential decays are also shown with their corresponding lifetimes.

The measurement results of silicided CS-SPAD pixels are shown in Fig. 5-17. The IRF (Fig. 5-17(a)) fluorescence decay (Fig. 5-17(b)) were measured at *VEX* = 0.9 V and 1.1 V. The range of excess voltages that could be measured for this pixel was limited by the front-end circuit, so the measured IRFs and fluorescence decays show little variation between the two excess voltages. However, the improved timing performance for the CS-SPAD pixel over the unbuffered pixel is apparent. The corresponding fluorescence lifetimes matched reasonably well with those obtained by the MPD SPAD.

### Ruby Crystal Lifetimes

Ruby crystal (Cr3+:Al2O3) makes an excellent specimen for fluorescence lifetime experiment. This is because it has a stable and well known fluorescence lifetime from its use in pulsed ruby lasers, high-pressure sensors, and fiber-optic temperature sensors [300]-[303]. Additionally, ruby is a very convenient sample with no sample preparation unlike the fluorescent R6G solutions. Also, it can be handled and stored easily and it is quite inexpensive.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑18: (a) The measurement set-up used to measure the fluorescence lifetime of ruby crystal. (b) A photograph of the laboratory set-up.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑40: a) The measurement set-up used to measure the fluorescence lifetime of ruby crystal b) A photograph of the laboratory set-up.

Ruby has two broad absorption bands at approximately 410 nm and 550 nm and the fluorescence emission profile of ruby has two peaks at 694.3 and 692.7 nm [302]. Thus, the fluorescence decay of ruby is of interest to study the NIR performance of the fabricated CMOS SPAD. Further, the milliseconds range lifetime of ruby permits the use of Hertz-range light modulation rather than picosecond pulsed lasers. Finally, the fluorescence intensity can be measured by pulse counting in a fixed integration time, which requires relatively simple digital logic circuits, rather than by TCSPC which requires high-performance TDC circuits. The optical instrumentation and electronic circuits are thus relatively inexpensive and can be set-up easily so that time-resolved fluorescence measurements using CMOS SPADs can be performed with affordable and relatively simple instrumentation [304].

Fig. 5-18 is a block diagram of the setup used. A low-cost 405 nm CW laser diode was chosen because it offers a cheap and reliable excitation source near the absorption band of ruby. To modulate the CW light, the light beam from the laser was aligned onto the chopper blade of a variable-speed beam chopper (Thor Labs, Chopper Head 220A). The chopper frequency was set at 20 Hz so that the resulting laser excitation period was 50 ms. It was important to align the excitation light onto the chopper blade so as to obtain a 50% duty cycle of the excitation signal and minimize the on-off transient flanks during the repetitive excitation. The ruby sample was placed in a quartz cuvette and the excitation light was directly incident on the ruby sample when passing through the chopper blades. The light emitted from the ruby was collected and collimated into a parallel beam with a single lens, passed through a 690-nm BPF, and finally focused onto the detector. An electrical signal synchronous to the chopper frequency was used as a trigger source for the oscilloscope and the measured SPAD signal traces were recorded by the oscilloscope for post-processing using MATLAB.

Calibration measurements performed with the reference MPD SPAD are shown in Fig. 5-19. The photon counts of the MPD SPAD (normalized by their amplitude) are plotted in Fig. 5-19(a). The blue traces represent the digital output pulses of the MPD SPAD corresponding to individual photon detections. The chopper reference signal is also shown (black trace) indicating the position of the chopper blade. When laser light passed through the rotating chopper blade, the intensity of the fluorescence signal at the detector is at a maximum and the chopper reference signal is high. When the chopper reference signal is low, the excitation light is blocked by the chopper blade and the resulting photon counts are due to the delayed fluorescence emission. Using a single 128 ms oscilloscope acquisition period, ~2.5 excitation cycles could be recorded in a single acquisition. The total number of oscilloscope acquisitions was 150, so ~375 excitation cycles were averaged in total.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑19: Measured data from the set-up in Fig. 5-10 with reference MPD SPAD. The intensity signals are shown for two different integration times. The chopper reference signal is also shown. (a) The acquisition time of the oscilloscope was set to 128 ms to record 2.5 excitation cycles using a 1 GS/s oscilloscope sample rate. (b) When the chopper reference signal is low the chopper blade blocks the incident light and the fluorescence decay can be obtained by fitting the decay portion of photon count rate with an exponential function.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 5‑42: Measured data from the set-up in Fig. 5-10 with reference MPD SPAD. The intensity signals are shown for two different integration times. The chopper reference signal is also shown. a) The acquisition time of the oscilloscope was set to 128 ms to record 2.5 excitation cycles using a 1 GS/s oscilloscope sample rate. b) When the chopper reference signal is low the chopper blade blocks the incident light and the fluorescence decay can be obtained by fitting the decay portion of photon count rate with an exponential function.

The average PCR corresponding to the average measured light signal intensity was obtained using two integration times, *TINT* = 100 and 500 μs. The number of photon counts within an integration time was calculated during post-processing of the individual SPAD photon counts. The number of counts occurring during the integration time was averaged over all the excitation cycles to obtain the average PCR. Both PCR curves show the same temporal behavior, but the signal obtained with *TINT* = 500 μm had higher SNR since more photons were counted during this interval.

The decay of the PCR corresponds to the fluorescence decay when the chopper blade blocks the incident light, as shown in Fig. 5-19(b). The data in this interval was fitted to an exponential decay model and the resulting decay time constant was approximately 3 ms, agreeing reasonably well with the reported lifetime of ruby at room temperature [301]-[303]. It is important to note that the precise lifetime of each ruby sample depends on level of Cr3+ doping level which was not known for the measured sample.

Following the reference measurements, the MPD SPAD was replaced in the set-up by the non-silicided SPAD test structure. The measured results are shown in Fig. 5-20 for two different integration times. Fig. 5-20(a) shows that the normalized PCR of the CMOS was accompanied with a reduction in SNR due to the lower PDE compared to the MPD SPAD. Nevertheless, a fluorescence lifetime of 2.97 ms was extracted from the measurements, agreeing well with the calibration measurements.

|  |  |  |
| --- | --- | --- |
|  | |  |
| (a) | (b) | |

Figure 5‑20: Measurement results for ruby lifetime obtained with the non-silicided SPAD pixel. (a) The measured SPAD output is plotted along with the extracted photon counting rates for two different integration times. (b) Fitting result is shown for the portion of the exponential decay in photon count rate.

|  |  |  |
| --- | --- | --- |
|  | |  |
| (a) | (b) | |

Figure 5‑44: Measurement results for ruby lifetime obtained with the non-silicided SPAD pixel a) The measured SPAD output is plotted along with the extracted photon counting rates for two different integration times b) Fitting result is shown for the portion of the exponential decay in photon count rate.

|  |  |  |
| --- | --- | --- |
|  | |  |
| (a) | (b) | |

Figure 5‑21: Measurement results for ruby lifetime obtained with the silicided SPAD pixel. (a) The measured SPAD output is plotted along with the extracted photon counting rates for two different integration times. (b) Fitting result is shown for the portion of the exponential decay in photon count rate.

|  |  |  |
| --- | --- | --- |
|  | |  |
| (a) | (b) | |

Figure 5‑46: Measurement results for ruby lifetime obtained with the silicided SPAD pixel a) The measured SPAD output is plotted along with the extracted photon counting rates for two different integration times b) Fitting result is shown for the portion of the exponential decay in photon count rate.

|  |  |  |
| --- | --- | --- |
|  | |  |
| (a) | (b) | |

Figure 5‑22: (a) Comparison of measured photon counts between MPD SPAD and silicided and non-silicided CMOS SPADs. (b) Corresponding plot of fluorescence decays on a y-axis log scale from which the fluorescence lifetime of ruby crystal was obtained.

|  |  |  |
| --- | --- | --- |
|  | |  |
| (a) | (b) | |

Figure 5‑48: a) The measured input photon flux striking the detector along with the uncorrected SPAD counting rate is plotted as a function of the optical attenuation b) The corrected SPAD counting rate

|  |  |  |
| --- | --- | --- |
|  | |  |
| (a) | (b) | |

Figure 5‑49: a) Comparison of measured photon counts between MPD SPAD and silicided and non-silicided CMOS SPADs b) Corresponding plot of fluorescence decays on a y-axis log scale from which the lifetime of ruby crystal was obtained.

|  |  |  |
| --- | --- | --- |
|  | |  |
| (a) | (b) | |

Figure 5‑50: a) The measured input photon flux striking the detector along with the uncorrected SPAD counting rate is plotted as a function of the optical attenuation b) The corrected SPAD counting rate

The silicided SPADs were tested as well with results shown in Fig. 5-21. The SNR was the poorest on account of the very low PDE, and the extracted lifetime was 3.22 ms. The error of the lifetime could be improved by increasing the number of excitation cycles in the measurement to average more photon counts. Fig. 5-22 shows a direct comparison of the measured photon counts for each measured device. It is clear that the measured fluorescence signal was the weakest for the silicided SPAD. However, all three measured SPADs were able to resolve the fluorescence decay of ruby.

# CMOS Time-to-Digital Converter Design, Simulation and Measurements

The unique operation principle of SPADs enables the detection of very weak optical power down to the single-photon level with sub-nanosecond timing resolutions. Therefore, the design of SPADs in a standard, low-cost, digital CMOS technology brings an entirely different set of challenges compared to that of CIS technology. In contrast to conventional APS, SPAD pixels do not produce an output voltage that is proportional to the overall photon flux. Rather, SPADs produce a stream of digital pulses corresponding to the absorption of single photons. As such, they are ideally suited for integration with high-performance mixed-signal circuits such as TDCs in CMOS to measure fluorescence lifetime decays with high resolution and accuracy.

A high-speed, compact and precise TDC prototype chip designed for SPAD imaging systems is presented in this chapter. The 6-bit TDC achieves sub-nanosecond resolution and high-speed operation with high accuracy. The maximum sampling rate and *TLSB* of the TDC are 60 MHz and 130.2 ps, respectively. Small circuit size and moderate power consumption are achieved by utilizing a coarse-fine delay-line architecture. The 0.61 LSB rms resolution measured over the dynamic range is attributed to the low non-linearity and low jitter of the coarse-fine delay lines. Compact size and robustness to PVT variations makes the TDC a suitable building block for multi-channel TCSPC and ToF systems such as FLIM. Full details of the design, simulation and measurements are described in this chapter.

## TDC Architecture and Design

TDCs that measure time intervals between two input signals are a fundamental building block of single-photon imaging systems [194]. In the past, TDCs for TCSPC systems were designed as stand-alone custom chips integrated separately from single-photon detectors [28]. With the advent of DSM CMOS technology, low-cost SoC integration of TDC with SPAD became feasible [43]-[49],[61],[62],[69]-[73],[196],[197],[221]. Important performance metrics of TDCs designed for high-speed TCSPC applications include high resolution and low non-linearity, as well as large dynamic range to accurately measure a wide range of timing signals. To facilitate multi-channel operation, low circuit area and power consumption, as well as robustness to process variations, are other important design prerequisites.

Traditionally, the main limitation of FLIM systems has been a slow frame rate due to the large SNR requirement for each pixel in order to distinguish the fluorescence lifetime signal from uncorrelated background noise [31]. An in-pixel TDC approach is commonly used to achieve high frame rates, whereby each SPAD and its associated front-end circuitry is connected to a dedicated TDC inside the pixel. Integrated arrays of SPADs with in-pixel TDCs have been developed for parallel acquisition in FLIM [42]-[46],[196],[221]. However, SPADs with in-pixel TDC are typically associated with very small fill-factor. On the other hand, TDC sharing among SPADs yields higher pixel fill-factor, while relaxing the TDC constraints on TDC area and power consumption. This results in TDCs that have lower INL and better uniformity [69][70],[197].

The proposed TDC is targeted for SPAD/TDC architectures which utilize TDC sharing for a reduction of silicon area and reduced power consumption. The performance must also be stable over different operating temperatures and voltage supply variations. For multi-channel realization, the variation of TDC performance between channels should also be reduced. Other important goals for the TDC design include < 1 LSB INL and high sampling rate (~50 MHz). Ideally, the precision of the TDC should be no worse than the FWHM of the SPAD intrinsic timing jitter, which has been measured to be between 70-150 ps. Thus, time digitization using multi-level interpolation can be carried out since time resolution range is achievable by VCDLs in 130 nm CMOS. IBM’s 130 nm CMOS technology was chosen for the fabrication of SPAD/TDC chips because this process provides a comprehensive suite of devices including thin and thick oxide FETs, triple-well and zero threshold voltage FETs and high integration density for high performance TDC circuits. In DSM technology, lower power consumption and gate delays as low as 10 ps are achievable, in addition to higher integration density compared to other technologies (180 and 350 nm) as a result of the STI and smaller lithographic feature sizes [224], [228]. On the other hand, the SPAD performance may be worse in DSM CMOS with respect to PDE and DCR performance, so the main benefit of using a standard digital process for single photon imaging is the much lower fabrication costs in conjunction with the integration of high performance mixed-signal circuits.

|  |
| --- |
| D:\PHD_2015\PAPERS\TCAS15\FIGURES\Fig 1 - Copy.tif |

Figure 6‑1: Coarse-fine TDC timing diagram. The fine interpolator calculates the time between rising edges of START2 and STOP2. This result *T2* is used to obtain the residue time *TR*, which is added to the coarse-interpolation result *T1* to digitize the timing interval *TM*.

|  |
| --- |
|  |

Figure 6‑2: Coarse-fine TDC timing diagram. The fine interpolator calculates the time between rising edges of START2 and STOP2. This result *T2* is used to obtain the residue time *TR*, which is added to the coarse-interpolation result *T1* to digitize the timing interval *TM*.

Fig. 6-1 illustrates the operational principle of a coarse-fine delay line TDC. Coarse-fine delay line interpolators are used to reduce the number of delay cells required to digitize the time interval, thus keeping the TDC core area small and the INL low (since a shorter VCDL is used). The reference clock period is sub-divided into *NC*×*NF* time intervals. The number of coarse/fine quantization steps, (*NC*/*NF*) are

|  |  |  |
| --- | --- | --- |
|  |  | (6-1) |

where *TCLK* is the reference clock period, *TC* is the delay of a single element in the coarse delay line.

In this work, the reference clock period is divided by a delay line into *NC* = 16 equally spaced coarse phases. Each coarse interval is further sub-divided into *NF* = 4 fine phases. Upon arrival of the STOP signal’s rising edge, the span of time since the preceding START signal’s rising edge is quantized with a thermometer code that represents the coarse interpolation result *T1*. The time elapsed between the STOP transition and the transition of the first successive phase of the coarse VCDL is measured by the fine TDC to obtain the fine interpolation result *T2*. Thus, the residue time *TR* is

|  |  |
| --- | --- |
| . | (6-2) |

The residue time *TR*is added to *T*1 to give the measurement time *TM*. With a 120 MHz clock, 16 coarse and 4 fine delay elements result in a 6-bit TDC with an LSB width of 130.2 ps according to eqns. (6-1) The coarse-fine TDC requires only 20 delay elements to achieve the given *TLSB* resolution compared to the 64 required with a single interpolating stage. The length of the delay line is thus kept relatively short, leading to better INL performance, since the INL tends to grow as the number of delay elements increases [199].

Fig. 6-2 illustrates the proposed architecture. *TDC*1 and *TDC*2 are comprised of coarse and fine VCDLs, respectively. The START signal is a PVT-invariant reference clock signal that propagates through the VCDL comprising *TDC*1. The clock is also fed to a replica VCDL of *TDC*2 (*VCDL*2). Regulation of the *VCDL*1/*VCDL*2 delay is done by *DLL*1/*DLL*2 consisting of two phase detectors (*PD*1/*PD*2) and charge pumps (*CP*1/*CP*2). The phase detectors compare the signals CLK and REF at their inputs, and generate UP/DOWN signals to their respective charge pump, depending on whether the CLK input is leading/lagging the REF input. The charge pump adjusts the control voltages *VCTRL*1/*VCTRL*2 to fine tune the delays of *VCDL*1/*VCDL*2*,* so that the inputs of the respective phase detectors are eventually locked in phase. This configuration ensures that the coarse and fine delays of the TDC are defined only by the external reference clock. The DLLs reduce the PVT sensitivity of the TDC, thus satisfying eqns. (2) and (3) over different operating conditions.

The multi-phase clock signal ϕ1 propagating through *VCDL*1, is fed into a synchronizer (*SYNC*) logic circuit to generate the precise timing signals for *TDC*2, based on *TDC*1coarse-interpolation result *O*1. *SYNC* generates a STOP2 signal for *TDC*2 to calculate the coarse-interpolation residual time, improving the coarse resolution by the factor *NF*. This circuit was designed with static logic to minimize its power and area consumption, and to produce STOP2for residual time calculation with minimal delay. A circuit replicating the *SYNC* delay is placed before *TDC*2 to compensate for the delay of *SYNC* logic, as described later in detail in section 6.1.2.

|  |
| --- |
| D:\PHD_2015\PAPERS\TCAS15\FIGURES\Fig 2 - full.tif |

Figure 6‑2: Architecture of a single channel of the proposed coarse-fine interpolating TDC. Only the circuits in the hatched portion of the TDC core need to be replicated for a multi-channel realization.

|  |
| --- |
|  |

Figure 6‑4: Architecture of a single channel of the proposed coarse-fine interpolating TDC. Only the circuits in the hatched portion of the TDC core need to be replicated for a multi-channel realization.

The TDC is designed to operate in the TCSPC mode, with the STARTsignal synchronized to a laser pulse. A STOP signal that arrives within a reference clock period signifies the detection of a photon. The 6-bit dynamic range of the TDC is suitable for FLIM applications that measures the lifetime of fluorophores up to several nanoseconds (ns). The maximum measurement range of the TDC in this work is 12.5 ns, corresponding to an 80 MHz reference clock. The measurement range could be increased by using a digital counter to count the number of clock cycles elapsed between START/STOP intervals longer than one reference clock period, but this was not implemented in the prototype described here.

### Voltage-Controlled Delay Lines (VCDL)

The objective in the design of the VCDL was adequate matching between the propagation delays of the delay cells. Due to the linear structure of these two delay lines, the measurement accuracy greatly depends on the matching of various delay cells [199]. Ideally the DLL provides as many evenly spaced clock signals within the clock cycle as there are delay cells in the delay line. However, the variation of the propagation delay of the delay cells caused by device mismatch is due to the variation of process parameters appears as nonlinearity of the delay line structures. Also the power supply noise caused by the bonding wires and wiring resistances introduce delay variations in the delay cells. The nonlinearity caused by the random variation between delay cells can be minimized at the circuit level by good device matching, and by restricting the number of delay cells in the delay line with the use of the coarse-fine architecture.

A VCDL delay element is shown in Fig. 6-3(a). The delay element is based on a differential-amplifier which improves power supply and substrate noise rejection compared to the single-ended inverter-based delay elements [305]. The delays of the differential amplifier cells are determined by the slew-rate limited rise/fall times of the signals at the drains of the pFET input transistors, which in turn depend on the capacitance of the nFET load transistors and the current supplied by transistor *PS*. A differential-to-single-ended converter (Diff.-to-S.E.) circuit restores the full logic swing of differential amplifier output signals. The full-swing, single-ended signal is converted back to differential format by an edge aligning circuit, ensuring that signal transitions are aligned and de-skewed [218].

Fig. 6-3(b) shows the structure of the coarse and fine TDCs. A replica bias circuit (V-to-I) was used to generate the bias current *IBIAS* from the control voltage *VCTRL* generated by the DLL, thereby allowing the VCDL delay to be regulated by the voltage across the DLL’s loop-filter capacitor as described in Section 6.1.3. Each phase of the delay line was converted to single-ended format by a buffer circuit to drive the sampling circuits and *SYNC* logic. It was also used to isolate the VCDL cells from delay mismatches due the different lengths of the metal interconnections used to route signals between *TDC1* and *SYNC*.

|  |
| --- |
| D:\PHD_2015\PAPERS\TCAS15\FIGURES\Fig 5a.tif |
| (a) |
| D:\PHD_2015\PAPERS\TCAS15\FIGURES\Fig 5b - comp.tif |
| (b) |
|  |
| (c) |

Figure 6‑3: Voltage Controlled Delay Line (VCDL) (a) Schematic of VCDL delay cell used for coarse and fine interpolation, comprised of a differential stage (diff. stage), differential to single-ended (diff. to SE) conversion stage, and edge aligner stage. (b) Block diagram of VCDL. A single-ended to differential (SE-to-DIFF) converter is used to convert the single-ended input signal to differential format. A Voltage-to-Current (V-to-I) converter generates a bias current from *VCTRL*. (c) Layout of coarse VCDL.

|  |
| --- |
|  |
| (a) |
|  |
| (b) |

Figure 6‑6: Voltage Controlled Delay Line (VCDL) schematics (a) Schematic of VCDL delay cell used for coarse and fine interpolation, comprised of a differential stage (diff. stage), differential to single-ended (diff. to SE) conversion stage, and edge aligner stage. (b) Block diagram of VCDL.

The layout of the coarse VCDL is shown in Fig. 6-3(c). To minimize the impact of variation of transistor parameters on the VCDL delays, transistors were interdigitated to ensure good matching of device parameters. Also, the differential stage transistors had larger gate areas and operated in strong inversion to reduce the statistical mismatches between delay elements, thus improving the linearity of the interpolators [209],[210]. Additionally, the routing of all metal interconnections along the delay-lines was carefully done to match the metal trace lengths and avoid delay mismatches and dummy elements were used to equalize parasitic capacitance loading between delay-line cells. Dummy elements were placed at the beginning and end of each VCDL to ensure equal loading of all delay elements. The individual cells were carefully laid out such that the inputs and outputs of the neighboring cells directly abut with the least amount of metal in order to minimize parasitic capacitance of the delay chain, thus minimizing the inverter delay. The balanced cell layout ensured equal delays of the complementary delay chains. Every other cell was flipped along the horizontal axis in order to further desensitize the complementary delay mismatches from process gradients and metal interconnect parasitics.

Symmetric rise/fall times for each delay element were designed for all corners of process variations in the IBM 130 nm CMOS technology so that the duty cycle for each phase of the delay line was constant. Symmetric rise/fall times prevent the signals from disappearing during propagation in the delay line, as in pulse-shrinking TDCs. Special attention was paid on the symmetry and matching of the VCDL circuits in the layout. Further, the power supply of the TDC was partitioned into analog and digital sections to reduce noise coupling from the digital logic into the sensitive analog circuits. Substrate guard rings and large on-chip decoupling capacitors (300 pF) were utilized to achieve good VCDL supply noise immunity.

Fig. 6-4 shows the post-layout simulation results of the coarse and fine delay element characteristics for slow, nominal and fast CMOS process corners. The results show that the control voltage of the delay cells can produce the targeted coarse/fine delay values for the slow, nominal and fast process variations. The nonlinearity and gain of the delay characteristic in the operating range corresponding to a 120 MHz reference clock were minimized to reduce the sensitivity of VCDL to small fluctuations of control voltage. Reduced VCDL gain also helped to minimize the change in DLL loop-gain due to process variations [17],[27].

|  |  |
| --- | --- |
| D:\PHD_2015\PAPERS\TCAS15\FIGURES\Fig 6a - comp.tif | D:\PHD_2015\PAPERS\TCAS15\FIGURES\Fig 6b - comp.tif |
| (a) | (b) |

Figure 6‑4: Post-layout simulation results of delay transfer characteristic of delay element and charge pump mismatch for (a) coarse and (b) fine interpolators. The range of VCDL control voltages that produce *TC* and *TLSB* at different process corners are indicated for 120 MHz reference clock frequency. In this range, the charge pump mismatch is at its lowest value, resulting in lower static phase error.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 6‑8: Post-layout simulation results of delay transfer characteristic of delay element and charge pump mismatch for (a) coarse and (b) fine interpolators. The range of VCDL control voltages that produce TC and TLSB at different process corners are indicated for 120 MHz reference clock frequency. In this range, the charge pump mismatch is at its lowest value, resulting in lower charge pump phase offset.

### Synchronizer Logic

The synchronizer circuit generates the proper timing signals so that *TDC2* can accurately calculate the residual time interval as illustrated in Fig. 6-1. *SYNC* logic selects the first multi-phase transition immediately successive to the STOP transition to produce the STOP2 signal for *TDC2*. The *SYNC* circuit was realized with combinational logic circuits which are advantageous because they consume no static power, are fast, and are of small area in DSM CMOS. The transistors were sized just greater than the minimum size to have sufficient capability to drive the inputs of successive logic elements and the interconnection load. Since there are no flip-flop elements in the synchronizer, then there are no potential set-up or hold time violations [201]-[203],[219]. As a result, the delay of the synchronizer is guaranteed to be below *TLSB* and is in fact defined by the propagation delay of the *SYNC* logic.

|  |  |
| --- | --- |
| D:\PHD_2015\PAPERS\TCAS15\FIGURES\Fig 3a - comp.tif | D:\PHD_2015\PAPERS\TCAS15\FIGURES\Fig 3b - comp.tif |
| (a) | (b) |

Figure 6‑5: (a) Synchronizer logic diagram and (b) post-layout simulation results for case when *TM* = 5.1 ns and *TLSB*=156.25 ps (corresponding to a 100 MHz reference clock). In this case, *TM*/*TC* > 8, and according to *SYNC* logic, STOP2=ϕS*=*ϕ1(9). The *SYNC* delay between ϕ1(9)and STOP2is Δ2.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 6‑10: a) Synchronizer logic diagram and b) post-layout simulation results for case when *TM* = 5.1 ns and *TLSB*=156.25 ps (corresponding to a 100 MHz reference clock). In this case, *TM*/*TC* > 8, and according to *SYNC* logic, STOP2=ϕS*=*ϕ1(9). The *SYNC* delay between ϕ1(9)and STOP2is Δ2.

The *SYNC* logic is illustrated in Fig. 6-5(a). The VCDL Sampler samples the state of the multi-phase clock, *ϕ1*, at the instant when STOP arrives, producing a thermometer code *O1* representing the coarse interpolation result. The Decode Delay logic detects the first ‘1’ to ‘0’ transition in O1, and generates a ‘one-hot’ encoded selection signal on bus S according to

|  |  |
| --- | --- |
| . | (6-2) |

The ‘one-hot' encoded vector signal S is sent to the *ϕ*-Select logic, to select the first multi-phase clock transition successive to STOP according to

|  |  |
| --- | --- |
| . | (6-3) |

When STOP arrives before any positive transition of the multi-phase clock, S is identically ‘0’, so the MUX gives *ϕS*= *ϕ1*(1). Status bit *N0C* indicates this special case. For all other cases, S is ‘one-hot’ encoded, and eqns. (6-2) and (6-3) apply. An important consideration in the synchronizer design is the effect of the propagation delay of the *SYNC* logic on the fine conversion. The delay of the SYNC logic causes STOP2 to be delayed by an amount Δ2 with respect to the first coarse VCDL transition successive to the STOP transition. If this delay is not properly accounted for, then the fine interpolator will get the wrong result and introduce an error in the residual calculation. This delay can be compensated by adding a delay Δ1 between STOP and START2, as shown in Fig. 6-5(b). The *SYNC* replica circuit in *TDC2* ensures that the delay between STOP and START2 is Δ1 ≈ Δ2. The time interval *T2* measured by the fine interpolator then accurately represents the residue time according to eq. (4).

### Delay-Locked-Loop (DLL)

Two DLLs are integrated with the TDC to regulate the delay of the VCDLs, as illustrated in Fig. 6-2. *PD*1 compares the phase difference between the reference clock and the last delay element of *VCDL*1, ϕ1(16). *CP*1 adds or removes charge from the loop filter capacitor based on the phase difference at the phase detector inputs. A large loop filter capacitor is preferred to minimize the DLL’s jitter. This comes at the cost of a longer DLL locking time, which is not a concern in this application. To prevent false locking, a start-control circuit is utilized. At start-up, the loop-filter capacitor is fully charged, forcing the total *VCDL*1 delay to be less than the reference clock period. Once *DLL*1 is enabled, the loop filter capacitor begins discharging by a fixed amount during each clock cycle, causing *VCTRL*1 to adjust until the delay between *PD*1 inputs is equal to one period of the reference clock, so that .

The primary goal for the design of the PD/CP circuits is to minimize the systematic static phase offset. With a large static phase offset, the ideal phase spacing between the multi-phase clock depicted in Fig. 6-1 no longer applies, resulting in higher quantization error and larger INL. The static phase offset is primarily determined by the mismatch between the UP and DOWN currents of the CP during the idle time of the PD. The static phase offset error can be expressed as

|  |  |
| --- | --- |
| , | (6-3) |

where *Tidle* is the idle time of the PD, *TCLK* is reference clock period, *ICP* is the average CP current and Δ*ICP* is the CP current mismatch. The idle time is due to both UP and DOWN signals being asserted when the input signals are in-phase, so that the PD’s dead-zone is eliminated [211],[212]. In this work, a dynamic PD consisting of two true-single-phase clock (TSPC) latches was selected for its short idle time of approximately 90 ps [212]. A current-steering charge pump was utilized because of its excellent high-speed and current matching performance [213],[214].

Fig. 6-4 illustrates the current mismatch of the coarse and fine interpolator charge pumps. With ideal circuits, the phase difference at the PD inputs would be zero once the DLLs are in the locked state. However, a small static phase offset is always present due to non-idealities of the PD/CP circuits, which contributes to the INL [53]. By reducing the mismatch of the PD/CP, the phase offset can be reduced, and the delays of the coarse and fine interpolators are given by eqns. (2) and (3). At the nominal control voltage values (between 0.5 and 1.2 V for coarse interpolator and 0.4 and 0.75 V for fine interpolator) which produced the desired coarse and fine delays across slow, nominal and fast process corners, the current mismatch was -1 μA and 0.75 μA, respectively. With charge pump currents of 15 μA and 7.5 μA for *CP1* and *CP2*, respectively, the static phase offset error for coarse and fine DLLs is approximately 0.14% and 0.22%, respectively. The DLLs were designed to operate with a reference clock frequency between 80 and 120 MHz, and *TLSB* between 130.2 and 195.3 ps, corresponding to a measurement range between 8.3 and 12.5 ns.

### Data Read-out scheme

The 6-bit output code of the TDC is comprised of 4 coarse interpolation bits, 2 fine interpolation bits, as well as two status bits, according to

|  |  |
| --- | --- |
| . | (6-4) |

The coarse VCDL bits are ‘one-hot’ encoded in the synchronization logic and then converted to binary number format by a 16-to-4 decoder. Similarly, the fine VCDL thermometer code is ‘one-hot’ encoded and then converted to binary number format by a 4-to-2 decoder.

The coarse and fine status bits indicate special cases of the synchronization logic. *N0C* indicates whether STOP1 signal has arrived before or after the transition of the first phase of the coarse interpolator *ϕ1*(1). *N0F* indicates whether the STOP2 signal arrives before or after the transition of the first phase of the fine interpolator *ϕ2*(1). For example, in Fig. 6-5(b), 8×*TC* = 5 ns is the coarse interpolation result. The coarse status bit is *N0C* = 1. The fine interpolator result is *NF* = 3, since STOP2 arrives after *ϕ2*(3) and before *ϕ2*(4). The fine status bit is *N0F* = 1, since STOP2 arrives after *ϕ2*(1). The corresponding residual is (4 – 4)×*TF* = 0. The status bits are also used to trigger the data read-out instruments indicating the occurrence of a TDC conversion by their LO-HI transitions. At 60 MHz sampling, the maximum data rate for the TDC was 480 Mbps.

A timing diagram of the TDC operation is shown in Fig. 6-6(a). Fig. 6-6(b) shows the post-layout simulation results of the TDC read-out. A master reset signal (RST) initializes all the on-chip read-out registers and enables the TDC for operation. Two controls signals, DLL1\_EN and DLL2\_EN, enable the DLLs. Once the DLLs are locked after approximately 7 μs the TDC is outputting valid data. The START signal is synchronous with a laser reference and functions as the reference clock of the TDC. For testing purposes, a delayed replica of the reference clock is used as the STOP signal. The START signal is fed to an on-chip divide-by-two circuit, which produces the read-out clock, CLKDIV, as well as the reset signal for the TDC registers.

|  |
| --- |
| D:\PHD_2015\PAPERS\TCAS15\FIGURES\Fig 4 - comp.tif |
| (a) |
|  |
| (b) |

Figure 6‑6: (a) Timing diagram of the TDC read-out. Once the DLLs are locked, the TDC outputs valid data M1, M2, M3, etc. (b) Post-layout simulation results of TDC read-out. Locking times of DLL1 and DLL2 are approximately 7 μs.

|  |
| --- |
|  |
| (a) |
|  |
| (b) |

Figure 6‑12: (a) Timing diagram of the TDC read-out. Once the DLLs are locked, the TDC outputs valid data M1, M2, M3, etc. (b) Post-layout simulation results of TDC read-out. Locking times of DLL1 and DLL2 are approximately 7 μs.

The TDC is designed to operate in time-interleaved mode as in [64],[197], whereby the TDC is enabled when CLKDIV is in the HI state, and reset when CLKDIV is in the LO state. The falling edge of the CLKDIV latches the TDC data from the TDC registers onto the read-out registers at the end of the reference clock cycle. So, while the TDC registers are being cleared for the next START-STOP measurement, the data from the previous START-STOP measurement is latched onto the output registers. However, the correct latching operation near the end of the reference clock cycle required that no code transitions occur during the set-up and hold times of the flip-flops. Therefore, approximately 1 ns of dynamic range (corresponding to 8 of the 64 TDC codes for *TLSB* = 130.21 ps) had to be used to ensure correct latching operation of the read-out circuits.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
|  | | |
| (c) | | |

Figure 6‑7: (a) Photograph of the TDC chip highlighting the TDC core area (b) Top level diagram of TDC chip and photograph of the printed circuit board (PCD) used for testing the TDC prototype chips. (c) TDC core layout with main components highlighted.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
|  | | |
| (c) | | |

Figure 6‑14: (a) Photograph of the TDC chip highlighting the TDC core area (b) Top level diagram of TDC chip and photograph of the printed circuit board (PCD) used for testing the TDC prototype chips. (c) TDC core layout with main components highlighted.

In this implementation, only one TDC channel was implemented. Using a 120 MHz reference clock the maximum sampling rate is 60 MHz. Implementation of a second TDC on the chip could double the maximum sampling rate to 120 MHz.

## Measured TDC Performance

The TDC was fabricated in a standard 130 nm digital IBM CMOS technology. The chip dimensions are 2 mm × 2 mm, with the TDC core occupying 0.04 mm2 of silicon area as shown in Fig. 6-7(a). Fig. 7-5(b) shows a top level diagram of the TDC chip, with the indicated arrangement of input, output, control and power supply pins, as well as photograph of the PCB that was used for testing. Fig. 6-7(c) illustrates the TDC core layout. The main components of the layout include on-chip decoupling capacitors (300 pF) used to improve the VCDL supply noise immunity and I/O blocks that were designed to drive off-chip 15 pF capacitive probes and provide input matching to the 50 Ω RF cables. The bias circuit was designed to provide stable voltage and current references for the TDC core over the temperature range between -40 and 40 °C. The silicon area of the TDC core area that would be replicated on the chip in a multi-channel realization (hatched area in Fig. 6-2) was only 0.0231 mm2.

The total power consumption of the chip was 35 mW for a 120 MHz reference clock and a 1.5 V supply. A significant portion of the power consumed is consumed by the I/O blocks which were designed to drive high-capacitance probes and cables. The TDC core power consumption was 7 mW from simulations. 40% of the TDC power was attributed to VCDL1 and VCDL2, which required large bias currents to attain the range of coarse and fine time resolution.

The TDC chip was packaged in a 68-pin ceramic PGA package and mounted on a PCB for testing purposes as shown in Fig. 6-7(b). Ten chips from the fabricated lot were measured to assess the chip-to-chip variation and the effect it has on performance of the TDC at room temperature and at -30 °C.

### Measurement Set-up

The measurements to characterize the TDC’s performance were performed at sampling rates of 1 MHz and 60 MHz. In both cases the TDC output was recorded by a LeCroy WaveRunner 640Zi mixed-signal oscilloscope. The measurement set-up for the TDC measurements at 1 MHz is illustrated in Fig. 6-8(a). A 120 MHz frequency was provided by a frequency synthesizer (Agilent, 83752A) and the reference clock signal (CLK) for the TDC was provided by a pulse pattern generator (PPG) (Anritsu, MP1763B). A Berkley Nucleonics BNC 745 Delay Generator (DG) was used to perform a linear delay sweep of the HIT signal relative to the CLK. The 50 Ω CLK and HIT signals from the PPG and DG were used as the START and STOP inputs of the TDC, respectively. The TDC chips digitized the time delay between the rising edges of CLK and HIT signals from which the TDC’s characteristics were obtained.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 6‑8: (a) Measurement set-up used for TDC characterization at 1 MHz sample rate. (b) Measured TDC input signals, CLK and HIT (left) and their respective cycle-to-cycle jitter (right).

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 6‑16: (a) Measurement set-up used for TDC characterization at 1 MHz sample rate. (b) Measured TDC input signals, CLK and HIT (left) and their respective cycle-to-cycle jitter (right).

Calibration measurements of the HIT and CLK signals acquired by the oscilloscope are shown in Fig. 6-8(b) and Fig. 6-9. The left panel of Fig. 6-8(b) displays the HIT and CLK waveforms with peak-to-peak amplitudes of 1.5 V. The right panel shows that the measured cycle-to-cycle rms jitter for CLK and HIT signals was 12.6 and 23.5 ps, respectively. Fig. 6-9(a) shows the histogram of measured delays between CLK and HIT obtained by sweeping the delay of the DG using a 1 ns increment. Although the rms jitter was less than 25 ps for each delay setting, there was a constant delay error of approximately 350 ps when the delays were greater than 5 ns (Fig. 6-9(b)). The effects of these offsets were removed in post processing of the measured TDC data, as shown in the next sub-section.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 6‑9: (a) Measured histogram of delays from set-up in Fig. 6-8(a). (b) Measured average delays show a deviation from the ideal behavior.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure 6‑18: (a) Measured histogram of delays from set-up in Fig. 6-8 (a). (b) Measured average delays show a deviation from the ideal behavior.

The maximum repetition rate of the DG was limited to 1 MHz. To obtain a higher sampling rate for the TDC, the DG was removed from the set-up. The TRIG signal from the PPG was used as the TDC’s STOP input to attain sampling at 60 MHz. On the PPG, the trigger (TRIG) signal corresponded to the CLK signal divided by 2. The delay of the trigger signal relative to CLK could be adjusted on the PPG in 1 ps increments. However, the maximum delay was only 1 ns and was not enough to characterize the entire TDC dynamic range. Thus, coaxial cables of different length were inserted between the TRIG output and TDC STOP input so that delays spanning the entire TDC dynamic range could be obtained. This set-up had approximately 2× lower jitter and 60× faster measurement rates, but the 1 MHz set-up using the DG could be fully automated, thus more chips could be measured to assess the TDC’s statistical variations.

### Transfer Characteristic

The TDC’s transfer characteristic includes all quantization and nonlinearity information required to assess the TDC’s performance. To obtain the TDC characteristic, the DG delay was swept over the TDC’s dynamic range in 10 ps increments, with 1000 samples measured by the oscilloscope at each input delay. In this manner, the effects of timing jitter on the TDC characteristic measurements could be averaged out. At each 10 ps delay step, the mean, representing the TDC’s accuracy, and standard deviation, representing its precision, were calculated.

Fig. 6-10(a) shows one sample of a measured transfer characteristic obtained at 1 MHz from the TDC as well as the transfer characteristic corresponding to an ideal 6-bit TDC with *TLSB* =130.2 ps. There are several imperfections of the TDC, as well as imperfections in the measurement set-up that affect the measured characteristic that will be discussed next.

|  |
| --- |
|  |

Figure 6‑10: Comparison between measured, corrected and ideal transfer characteristics of a 6-bit TDC. Figure inset shows an example of non-monoticitiy in the measured characteristic due to timing jitter. Measured TDC characteristics was obtained using a 120 MHz reference clock frequency corresponding to *TLSB* = 130.2 ps.

|  |
| --- |
|  |

Figure 6‑20: Comparison between measured, corrected and ideal transfer characteristics of a 6-bit TDC. Figure inset shows an example of non-monoticitiy in the measured characteristic due to timing jitter. Measured TDC characteristics was obtained using a 120 MHz reference clock frequency corresponding to *TLSB* = 130.2 ps.

i) The effects of timing jitter of the CLK and HIT signals as well as the jitter of the TDC cause the non-monotonic behavior seen in the measured characteristic (inset of Fig. 6-10(a)). The source jitter causes variations of the TDC output near the code transitions. The effect of jitter on the TDC characteristic was averaged out by repetitively measuring the same input time interval 1000 times and taking the average output code.

ii) The inaccuracy of the DG as the input delay increases beyond 4.5 ns (as shown in Fig. 6-9) caused the measured TDC characteristic to deviate considerably from its true characteristic. In order to correct for this error, 2 LSBs were subtracted from the measured TDC output code whenever the input delay exceeded 4.5 ns. The DG error of approximately 350 ps was accounted for in this manner. Fig. 6-9(b) shows the effect of DG inaccuracy on the measured characteristic data as well as the results of the correction.

iii) As the input delay increased above 7300 ps, the TDC is no longer responsive to changes of the input delay and the maximum output code was reached. This occurred because the flip-flops in the read-out circuits required approximately 1 ns of set-up and hold time to properly latch the output code bits. During this time the TDC was reset to ensure no flip-flop set-up and hold time were violated, which could lead to metastability and result in the arriving HIT signals not being registered. Also, because of the timing jitter, the TDC may also output codes near zero when the maximum TDC code is reached and HIT signals occur near the end of the timing interval.

iv) The two status bits of the TDC were used to trigger the oscilloscope and this caused variations within the first coarse TDC code. Based on the value of the input delay, the coarse status bit *N0C* indicate whether the HIT signal arrives before the first LO-HI transition in the coarse VCDL. If HIT arrives after, then a LO-HI transition of *N0C* is guaranteed in this case. When the HIT signal arrives before the first low-to-high (LO-HI) transition of the fine VCDL, then a LO-HI transition is guaranteed for fine status bit *N0F*. Thus the oscilloscope was programmed to switch the trigger source from *N0F* to *N0C* whenever the input delay exceeds 4 LSBs to ensure that it has a valid trigger for all input delays. In the ideal case, the oscilloscope would always have the proper trigger at each input delay setting. However, the trigger source was not always switched at the proper instant due to timing jitter of the instruments and nonlinearity of the TDC, and this caused variations in the measured output code when the input delay was near 4 LSB. The non-monotonicity of the transfer characteristic due to this effect was corrected during the post-processing

v) The first code transition of the TDC ideally occurs at the first LSB. However, this transition may be shifted in the measurements due to non-idealities such as timing jitter and non-linearity. The complete transfer characteristic is then shifted along the time axis as a result and the converter is said to have an offset error. The offset error causes an increase of the INL because INL is defined as the difference in time between the measured and ideal code transitions. However, as this portion of the INL is systematic and is common to all code transitions, it may be subtracted during post processing. The corrected data shown in Fig. 6-10(a) was also corrected for the offset error.

Fig. 6-11(a) shows the TDC characteristics of 10 measured chips and the ideal characteristic of a 6-bit TDC with *TLSB* = 130.2 ps. The measurements were obtained at a 1 MHz sample rate with a 120 MHz reference clock frequency. The corresponding quantization error (Fig. 6-11(b) top) was extracted from the characteristics by calculating difference between the measured TDC output and the input delay. The measured quantization error includes the all the error due to non-linearity, so the standard deviation of the measured quantization error corresponds to the rms time resolution of the TDC. The best and worst case rms resolutions were 0.48 and 0.86 LSB, for chip 8 and chip 6, respectively. Averaged over all chips, the rms resolution of the TDC was 0.61 LSB at room temperature.

|  |
| --- |
|  |
| (a) |
|  |
| (b) |

Figure 6‑11: Measured TDC characteristics of 10 chips using a 120 MHz reference clock frequency corresponding to *TLSB* = 130.2 ps. (a) Comparison between measured, corrected and ideal transfer characteristics of a 6-bit TDC. Figure inset shows an example of non-monoticitiy in the measured characteristic due to timing jitter. (b) Measured TDC quantization error (top) and associated histogram (bottom). The average rms resolution of the 10 measured chips was 0.61 LSB.

|  |
| --- |
|  |
| (a) |
|  |
| (b) |

Figure 6‑22: Measured TDC characteristics of 10 chips using a 120 MHz reference clock frequency corresponding to *TLSB* = 130.2 ps. (a) Comparison between measured, corrected and ideal transfer characteristics of a 6-bit TDC. Figure inset shows an example of non-monoticitiy in the measured characteristic due to timing jitter. (b) Measured TDC quantization error (top) and associated histogram (bottom). The average rms resolution of the 10 measured chips was 0.61 LSB.

### Non-linearity

The DNL and INL were extracted from the TDC characteristics in Fig. 6-11(a) by using eq. (2-2). The width of each TDC bin value was compared to the ideal *TLSB* value and the difference in time (in LSB units) between the measured code transitions and the ideal code transitions was calculated.

|  |
| --- |
|  |

Figure 6‑12: Top and bottom graphs are the DNL and INL, respectively, of 10 measured chips at room temperature.

|  |
| --- |
|  |

Figure 6‑24: Top and bottom graphs are DNL and INL, respectively, of 10 measured chips at room temperature.

The nonlinearity for 10 measured chips is shown in Fig. 6-12. DNL for all TDC codes for all the chips is within -1 and +1.15 LSB, and INL is within -0.84 LSB and +1 LSB. The DNL is evenly distributed about 0 LSB, implying low INL. The DNLs for all the measured chips have similar shape but the chip-to-chip variation of INL is considerable. Nevertheless, none of the measured chips had the occurrence of a missing code in the characteristic thereby ensuring the accuracy of the measurements to within 1 LSB across the entire dynamic range. The best and worst case rms INL values were 0.25 and 0.46 LSB, for chip 3 and chip 8, respectively. The rms DNL and INL values, averaged over all ten chips, were 0.4 and 0.3 LSB, respectively.

All the measured chips had similar systematic variation in DNL/INL, occurring with a period of 4 LSBs due to unequal matching of delays in the fine interpolator. Simulation results in Fig. 6-5(b), indicate that the delay between START2 and the first fine interpolator phase is slightly longer than the delays between the other 3 phases. This is due to the unavoidable mismatch in delay caused by the interconnection routing.

### Jitter

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 6‑13: TDC jitter measured at 1 MHz sample rate. (a) The TDC jitter across the entire measurement range was 0.385 LSB rms. (b) Close-up view of the jitter due to a coarse code transition. In this shorter time range, the rms jitter was 0.8 LSB.

|  |  |
| --- | --- |
|  | |
| (a) | (b) |

Figure 6‑26: TDC jitter measured at 1 MHz sample rate. (a) The TDC jitter across the entire measurement range was 0.385 LSB rms. (b) Close-up view of the jitter due to a coarse code transition. In this shorter time range, the rms jitter was 0.8 LSB.

Besides the non-linearity, the TDC measurement accuracy depends on the source jitter and on the jitter of the coarse and fine VCDLs. The standard deviation of the repetitive delay measurements used to obtain the TDC characteristics represents the timing jitter. The timing jitter for one TDC chip (chip 1) measured at 1 MHz sampling rate is shown in Fig. 6-13(a). The rms value of the jitter over all time intervals reveals the overall jitter performance of the TDC. The measured rms value of jitter across the entire range was 0.39 LSB, and the maximum and minimum jitter values were 2.1 and 0 LSB, respectively. The best jitter occurs when the input time interval occurs halfway between any two code transitions. In this case, all the measurements of the timing interval give the same result when the time interval is measured repeatedly, so the source and VCDL jitter has no effect on the TDC’s precision. Conversely, the worst jitter is for input time intervals that occur near a code transition. When this happens, a constant time interval measured repeatedly gives different results each time and degrades the precision.

|  |
| --- |
|  |

Figure 6‑14: TDC jitter measured at 60 MHz sample rate using 100 MHz reference clock.

The measured TDC jitter in Fig. 6-13 exhibits a periodic behavior where the jitter peaks at every 4th LSB corresponding to each coarse code transition. This behavior is on account of the different jitter behavior of the coarse and fine VCDLs. The coarse VCDL was designed to span a much longer range of delays for the same range of control voltages, so it had a much higher gain than the fine VCDL (Fig. 6-4). As a result, small variations on the control voltage or on the power supply will translate to greater variations of the delay for the coarse VCDL compared the fine VCDL. For the fine VCDL, the peak jitter was approximately 0.5 LSB, while for coarse VCDL the peak jitter was roughly 3× larger. The jitter histogram for the coarse VCDL in Fig. 6-13(b) shows a rms jitter of 0.8 LSB in this range.

Jitter performance of the TDC was further assessed at a 100 MHz sampling rate. In this case, the source jitter was approximately 2× lower since the delays were being generated by the PPG and coaxial cables instead of the DG. The measured jitter in Fig. 6-14 shows the same periodic behavior, but the peak rms jitter is reduced by about one half as expected from the calibration results in Fig. 6-8(b). The good jitter performance of the TDC was on account of the low DNL (< 1 LSB), that led to less variation of the time jitter for different input time delays. Also the short length of the VCDL led to less jitter accumulation along the delay line. The analysis reported in Figs. 6-13 and 6-14 confirms that the TDC jitter does not degrade the measurement accuracy and that the TDC LSB is equal to its resolution.

# Conclusion

The following sections provide a summary and some concluding remarks, as well as recommendations for future work.

## Summary and Discussion

In this work, single-photon avalanche diodes (SPAD) and time-to-digital converters (TDC) were developed for time-resolved fluorescence analysis in a low-cost, standard digital deep-submicron (DSM) complementary metal-oxide-semiconductor (CMOS) process. Although CMOS technology offers very high-speed, low power, and highly integrated digital circuits, it is mostly beneficial for high-performance TDC design. The realization of SPADs with high-sensitivity and wide dynamic range remains a significant challenge since standard CMOS technology is not optimized for single-photon imaging. However, besides the lower fabrication costs, the main driving factors of CMOS SPADs are unparalleled levels of miniaturization and portability, as well as potential for system-on-chip (SoC) integration as evidenced by the growth in the functionality and performance of time-resolved single-photon imaging applications over the years.

This research focused on the design, fabrication and characterization of SPAD pixels in a standard low-cost DSM CMOS process and the assessment their time-resolved single-photon detection capabilities. High-speed, compact and precise TDC prototype integrated circuits (ICs) with sub-nanosecond time resolution were developed as well. The SPAD pixels were designed to be are suitable for integration with multi-channel TDCs in order to realize a low-cost and fully miniaturized single-photon camera for biomedical imaging applications.

Chapter 2 reviewed the key system requirements and main measurement techniques used in single-photon imaging applications, particularly fluorescence lifetime imaging (FLIM). Some of the current competing technologies, such as PMTs and CCDs, were reviewed and their advantages and disadvantages were comparatively assessed. The principle of SPAD operation was introduced and an extensive literature review of CMOS SPADs was presented. The advantages of CMOS SPADs fabricated in DSM technology have been given particular focus and their main performance characteristics were studied and compared for different CMOS fabrication technologies. CMOS imaging technologies with dedicated features for optimum SPAD detector performance have been identified and contrasted with standard low-cost CMOS technology. For FLIM applications, the timing information of photons is required. A review of TDC concepts and circuit architectures was presented and the specifications of the TDC prototype ICs were outlined.

In Chapter 3, the design of SPADs in a 130 nm IBM CMOS process was presented featuring photosensitive areas larger than previously reported in the literature for this technology. Important technological considerations such as triple-well isolation, shallow-trench isolation (STI), silicide, and final chip passivation were considered for successful SPAD realization. The developments were mainly concerned with optimization of the SPAD structure (separating the STI from the active region and removing the silicide layer) which lead to a considerable improvement in the noise and the sensitivity performance.

Four different passively quenched SPAD pixels were characterized in this work. Unbuffered test structures were initially fabricated to measure the breakdown voltage characteristics of silicided and non-silicided SPADs, as well to investigate the voltage headroom limitations of SPADs in DSM CMOS. Measurements of output pulses from unbuffered test structures validated the SPAD circuit model used for simulations. An SPAD pixel with source-follower front-end was then designed in order to minimize the capacitive loading effects and to study the behavior of output pulses at different temperatures and bias voltages. However, this pixel had a high power consumption and large timing jitter, so a low-power and low-jitter front-end circuit based on a common-source amplifier was realized. This pixel structure was extensively characterized. Analysis of the dark counts and afterpulses as a function of temperature and excess voltage revealed important information about the noise generation mechanisms of SPADs in CMOS technology.

To circumvent the limitations of dark noise and afterpulsing, the time-gated mode of SPAD operation has been exploited in Chapter 4, and encouraging results have been reported in comparison to the free-running mode. Analysis based on the measured inter-arrival time data was performed to fully characterize the afterpulsing, enabling optimal operating temperature and excess voltage conditions to be found. The results of the afterpulsing measurements suggest that the time-gated mode of operation is essential for optimal performance when cooled to minimize the thermal noise. It was shown that time-gated operation at -30 °C could reduce the probability of a dark count occurring within the time gate to be as low as 10-4 % and requiring a hold-off time of 160 ns to completely eliminate the afterpulsing. In contrast, a hold-off time in the microsecond range was required to minimize afterpulsing in the free-running mode for the same excess voltage and temperature.

The characterization of the SPAD to low-level light and high-intensity light was performed in Chapter 5. While operating the SPAD in the time-gated mode was beneficial in reducing the dark counts, the sensitivity was shown to be considerably reduced. Although removal of the silicide layer from the SPAD’s photosensitive area benefits the detection efficiency, the consequence of fabricating SPADs in a non-imaging CMOS process meant that the peak detection efficiency was fundamentally limited by the technology. Nevertheless, the timing performance attained by the SPAD pixels with common-source front end was comparable to a commercially available device. The good SPAD timing jitter and improved PDE and DCR performance of the non-silicided pixels in this standard CMOS technology opens up the potential for their development as low-cost instruments for fluorescence lifetime analysis.

In Chapter 5, the SPAD’s performance in fluorescence lifetime analysis was demonstrated. The fabricated SPAD pixels were used to measure the lifetimes of two fluorescent calibration standards, Rhodamine 6G (R6G) and ruby crystal. R6G was chosen because its peak emission wavelength coincided with the SPAD’s maximum PDE wavelength and its nanosecond scale lifetime was suitable to demonstrate the SPAD’s time-resolved capability. The ruby crystal was chosen since it emits fluoresces in the near-infrared (NIR) spectrum and could thus demonstrate the SPAD’s capability in detecting NIR light, which is very important for biomedical imaging. The time-correlated single-photon counting (TCSPC) technique was utilized to reconstruct the lifetime of R6G over a nanosecond time scale and the lifetimes of R6G could be identified for concentrations down to 10-5 M. For the ruby lifetime measurements, very low-cost instruments were used to assess the SPAD’s performance in the detection of fluorescence on a millisecond time scale. The time-resolved fluorescence measurements of samples with well-known fluorescence lifetimes validate the applicability of the standard CMOS SPADs prototype chips that were fabricated in this work.

The final phase of the research continued by developing TDC circuits suitable for SPAD applications. The proposed 6-bit coarse-fine interpolating TDC utilized a differential amplifier as a delay line for both coarse and fine resolution. A synchronizer circuit using static CMOS logic gates was utilized to simplify synchronization between coarse and fine interpolators and to reduce the trade-offs between time resolution, linearity, sample rate and circuit size. The TDC prototypes were characterized by their ability to digitize input delay times with sub-nanosecond resolution and accuracy. The TDC required only 20 delay elements and 0.04 mm2 of silicon area for 130.2 ps least-significant-bit (LSB) resolution, < 1 LSB integral non-linearity (INL) and 60 MHz maximum sampling rate. Two delay-locked-loop (DLL) circuits ensured high linearity and good jitter performance and reduced process-voltage-temperature (PVT) sensitivity.

## Recommendations for Future Work

The SPADs and TDCs developed in this work were targeted for biomedical applications such as FLIM that require low-cost, low-power, miniaturized sensors that are able to detect light down to the single-photon level with sub-nanosecond timing resolution. The starting point for the development of such sensors was represented by the design and characterization of single-pixel SPADs, investigation of their performance for time-resolved fluorescence analysis, and the realization of high-performance single-channel TDC prototype ICs. Future developments are oriented towards full integration of SPAD and TDC on a single chip. However, there are improvements to the SPAD pixels and TDC prototype that can be made for the realization of fully integrated multi-channel, compact and low-cost CMOS sensors with time-resolved single-photon detection capabilities.

First, although the improved PDE together with the reduced DCR of the non-silicided pixels opened up the capability of using standard CMOS SPADs for high-sensitivity fluorescence lifetime imaging applications, more improvements in PDE and DCR performance are required in order to attain the levels of sensitivity that was demonstrated for the commercial SPAD detectors. Deep cooling past -30 °C may have additional benefits to reducing the DCR in the free-running mode if the afterpulsing probability can be correspondingly minimized. Optimized AQR front-end circuits are therefore required for the free-running SPAD pixels in the pixel to precisely control the hold-off time so that afterpulsing is reduced at lower temperatures. Operation of the SPAD’s in time gated mode was shown to be beneficial at low temperatures. However, operating the SPAD in time-gated mode involves a trade-off of the detection efficiency. Therefore, future work is needed to develop an array of time-gated SPAD pixels that operate in parallel so that the detection efficiency can be improved in the time-gated mode.

Second, for the SPAD pixel design, more research is required to more accurately model the dynamic behavior of the SPAD by considering the voltage dependence of the pn junction capacitance as well as the parasitic pn junction capacitance associated with the SPAD guard rings. SPAD models that take into account the statistical effects such as the avalanche triggering probability, the DCR and the afterpulsing phenomena and their dependence on temperature and excess voltage should be developed and implemented using a hardware description language (HDL) such as Verilog-A, which has the advantage that it can be used in many commercial circuit simulators.

Third, a multichannel TDC implementation is required in order to obtain a sample rate that can equal, or exceed, the reference clock rate. Improvements need to be made to reduce the core TDC’s power consumption to enable multi-channel operation. Immunity to electrical crosstalk interference from high-speed I/O pad drivers distributed throughout the chip and resulting simultaneous switching noise (SSN) should be also improved. Otherwise, jitter and linearity performance can be jeopardized when a large number of TDCs are integrated together on the same chip. Also, more research should be focused on improving the resolution of the TDC down to the picosecond range without incurring a reduction in sample rate or an increase in either the power consumption or circuit area.

Finally, another issue concerning the employment of SPAD/TDC arrays for multi-channel TCSPC operating at high sample rates is the challenge of processing the large amount of data generated by the TDCs. The flash architecture and simple synchronization scheme resulted in a 60 MHz sample rate, which corresponds to a data rate of 480 Mbps, for the designed TDC. Directly transferring the TDC conversions to an external computer to obtain the histograms represents a significant bottle-neck for real-time determination of fluorescence decays. Therefore, more work is required to implement on-chip histogram techniques that can significantly reduce the data read-out rate.

# References

1. J. R. Lakowicz, *Principles of Fluorescence Spectroscopy, 3rd Ed.*, Springer, New York, 2006.
2. B. Valeur, *Molecular Fluorescence – Principles and Applications,* Wiley-VCH, New York, 2002.
3. L. Marcu, et al., (Eds.) *Fluorescence Lifetime Spectroscopy and Imaging – Principles and Applications in Biomedical Diagnosics,* CRC Press, Boca Raton, 2015.
4. L. Marcu, B. A. Hartl, “Fluorescence Lifetime Spectroscopy and Imaging in Neurosurgery”, *IEEE J. Sel. Top. Quant. Elec.*, vol. 18, no.4, pp. 1465-1477, 2012.
5. Y. Sun, et al., “Endoscopic Fluorescence Lifetime Imaging for In Vivo Intraoperative Diagnosis of Oral Carcinoma”, *Microscopy and Microanalysis,* vol. 19, no. 4, pp. 791-798, 2013.
6. N. P. Galletly, et al., “Fluorescence lifetime imaging distinguishes basal cell carcinoma from surrounding uninvolved skin”, *British Journal of Dermatology*, vol. 159, pp. 152-161, 2008.
7. J. Siegel, et al., “Studying biological tissue with fluorescence lifetime imaging: microscopy, endoscopy, and complex decay profiles”, *Appl. Opt.*, vol. 42, no. 16, pp. 2995-3004, 2003.
8. R. Cubeddu, et al., “Time resolved fluorescence imaging in biology and medicine,” *J. Phys. D., Appl. Phys.*, vol. 35, pp. 61–76, 2002.
9. K. Suhling, et al., “Time-resolved fluorescence microscopy,” *Photochem. Photobiol. Sci.*, vol. 12, pp. 12–22, 2005.
10. J. W. Borst and J. W. G. Visser, “Fluorescence lifetime imaging microscopy in life sciences,” *Meas. Sci. Tech.*, vol. 21, pp. 102002, 2010.
11. M. Y. Berezin, S. Achilefu, “Fluorescence Lifetime Measurements and Biological Imaging”, *Chem. Rev.*, vol. 110, no. 5, pp. 2641-2684, 2010.
12. D. K. Bird, et al., “Time-resolved fluorescence microscopy of gunshot residue: an application to forensic science”, *J. of Microscopy*, vol. 226, no. 1, pp. 18-25, 2006.
13. A. Ehn, et al., “Fluorescence lifetime imaging in a flame”, *Proceedings of the Combustion Institute*, vol. 33, no. 1, pp. 807-813, 2011.
14. R. K. P. Benninger, et al., “Time-resolved fluorescence imaging of solvent interactions in microfluidic devices”, *Opt. Exp.*, vol. 13, no. 16, pp. 6275-6285, 2005.
15. E. M. Graham, et al., “Quantitative mapping of aqueous microfluidic temperature with sub-degree resolution using fluorescence lifetime imaging microscopy”, *Lab on Chip*, vol. 10, pp. 1267-1273, 2010.
16. D. Comelli, et al., "Fluorescence lifetime imaging and spectroscopy as tools for nondestructive analysis of works of art," *Appl.Opt.*, vol. 43, no. 10, pp. 2175-2183, 2004.
17. X. F. Wang, et al., “Fluorescence Lifetime Imaging Microscopy (FLIM): Instrumentation and Applications”, *Critical Reviews in Analytical Chemistry*, vol. 23, no. 5, pp. 369-395, 1992.
18. P. Herman, J. R. Lakowicz, “Lifetime-based imaging” in H. J. Lin (Ed.) *Biomedical Photonics Handbook, 2nd Ed.: Fundamentals, Devices, and Techniques,* CRC Press, New York, 2014.
19. W. Becker, “Fluorescence lifetime imaging – techniques and applications”, *Journal of Microscopy*, vol. 247, no. 2, pp. 119-136, 2012.
20. N.Boens, et al., “Fluorescence Lifetime Standards for Time and Frequency Domain Fluorescence Spectroscopy”, *Anal. Chem*., vol. 79, pp. 2137-2149, 2007.
21. H. C. Gerritsen, et al., “Fluorescence Lifetime Imaging in Scanning Microscopy”, in J. B. Pawley, *Handbook of Biological Confocal Microscopy, 3rd Ed.*, Springer, New York, 2006.
22. E. Gratton, et al., “Fluorescence lifetime imaging for the two-photon microscope: time domain and frequency domain methods”, *J. of Biom. Opt.*, vol. 8, no. 3, pp. 381-390, 2003.
23. J. Philip, K. Carlsson, “Theoretical investigation of the signal-to-noise ratio in fluorescence lifetime imaging”, *Journal of the Optical Society of America A*, vol. 20, no. 2, pp. 368-378, 2003.
24. D. V. O’Connor, D. Phillips, *Time-correlated Single Photon Counting*, Academic Press, London, 1984.
25. D. J. S. Birch, R. E. Imhof, “Time-Domain Fluorescence Spectroscopy Using Time-Correlated Single-Photon Counting”, in J. R. Lakowicz (Ed.), *Topics in Fluorescence Spectroscopy, Vol. 1: Techniques*, Plenum Press, New York, 1991.
26. W. Becker, *Advanced Time-Correlated Single Photon Counting Techniques*, Springer Verlag, New York, 2005.
27. P. Seitz, A. J. P. Theuwissen, (Eds.) *Single-Photon Imaging*, Springer, New York, 2011.
28. W. Becker, *The bh TCSPC Handbook, 6th Ed.*, Becker & Hickl GmbH, 2014.
29. P. Kapusta, et al., (Eds.), *Advanced Photon Counting – Applications, Methods, Instrumentation*, Springer, New York, 2015.
30. J. McGinty, et al., "Wide-field fluorescence lifetime imaging of cancer." *Biom. Opt. Exp.*, vol. 1, no. 2, pp. 627-640, 2010.
31. V. Katsoulidou, et al., “How fast can TCSPC FLIM be made*?”, Proc. SPIE*, vol. 6771, pp. 67710B-1–7, 2007.
32. W. Becker, et al., “Picosecond Fluorescence Lifetime Microscopy by TCSPC Imaging”, *Proc. SPIE*, vol. 4262, pp. 414-419, 2001.
33. W. Becker, et al., “Fluorescence Lifetime Imaging by Time-Correlated Single-Photon Counting”, *Microscopy Research and Technique*, vol. 63, pp. 58-66, 2004.
34. W. Becker, et al., “Spatially Resolved Recording of Transient Fluorescence-Lifetime Effects by Line-Scanning TCSPC”, *Proc. SPIE*, vol. 8226, 6 pp., 2012.
35. H. Studier, et al., “Megapixel FLIM”, *Proc. SPIE*, vol. 8948, pp. 89481K1-9, 2014.
36. M. Wahi, et al., “Hardware solution for continuous time resolved burst detection of single molecules in flow”, *Proc. SPIE*, vol. 3259, pp. 173-178, 1998.
37. S. Felekyan, et al., “Full correlation from picoseconds to seconds by time-resolved and time-correlated single photon detection”, *Rev. Sci. Instrum.*, vol. 76, pp. 083104, 2005.
38. M. Wahl, et al., “Fast calculation of fluorescence correlation data with asynchronous time-correlated single-photon counting” *Opt. Exp.*, vol. 11, no. 26, pp. 3583-3591, 2003.
39. Q. Li, S. Seeger, “Autofluorescence Detection in Analytical Chemistry and Biochemistry”, *Applied Spectroscopy Reviews*, vol. 45, no. 1, pp. 12-43, 2010.
40. D. V. O’Connor, W. R. Ware, “Deconvolution of Fluorescence Decay Curves. A Critical Comparison of Techniques”, *J. of Physical Chemistry*, vol. 83, no. 10, pp. 1333-1343, 1979.
41. J. Arlt, et al., “A study of pile-up in integrated time-correlated single photon counting systems”, *Rev. Sci. Instr*., vol. 84, pp. 103105, 2013.
42. D.-U. Li, et al., "FPGA implementation of a video-rate fluorescence lifetime imaging system with a 32×32 CMOS single-photon avalanche diode array," *IEEE Int. Symp. Circ. Sys.*, pp. 3082- 3085, 2009.
43. C. Veerappan, et al., "A 160×128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter," *IEEE Int. Sol.-St. Cir. Conf. Dig. Tech. Papers (ISSCC)*, pp.312-314, 2011.
44. M. Gersbach, et al., "A time-resolved, low-noise single-photon image sensor fabricated in deep-submicron CMOS technology," *IEEE J. Sol. St. Circ.*, vol.47, no. 6, pp. 1394-1407, 2012.
45. N. Krstajic, et al., “Improving TCSPC data acquisition from CMOS SPAD arrays,” *Proc. SPIE*, vol. 8797, 879709, 8 pp., 2013.
46. D. Tamborini, et al., "Time-resolved optical spectrometer based on a monolithic array of high-precision TDCs and SPADs." *Proc. SPIE*, vol. 8993, pp. 89932I-89932I, 2013.
47. R. M. Field, et al., “A 100 fps, time-correlated single-photon counting-based fluorescence-lifetime imager in 130 nm CMOS,” *IEEE J. Sol. St. Circ.*, vol. 49, no. 4, pp. 867-880, 2014.
48. N. Krstajić, et al., "256 × 2 SPAD line sensor for time resolved fluorescence spectroscopy," *Opt. Exp.*, vol. 23, pg. 5653-5669, 2015.
49. D. E. Schwartz, et al., “A Single-Photon Avalanche Diode Array for Fluorescence Lifetime Imaging Microscopy”, *IEEE J. Sol.-St. Circ.,* vol. 43, no. 11, pp. 2546-2557, 2008.
50. D. Stoppa, et al., "Single-photon avalanche diode CMOS sensor for time-resolved fluorescence measurements," *IEEE Sensors J.*, vol.9, no.9, pp.1084-1090, 2009.
51. Y. Maruyama, E. Charbon, "An all-digital, time-gated 128×128 spad array for on-chip, filter-less fluorescence detection,” *16th Int. Sol-St Sensors, Actuators & Microsys. Conf.*, pp.1180-1183, 2011.
52. G. Boso, et al., “Fast-gated single-photon detection module with 200 ps transitions running up to 50 MHz with 30 ps resolution,” *Proc. SPIE*, vol. 8268, pp.82681U1-6, 2012.
53. Z. Li, M. J. Deen, “Towards a portable Raman spectrometer using a concave grating and a time-gated CMOS SPAD,” *Opt. Exp.*, vol. 22, no. 15, pp. 18736–18747, 2014.
54. Y. Maruyama, et al., “A 1024 x 8, 700-ps time-gated SPAD line sensor for planetary surface exploration with laser Raman spectroscopy and LIBS”, *IEEE J. Sol.-St. Circ.*, vol.49, no.1, pp. 179-189, 2014.
55. S. Burri, et al., "Architecture and applications of a high resolution gated SPAD image sensor." *Opt. Exp.*, vol. 22, no. 14, pp. 17573-17589, 2014.
56. I. Nissinen, et al., “A 2×(4)×128 Multitime-Gated SPAD Line Detector for Pulsed Raman Spectroscopy”, *IEEE Sens. J.*, vol.15, no. 3, pp.1358-1365, 2015.
57. M. Benetti, et al., "Highly parallel SPAD detector for time-resolved lab-on-chip." *Proc. SPIE*, vol. 7723, pp. 77231Q-77231Q, 2010.
58. B. R. Rae, et al., “A vertically integrated CMOS microsystem for time-resolved fluorescence analysis”, *IEEE Trans. Biomed. Circ. and Sys.*, vol. 4, no. 6, pp. 437 – 444, 2010.
59. N. A. W. Dutton, et al., "9.8 µm SPAD-based Analogue Single Photon Counting Pixel with Bias Controlled Sensitivity”. *International Image Sensor Workshop,* 2013.
60. L. Pancheri, et al., “SPAD Image Sensor With Analog Counting Pixel for Time-Resolved Fluorescence Detection”, *IEEE Trans. Elect. Dev.*, vol. 60, no. 10, pp. 3442-3449, 2013.
61. D. P. Palubiak, M. J. Deen, "CMOS SPADs: Design Issues and Research Challenges for Detectors, Circuits, and Arrays," *IEEE J. Sel. Top. Quant. Elect.*, vol.20, no.6, pp.409-426, 2014.
62. E. Charbon, "Single-photon imaging in complementary metal oxide semiconductor processes." *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 372, no. 2012, pp. 20130100, 2014.
63. J. R. Meijlink, et al., “First measurement of scintillation photon arrival statistics using a high-granularity solid-state photosensor enabling time-stamping of up to 20,480 single photons,” *IEEE Nucl. Sci. Symp. Med. Imag. Conf. (NSS/MIC)*, pp.2254-2257, 2011
64. L. H. C. Braga, et al., "A fully digital 8 x 16 SiPM array for PET applications with per-pixel TDCs and real-time energy output," *IEEE J. Sol. St. Circ.*, vol.49, no.1, pp.301-314, 2014.
65. C. Bruschini et al., “SPADnet: a fully digital, scalable and networked photonic component for time-of-flight PET applications,” *Proc. SPIE*, vol. 9129, 912913, 11 pp. 2014.
66. G. S. Buller, A. M. Wallace, "Ranging and Three-Dimensional Imaging Using Time-Correlated Single-Photon Counting and Point-by-Point Acquisition," *IEEE J. Sel. Top. Quant. Elect.*, vol.13, no.4, pp.1006-1015, 2007
67. A. McCarthy, et al., "Long-range time-of-flight scanning sensor based on high-speed time-correlated single-photon counting," *Appl. Opt.*, vol. 48, no. 32, pp. 6241-6251, 2009.
68. R. Shu, et al., “Multi-channel photon counting three-dimensional imaging laser radar system using fiber array coupled Geiger-mode avalanche photodiode,” *Proc. SPIE*, vol. 8542, pp. 85420C-1–85420C-9, 2012.
69. C. Niclass, et al., "A 128 x 128 single-photon image sensor with column-level 10-bit time-to-digital converter Array," *IEEE J. Sol. St. Circ.*, vol.43, no.12, pp. 2977-2989, 2008.
70. C. Niclass, et al., “A 0.18-μm CMOS SoC for a 100-m-Range 10-Frame/s 200x96-Pixel Time-of-Flight Depth Sensor”, *IEEE J. Sol. St. Circ.*, vol. 49, no. 1, pp. 315 – 329, 2014.
71. D. Bronzi, et al. "100 000 Frames/s 64× 32 Single-Photon Detector Array for 2-D Imaging and 3-D Ranging." *IEEE J. Sel. Top. Quant. Elect.*, vol. 20, no. 6, pp. 354-363, 2014.
72. I. Vornicu, et al., "A SPAD-based 3D imager with in-pixel TDC for 145ps-accuracy ToF measurement." *Proc. SPIE*, vol. 9403, pp. 94030I-94030I, 2015.
73. D. Bronzi, et al., “SPADAS: a high-speed 3D Single-Photon camera for Advanced Driver Assistance Systems”, *Proc. SPIE*, vol. 9366, pp. 93660M-1–93660M-7, 2015.
74. H. C. Gerritsen, et al., “Fluorescence Lifetime Imaging of Oxygen in Living Cells”, *J. of Fluorescence*, vol. 7, no. 1, pp. 11-15, 1997.
75. J. Sytsma, et al., “Time-gated fluorescence lifetime imaging and microvolume spectroscopy using two-photon excitation”, *J. of Microscopy*, vol. 191, pt. 1, pp. 39-51, 1998.
76. K. Dowling, et al., “Fluorescence lifetime imaging with picosecond resolution for biomedical applications,” *Opt. Lett.*, vol. 23, no. 10, pp. 810–812, 1998.
77. K. K. Sharman, et al., “Error Analysis of the Rapid Lifetime Determination Method for Double-Exponential Decays and New Windowing Schemes”, *Anal. Chem.*, vol. 71, pp. 947-952, 1999.
78. D. D.-U. Li, et al., “Time-Domain Fluorescence Lifetime Imaging Techniques Suitable for Solid-State Imaging Sensor Arrays”, *Sensors*, vol. 12, pp. 5650-5669, 2012.
79. H. C. Gerritsen, et al., “Fluorescence lifetime imaging in scanning microscopies: acquisition speed, photon economy and lifetime resolution”, *J. of Microscopy*, vol. 206, no. 3, pp. 218-224, 2002.
80. A V Agronskaia, et al., “High frame rate fluorescence lifetime imaging”, *J. Phys. D: Appl. Phys.*, vol. 36, pp. 1655-1662, 2003.
81. A. V. Agronskaia, et al., “Fast fluorescence lifetime imaging of calcium in living cells”, *J. of Biomedical Optics*, vol. 9, no. 6, pp. 1230-1237, 2004.
82. J. McGinty, et al., “Signal-to-noise characterization of time-gated intensifiers used for wide-field time-domain FLIM”, *J. Phys. D: Appl. Phys.*, vol. 42, 135103, 9 pp., 2009.
83. M. M. El-Desouki, et al., “A Novel, High Dynamic Range, High Speed and High Sensitivity CMOS Imager Using Time-Domain Single Photon Counting and Avalanche Photodiodes,” *IEEE Sensors J*., vol. 11, pp. 1078 – 1083, 2011.
84. D. Palubiak, et al., "High-Speed, Single-Photon Avalanche-Photodiode Imager for Biomedical Applications," *IEEE Sensors J.*, vol.11, no.10, pp.2401-2412, 2011.
85. N. Faramarzpour, et al., "Fully integrated single photon avalanche diode detector in standard CMOS 0.18-μm technology," *IEEE Trans. Elec. Dev.*, vol.55, no.3, pp. 760-767, 2008.
86. D. P. Palubiak, et al., "Afterpulsing Characteristics of Free-Running and Time-Gated Single-Photon Avalanche Diodes in 130-nm CMOS," *IEEE Trans. Elec. Dev.*, vol. 62, no. 11, pp. 3727-3733, 2015.
87. M. Ghioni, et al., “Progress in silicon single-photon avalanche diodes,” *IEEE J. Sel. Top. Quant. Electron.*, vol. 13, no. 4, pp. 852-862, 2007.
88. X. Michalet, et al., “Detectors for single-molecule fluorescence imaging and spectroscopy”, *J. Mod. Opt.*, vol. 54, no. 2-3, pp. 239 – 281, 2007.
89. N. Faramarzpour, et al., “CMOS imaging for biomedical applications,” *IEEE Potentials*, vol. 27, no. 3, pp. 31–36, 2008.
90. D. Renker, E. Lorenz, “Advances in solid state photon detectors,” *J. Instrum.*, vol. 4, no. 4, pp. P04004, 2009.
91. X. Michalet, et al., “Development of new photon-counting detectors for single-molecule fluorescence microsocopy”, *Phil. Trans. R. Soc. B.*, vol. 368, pp. 20120035, 2012.
92. S. Donati, T. Tambosso, “Single-Photon Detectors: From Traditional PMT to Solid-State SPAD-Based Technology”, *IEEE J. Sel. Top. Quant. Elect.*, vol. 20, no. 6, pp. 3805008, 2014.
93. S. Weiss, “Fluorescence Spectroscopy of Single Biomolecules”, *Science*, Vol. 283, no. 5408, pp. 1676–1683, 1999
94. W. E. Moerner, D. P. Fromm, “Methods of single-molecule fluorescence spectroscopy and microscopy”, *Rev. Sci. Instrum.*, vol. 74, no. 8, pp. 3597–3619, 2003.
95. C. Joo, et al., “Advances in Single-Molecule Fluorescence Methods for Molecular Biology”, *Annual Review of Biochemistry*, vol. 77, no. 51-76, pp. 51–76, 2008
96. L. Alaverdian et al., “A family of novel DNA sequencing instruments based on single-photon detection”, *Electropheresis*, vol 23, pp. 2804-2817, 2002.
97. I. Rech, et al., “Microchips and single-photon avalanche diodes for DNA separation with high sensitivity”, *Electrophoresis*, vol. 27, no. 19, pp. 3797-3804, 2006
98. Y. Hämisch, “Molecular imaging with PET: New insights into the molecular basis of health and disease,” *MedicaMundi*, vol. 71, no. 1, pp. 18–27, 2003.
99. W. W. Moses, “Recent advances and future advances in time-of-flight PET,” *Nucl. Inst. Methods Phys. Res. A*, vol. 580, pp. 919–924, 2007.
100. P. Zanzonico, “Principles of nuclear medicine imaging: Planar, SPECT, PET, multi-modality, and autoradiography systems,” *Radiation Res.*, vol. 177, pp. 349–364, 2012.
101. M. Höbel, J. Ricka, “Deadtime and afterpulsing correction in multiphoton timing with nonideal detectors,” *Rev. Sci. Instr.*, vol. 62, no. 7, pp. 2326,2336, 1994.
102. D. Chitnis, S. Collins, “A SPAD-Based Photon Detecting System for Optical Communications”, *J. of Lightwave Technology*, vol.32, no.10, pp.2028-2034, 2014.
103. E. B. Hanlon, et al., “Prospects for in vivo Raman spectroscopy,” *Phys. Med. Biol.*, vol. 45, pp. 1–59, 2000.
104. A. Tripathi, et al., “Detection and identification of a water mixture of E. coli cells and B. subtilis spores with Raman chemical imaging microscopy,” *Proc. SPIE*, vol. 6554, pp. 65540J-1–65540J-9, 2007.
105. Z. Li, et al. "Raman Spectroscopy for In-Line Water Quality Monitoring—Instrumentation and Potential." *Sensors*, vol. 14, no. 9, pp. 17275-17303, 2014.
106. G. F. Knoll, *Radiation Detection and Measurement, 4th Ed.*, Wiley, New Jersey, 2010.
107. P. Buzhan, et al., “Silicon photomultiplier and its possible applications”, *Nucl. Instrum. and Meth. Phys. Res. Sec. A*, vol. 504, 48-52, 2003.
108. V. Golovin, V. Saveliev, “Novel type of avalanche photodetector with Geiger mode operation”, *Nucl. Instrum. and Meth. in Phys. Res. Sec. A*, vol. 518, no. 1–2, pp. 560-564, 2004.
109. R. Pestotnik, et al., “Silicon photomultiplier as a detector of Cherenkov photons”, *IEEE Nucl. Sci. Symp. Conf. Record*, vol.3, pp.2093-2096, 2007.
110. K. Tada, "Einstein’s Photon Hypothesis and Its Impact on Science and Technology," *AAPPS Bulletin*, vol. 15, no. 2, pp. 32-38, 2005.
111. *Photomultiplier Tubes – Basics and Applications, 3rd Ed.*, Hamamatsu Photonics K. K., 2006.
112. M.Ito, et al., “Computer Analysis of the timing properties in micro channel plate photomultiplier tubes”, *IEEE Trans. Nucl. Sci.*, vol. NS-31, no. 1, pp. 408-412, 1984.
113. C. D. Mackay, “Charge-Coupled Devices in Astronomy”, *Ann. Rev. Astron. Astrophys*., vol. 24, pp. 255-283, 1986.
114. J. R. Janesick, Scientific charge-coupled devices, *SPIE Press*, Bellingham, WA, 2001.
115. S. M. Sze, *Physics of Semiconductor Devices, 2nd Ed.*, Wiley-Interscience New York, NY, USA, 1981.
116. E. A. Gutierrez-D., et al., *Low Temperature Electronics – Physics, Devices, Circuits, and Applications*, Academic Press – New York, 2001.
117. P. Vu, et al., “Wafer-scale scientific CCDs at Fairchild Semiconductor”, *Proc. SPIE*, Vol. 5499, pp. 250-257, 2004.
118. O. Daigle, et al., “L3CCD results in pure photon counting mode”, *Proc. SPIE*, vol. 5499, pp. 219-227, 2004.
119. O. Djazovski, et al., “Electron-Multiplying CCDs for Future Space Instruments”, *Proc. SPIE*, vol. 8915, no. 8915Q, 13 pp., 2013.
120. M. S. Robbins, B. J. Hadwen, "The noise performance of electron multiplying charge-coupled devices," *IEEE Trans. Elect. Dev.*, vol.50, no.5, pp.1227-1232, 2003.
121. J. Hynecek, N. Nishiwaki, “Excess noise and other important characteristics of low light level imaging using charge multiplying CCDs,” *IEEE Trans. Elect. Dev.*, vol. 50, no. 1, pp. 239–245, 2003.
122. T. Isoshima, et al., “Ultrahigh sensitivity single-photon detector using a Si avalanche photodiode for the measurement of ultraweak biochemiluminescence”, *Rev. Sci. Instr.*, vol. 66, no. 4, pp. 2922-2926, 1995.
123. E. B. Johnson, et al., “Characteristics of CMOS Avalanche Photodiodes at Cryogenic Temperatures,” *IEEE Nucl. Sci. Symp. Conf. Rec.*, vol. N36, no. 1, pp. 2108-2114, 2009.
124. H. Dautet, et al., "Photon counting techniques with silicon avalanche photodiodes," *Appl. Opt*. vol. 32, no. 21, pp. 3894-3900, 1993.
125. S. Cova, et al., “Evolution and prospects for single-photon avalanche diodes and quenching circuits”, *J. Mod. Opt.*, vol. 51, no. 9-10, pp. 1267-1288, 2004.
126. F. Zappa, et al., “Principles and features of single-photon avalanche diode arrays”, *Sensors and Actuators A*, vol. 140, pp. 103-112, 2007
127. W. Mönch, “On the Physics of Avalanche Breakdown in Semiconductors”, *Physica Status Solidi (b)*, vol. 36 no.1, pp. 9-48, 1969.
128. R. J. McIntyre, “On the Avalanche Initiation Probability of Avalanche Diodes Above the Breakdown Voltage”, *IEEE Trans. Elect. Dev.*, vol. 20, no. 7, pp. 637-641, 1973.
129. Y. Taur, T. H. Ning, *Fundamentals of Modern VLSI Devices, 2nd Ed.*, Cambridge University Press: New York, 2009.
130. A. Spinelli, A. L. Lacaita, "Physics and numerical simulation of single photon avalanche diodes," *IEEE Trans. Elect. Dev.*, vol.44, no.11, pp.1931-1943, 1997.
131. J. S. Ng, et al., “Simulations of avalanche breakdown statistics: probability and timing”, *Proc. SPIE*, vol. 7681, pp. 76810K1-8, 2010.
132. I. Rech, et al., “Photon-timing detector module for single-molecule spectroscopy with 60-ps resolution,” *IEEE J. Sel. Top. Quant. Electron.*, vol. 10, no. 4, pp. 788-795, 2004.
133. N. Bertone, A. Giudice, “Developments in single photon avalanche photodiodes with fast timing resolution,” *Proc. SPIE*, vol. 6119, 611907, 8 pp., 2006.
134. S. Tisa, et al., “Electronics for single photon avalanche diode arrays”, *Sensors and Actuators A*, vol. 140, pp. 113-122, 2007.
135. S. Cova, et al., “Avalanche photodiodes and quenching circuits for single-photon detection”, *Appl. Opt.*, vol. 35, no. 12, pp. 1956-1976, 1996.
136. R. Mita, G. Palumbo, "High-Speed and Compact Quenching Circuit for Single-Photon Avalanche Diodes," *IEEE Trans. Instrum. and Meas.*, vol.57, no.3, pp.543-547, 2008.
137. A. Gallivanoni, et al., "Progress in Quenching Circuits for Single Photon Avalanche Diodes," *IEEE Trans. Nucl. Sci.*, vol.57, no.6, pp.3815-3826, 2010.
138. A. Eisele, et al., “185 MHz Count Rate, 139 dB dynamic range SPAD with active quenching circuit in 130 nm CMOS technology”, *Proc. Int. Image Sensor Workshop (IISW)*, R43, 2011.
139. A. Lacaita, et al., “Double epitaxy improves single-photon avalanche diode performance,” *Electron. Lett.*, vol. 25, pp. 841–843, 1989.
140. G. Kell, et al., “τ-SPAD: A new red sensitive single photon counting module”, *Proc. SPIE*, vol. 8033, 803303, 8 pp., 2011.
141. A. Giudice, et al., “High-detection efficiency and picoseconds timing compact detector modules with red-enhanced SPADs”, *Proc. SPIE*, vol. 8735, 83750P, 8 pp., 2012.
142. P. Finocchiaro, et al, “SPAD arrays and micro-optics: towards a real single photon spectrometer,” *J. Mod. Opt.*, vol. 54, pp. 119-212, 2007.
143. I. Rech, et al, “High-performance SPAD array detectors for parallel photon timing applications”, *Proc. SPIE*, vol. 8155, 12 pp., 2011.
144. S. Antonioli, et al., “Ultra-compact 32-channel system for time-correlated single-photon counting measurements”, *Proc. SPIE*, vol. 8773, 87730D, 11 pp., 2013.
145. A. Arbat, et al., “High voltage vs. high integration: a comparison between CMOS technologies for SPAD cameras,” *Proc. SPIE*, vol.7780, 77801G, 8 pp., 2010.
146. E. Charbon, “Single-photon Imaging in CMOS”, *Proc. SPIE*, vol. 7780, 77801D, 15 pp., 2010.
147. G.-F. Dalla Betta, et al., “Avalanche Photodiodes in Submicron CMOS Technologies for High-Sensitivity Imaging”, *Advances in Photodiodes*, G.-F. Dalla Betta Ed., IntechOpen, 2011, pp. 226-248.
148. A. Rochas, et al., "A Geiger Mode Avalanche Photodiode Fabricated in a Conventional CMOS Technology," *Proc. of 31st Eu. Sol.-St. Dev. Res. Conf.*, 2001, pp.483-486, 2001.
149. J. C. Jackson, et al., “Characterization of Geiger mode avalanche photodiodes for fluorescence decay measurements”. *International Society for Optics and Photonics Symposium on Integrated Optoelectronic Devices*, pp. 55-66, 2002.
150. A. Rochas, et al., "Single photon detector fabricated in a complementary metal–oxide–semiconductor high-voltage technology," *Rev. Sci. Instr.*, vol.74, no.7, pp.3263-3270, 2003.
151. A. Rochas, et al., "First fully integrated 2-D array of single-photon detectors in standard CMOS technology," *IEEE Photon. Tech. Lett.*, vol.15, no.7, pp.963-965, 2003.
152. E. Sciacca, et al. "Silicon planar technology for single-photon optical detectors." *IEEE Trans. Elect. Dev.*, vol. 50, no. 4, pp. 918-925, 2003.
153. F. Zappa, et al., “Complete single-photon counting and timing module in a microchip,” *Opt. Lett.*, vol. 30, no. 11, pp. 1327–1329, 2005.
154. C. Niclass, et al., "Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes," *IEEE J. Sol. St. Circ.*, vol.40, no.9, pp.1847-1854, 2005.
155. A. Rochas, “Single Photon Avalanche Diodes in CMOS technology”, *PhD thesis*, École polytechnique fédérale de Lausanne (EPFL), 2003.
156. A. Zanchi, et al., "A probe detector for defectivity assessment in p-n junctions," *IEEE Trans. Elect. Dev.*, vol. 47, no. 3, pp.609-616, 2000.
157. B. Nouri, et al., “Large-area low-noise Single-Photon Avalanche Diodes in standard CMOS,” *IEEE Sensors*, pp. 1-5, 2012.
158. D. Bronzi, et al., “Figures of Merit for CMOS SPADs and arrays”, *Proc. SPIE*, vol. 8773, 877304, 7 pp., 2013.
159. C. Niclass, et al., "A single photon avalanche diode array fabricated in 0.35 μm CMOS and based on an event-driven readout for TCSPC experiments," *Proc. SPIE*, vol.6372, 63720S, 12 pp., 2006.
160. D. Bronzi, et al., "Low-noise and large-area CMOS SPADs with timing response free from slow tails," *Proc. Eu. Sol.-St. Dev. Rese. Conf. (ESSDERC)*, pp. 230-233, 2012.
161. D. Bronzi, et al. “Large-area CMOS SPADs with very low dark counting rate”, *Proc. SPIE*, vol. 8631, 86311B, 8 pp., 2013.
162. C. Scarcella, et al., “Low-noise low-jitter 32-pixels CMOS single-photon avalanche diodes array for single-photon counting from 300 nm to 900 nm”, *Rev. Sci. Instrum.*, vol. 84, 123112, 2013.
163. C. Niclass, M. Soga, "A miniature actively recharged single-photon detector free of afterpulsing effects with 6ns dead time in a 0.18 µm CMOS technology," *IEEE Int. Elec. Dev. Mtg.*, pp.14.3.1-14.3.4, 2010.
164. T. Leitner, et al., “Measurements and simulations of low dark count rate single photon avalanche diode device in a low voltage 180-nm CMOS image sensor technology,” *IEEE Trans. Electron. Dev.*, vol. 60, no. 6, pp. 1982-1988, 2013
165. L. Pancheri, D. Stoppa, "Low-noise single photon avalanche diodes in 0.15 μm CMOS technology," *Proc. Eu. Sol.-St. Dev. Res. Conf. (ESSDERC)*, pp.179-182, 2011.
166. C. Niclass, et al., "A single photon avalanche diode implemented in 130-nm CMOS technology," *IEEE J. Sel. Top. Quant. Elec.*, vol.13, no.4, pp.863-869, 2007.
167. M Gersbach, et al “A low-noise single-photon detector implemented in a 130nm CMOS imaging process,” *Sol.-St. Elec.*, vol. 53, pp. 803-8, 2009.
168. R. K. Henderson, et al., “Reduction of band-to-band tunneling in deep-submicron CMOS single photon avalanche photodiodes,” *Intl. Image Sensors Workshop (IISW)*, 4 pp, 2009.
169. J. A. Richardson, et al., "Low dark count single-photon avalanche diode structure compatible with standard nanometer scale CMOS technology," *IEEE Photon. Tech. Letts.*, vol.21, pp.1020-1022, 2009.
170. R.M. Field, et al., “A low-noise, single-photon avalanche diode in standard 0.13 μm complementary metal-oxide-semiconductor process”, *Appl. Phys. Lett.*, vol. 97, pp. 211111, 2010.
171. J. A. Richardson, et al., "A 2um diameter, 9Hz dark count, single poton avalanche diode in 130nm CMOS technology," *Proc. Eu. Sol.-St. Dev. Res. Conf. (ESSDERC)*, pp.257-260, 2010.
172. J. A. Richardson, et al., "Scaleable single-photon avalanche diode structures in nanometer CMOS technology," *IEEE Trans. Elec. Dev.*, vol.58, no.7, pp.2028-2035, 2011.
173. E.A.G. Webster, et al., "A high-performance single-photon avalanche diode in 130-nm CMOS imaging technology." *IEEE Elec. Dev. Lett.*, vol. 33, no. 11, pp. 1589-1591, 2012:
174. E. A. G. Webster, et al., “Single-photon avalanche diodes in 90 nm CMOS imaging technology with sub-1Hz median dark count rate,” *Proc. Int. Image Sensor Workshop (IISW)*, R39, 2011.
175. E. A. G. Webster, et al., “A single-photon avalanche diode in 90-nm CMOS imaging technology with 44% Photon Detection Efficiency at 690 nm," *IEEE Elect. Dev. Lett.*, vol.33, no.5, pp.694-696, 2012.
176. M. A. Karami, et al., “Single-Photon Avalanche Diodes in sub-100nm Standard CMOS Technologies,” *Proc. Int. Image Sensor Workshop (IISW)*, p. 16, 2011.
177. E. Charbon, et al., “A Geiger mode APD fabricated in standard 65 nm CMOS technology,” *2013 IEEE Int. Elect. Dev. Meet.*, p. 27.5.1-27.5.4, 2013
178. S. Cova, et al., “Trapping Phenomena in Avalanche Photodiodes on Nanosecond Scale,” *IEEE Elect. Dev. Lett.*, vol. 12, no. 12, pp. 685-687, 1991.
179. W. J. Kindt, H. W. van Zeijl, “Modelling and fabrication of Geiger mode avalanche photodiodes,” *IEEE Trans. Nucl. Sci.*, vol. 45, no. 3, pp. 715-719, 1998
180. A. C. Guidice, et al., “A process and deep level evaluation tool: afterpulsing in avalanche junctions,” *33rd Conf on Eur. Sol.-St. Dev. Res.,* (*ESSDERC*), pp. 347-350, 2003
181. D. Bronzi, et al., "Fast sensing and quenching of CMOS SPADs for minimal afterpulsing effects," *IEEE Photon. Tech. Lett.*, vol.25, no.8, pp.776-779, 2013
182. S. M. Sze, G. Gibbons. "Effect of junction curvature on breakdown voltage in semiconductors." *Sol.-St. Elect.*, vol. 9, pp. 831-845, 1966.
183. C. Basavanagoud, K. N. Bhat. "Effect of lateral curvature on the breakdown voltage of planar diodes." *IEEE Elect. Dev. Lett.*, vol. 6, no. 6 pp. 276-278, 1985.
184. G. Bonanno, et al., “Precision measurements of Photon Detection Efficiency for SiPM detectors,” *Nucl. Instrum. Meth. Phys. Res. A*, vol. 610, no. 1, pp. 93-97, 2009.
185. A. Gulinatti, et al., “Modeling photon detection efficiency and temporal response of single photon avalanche diodes,” *Proc. SPIE*, vol. 7355, 73550X, 17 pp., 2009.
186. J.L. Regolini, et al., "Passivation issues in active pixel CMOS image sensors." *Microelectronics Reliability*, vol. 47, pp. 739-742, 2007.
187. J-P. Carrere, et al., "CMOS image sensor: Process impact on dark current." *IEEE International Symposium on Reliability Physics*, pp. 3C-1, 2014.
188. F. Tavernier, M.S.J. Steyaert, "High-speed optical receivers with integrated photodiode in 130 nm CMOS." *IEEE J. Sol.-St. Circ.*, vol. 44, no.10, pp. 2856-2867, 2009.
189. K. Bach, et al., "Integrated Photodetectors in CMOS Chips and their Spectral Sensitivity", www.xfab.com
190. S. Donati, et al., “Microconcentrators to recover fill-factor in image photodetectors with pixel on-board processing circuits,” *Opt. Exp.*, vol. 15, no. 26, pp. 18066-18075, 2007.
191. E. Randone, et al., "SPAD-array photoresponse is increased by a factor 35 by use of a microlens array concentrator," *IEEE LEOS Annual Meeting Conf.* Proc., pp.324-325, 2009.
192. F. Acerbi, et al., “Characterization of Single-Photon Time Resolution: From Single SPAD to Silicon Photomultiplier”, *IEEE Trans. Nucl. Sci.*, vol. 61, no. 5, pp. 2678-2686, 2014.
193. J. Kalisz, “Review of methods for time interval measurements with picosecond resolution”, *Metrologia*, vol. 41, pp. 17-32, 2004.
194. S. Henzler, *Time-to-Digital Converters*, Springer-Verlag, New York, 2010.
195. P. Carbone, et al., eds*. Design, Modeling and Testing of Data Converters*. Springer, 2014.
196. J. Richardson, et al., "A 32×32 50ps resolution 10 bit time to digital converter array in 130nm CMOS for time correlated imaging," *IEEE Cust. Integ. Circ. Conf.*, pp.77-80, 2009
197. D. Tyndall, et al., "A high-throughput time-resolved mini-silicon photomultiplier with embedded fluorescence lifetime estimation in 0.13 μm CMOS," *IEEE Tran. Biomed. Cir. Sys.*, vol.6, pp.562-570, 2012.
198. Ł. Zaworski, et al., “Quantization error in time-to-digital converters”, *Metrology and Measurement Systems*, vol. 19, no. 1, pp. 115-122, 2012.
199. F. Baronti, et al., "On the differential nonlinearity of time-to-digital converters based on delay-locked-loop delay lines," *IEEE Trans. on Nucl. Sci.*, vol.48, no.6, pp.2424-2431, 2001.
200. Texas Instruments. "Understanding data converters." SLAA013, 1995.
201. M. J. M. Pelgrom, *Analog-to-digital Conversion*. Springer Netherlands, 2010.
202. J.-P. Jansson, “A stabilized multi-channel CMOS time-to-digital converter based on a low frequency reference”, *PhD thesis*, Oulu, 2012.
203. A. Mantyniemi, “An integrated CMOS high precision TDC based on stabilized three-stage delay line interpolation”, *PhD thesis*, Oulu, 2004.
204. M. Crotti, et al., "Four Channel, 40 ps Resolution, Fully Integrated Time-to-Amplitude Converter for Time-Resolved Photon Counting," *IEEE J. Sol.-St. Circ.*, vol.47, no.3, pp.699-708, 2012.
205. D. Stoppa, et al., "A 32x32-pixel array with in-pixel photon counting and arrival time measurement in the analog domain," *Proc. ESSCIRC*, pp.204-207, 2009.
206. T. E. Rahkonen, J T. Kostamovaara, "The use of stabilized CMOS delay lines for the digitization of short time intervals," *IEEE J. Sol.-St. Circ.*, vol.28, no.8, pp.887-894, 1993
207. *Fundamentals of Time Interval Measurement*, Hewlett-Packard App. Note 200-3, 1997.
208. R. Szplet, M. Gołaszewski, “Integrated time-to-digital converter with the use of the counter method and a multiphase clock”, *Pomiary, Automatyka, Kontrola*, vol. 54, no. 8, pp. 591-593, 2008 (in Polish)
209. J. Christiansen, "An integrated high resolution CMOS timing generator based on an array of delay locked loops," *IEEE J. Sol.-St. Circ.*, vol.31, no.7, pp.952-957, 1996
210. M. Mota, J. Christiansen, "A high-resolution time interpolator based on a delay locked loop and an RC delay line," *IEEE J. Sol.-St. Circ.*, vol.34, no.10, pp.1360-1366, 1999.
211. W. Gao, et al., "Precise Multiphase Clock Generation Using Low-Jitter Delay-Locked Loop Techniques for Positron Emission Tomography Imaging," *IEEE Trans. Nucl. Sci.*, vol.57, no.3, pp.1063-1070, 2010.
212. K. Arshak, et al., “Design and simulation difference types CMOS phase frequency detector for high speed and low jitter PLL,” *Proc. 5th* *IEEE Intl. Conf. Dev., Circ. & Syst.*, vol. 1, pp.188-191, 2004.
213. W. Rhee, "Design of high-performance CMOS charge pumps in phase-locked loops," *Proceedings of the 1999 IEEE Intern. Symp. Circ. Sys.*, vol. 2, pp. 545-548. 1999.
214. J. G. Maneatis, “Low-Jitter Process-Independent DLL and PLL Based on Self-Biased Techniques,” *IEEE J. Sol. St. Circ.*, vol. 31, no. 11, pp.1723-1732, 1996.
215. H.-H. Chang, et al., "A wide-range delay-locked loop with a fixed latency of one clock cycle," *IEEE J. Sol. St. Circ.*, vol.37, no.8, pp.1021-1027, 2002
216. J.-P. Jansson, "A CMOS time-to-digital converter with better than 10 ps single-shot precision," *IEEE J. Sol. St. Circ.*, vol.41, no.6, pp.1286-1296, 2006.
217. R. Tonietto, et al., "A 3MHz Bandwidth Low Noise RF All Digital PLL with 12ps Resolution Time to Digital Converter," *Proc. 32nd Eu. Sol. St. Circ. Conf., ESSCIRC* pp.150-153, 2006
218. R. B. Staszewski, et al.. "1.3 V 20 ps time-to-digital converter for frequency synthesis in 90-nm CMOS" *IEEE Trans. Circ. Sys. II: Express Briefs*, vol. 53, no. 3 pp. 220-224, 2006.
219. J.-P. Jansson, et al.. "Synchronization in a multilevel CMOS time-to-digital converter." *IEEE Trans. Circ. Sys. I: Regular Papers*, vol. 56, no. 8, pp. 1622-1634, 2009.
220. C.-S. Hwang, et al., "A high-precision time-to-digital converter using a two-level conversion scheme," *IEEE Trans. Nucl. Sci.*, vol.51, no.4, pp.1349,1352, Aug. 2004
221. B. Markovic, et al., "A High-Linearity, 17 ps Precision Time-to-Digital Converter Based on a Single-Stage Vernier Delay Loop Fine Interpolation," *IEEE Trans. Circ. Sys. I: Regular Papers*, vol.60, no.3, pp.557,569, March 2013
222. X.-J. Zhang, et al., "A low-power coarse-fine time-to-digital converter in 65nm CMOS," *2015 Intern. Symp. Sig., Circ. Sys. (ISSCS)*, pp.1-4, 9-10, 2015.
223. Z. Cheng, et al., "A Low-Power Gateable Vernier Ring Oscillator Time-to-Digital Converter for Biomedical Imaging Applications", *IEEE Trans. Biomed. Circ. Sys*, (in press), 2015.
224. B. B. Calhoun, et al., “Digital circuit design challenges and opportunities in the era of nanoscale CMOS*”, Proc. IEEE*, vol. 96, no. 2, 343-365, 2008.
225. T. Hamamoto, "Sidewall damage in a silicon substrate caused by trench etching." *Appl. Phys. Lett.*, vol. 58, no. 25, pp. 2942-2944, 1991.
226. S. Tyagi, et al. "A 130 nm generation logic technology featuring 70 nm transistors, dual Vt transistors and 6 layers of Cu interconnects." *Intern. Elect. Dev. Meet. Tech. Digest IEDM*, 2000.
227. N. Zamdmer, et al, "A 0.13-/spl mu/m SOI CMOS technology for low-power digital and RF applications*," Symp. VLSI Tech., Digest of Technical Papers*. pp.85-86, 2001
228. S. Thompson, et al, '130nm Logic Technology Featuring 60nm Transistors, Low-K Dielectrics, and Cu Interconnects', *Intel Tech. J.*, vol. 6, no. 2, pp. 5-13, 2002.
229. P. M. Zeitzoff, "Circuit, MOSFET, and front end process integration trends and challenges for the 180 nm and below technology generations: an International Technology Roadmap for Semiconductors perspective," *Proc. 6th Intern. Conf. Sol. St. I.C. Tech.*, vol.1, pp.23-28 2001
230. H. S. Wong, “Technology and device scaling considerations for CMOS imagers”, *IEEE Trans. Elect. Dev.*, vol. 43, no. 12, pp. 2131-2142, 1996.
231. A. Hoffman, et al., "CMOS detector technology." *Experimental Astronomy*, vol. 19, nos. 1-3, pp. 111-134, 2003.
232. S. Mandai, et al., “A wide spectral range single-photon avalanche diode fabricated in an advanced 180 nm CMOS technology”, *Opt. Exp.*, vol. 20, no. 6, pp. 5849-5857, 2012.
233. E.A.G. Webster, et al., “Transient Single-Photon Avalanche Diode Operation, Minority Carrier Effects, and Bipolar Latch Up,” *IEEE Trans. Elec. Dev.*, vol.60, no.3, pp.1188-1194, 2013.
234. R. J. Walker, et al., "High fill factor digital silicon photomultiplier structures in 130nm CMOS imaging technology," *IEEE Nucl. Sci. Symp. Med. Imag. Conf. (NSS/MIC)*, pp.1945-1948, 2012.
235. F.-P. Chou, et al. "Silicon photodiodes in standard CMOS technology." *IEEE Journal of Sel. Top. Quant. Elect.*, vol. 17, no. 3, pp. 730-740, 2011.
236. E. Kamrani, et al., "Premature edge breakdown prevention techniques in CMOS APD fabrication." *IEEE 10th Intern. New Circ. Sys. Conf. (NEWCAS)*, 2012.
237. M.-J. Lee, et al., "Effects of guard-ring structures on the performance of silicon avalanche photodetectors fabricated with standard CMOS technology." *IEEE Elect. Dev. Lett.*, vol. 33, no. 1, pp. 80-8, 2012.
238. Q. Pan, et al. "A 65-nm CMOS P-well/Deep N-well avalanche photodetector for integrated 850-nm optical." *2013 IEEE 10th Intern. Conf. on ASIC (ASICON)*, 2013.
239. M. Atef, et al., "Avalanche double photodiode in 40-nm standard CMOS technology." *IEEE J. Quant. Electr.*, vol. 49, no. 3, pp. 350-356, 2013.
240. H. I. Kwon, et al. "The analysis of dark signals in the CMOS APS imagers from the characterization of test structures." *IEEE Trans. Elect. Dev.*, vol. 51, no. 2, pp. 178-184, 2004.
241. N. Teranishi, "Effect and Limitation of Pinned Photodiode", *IEEE Trans. Elect. Dev.*, vol. 51, no. 2, pp. 178-184, 2015.
242. H. Finkelstein, et al., "STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology." *IEEE Elect. Dev. Lett.*, vol. 27, no. 11, pp. 887-889, 2006.
243. M. Dandin, et al., “Characterization of single-photon avalanche diodes in a 0.5 µm standard CMOS process - Part1: Perimeter breakdown suppression”, *IEEE Sensors J*, vol. 10, no. 11, pp.1682-1690, 2010
244. B.-Y. Tsui, et al., "Impact of silicide formation on the resistance of common source/drain region." *IEEE Elect. Dev. Lett.*, vol. 22, no. 10, pp. 463-465, 2001.
245. S-L. Zhang, U. Smith, "Self-aligned silicides for Ohmic contacts in complementary metal–oxide–semiconductor technology: TiSi2, CoSi2, and NiSi." *J. Vac. Sci. Tech. A,* vol. 22, no. 4, pp. 1361-1370, 2004.
246. K. Goto, et al. "A new leakage mechanism of Co salicide and optimized process conditions [for CMOS]." *IEEE Trans. Elect. Dev.*, vol. 46, no. 1, 117-124, 1999.
247. H.-D. Lee, "Characterization of shallow silicided junctions for sub-quarter micron ULSI technology. Extraction of silicidation induced Schottky contact area." *IEEE Trans Elect. Dev.*, vol. 47, no. 4, pp. 762-767, 2000.
248. D. Codegoni, et al. "Leakage current and deep levels in CoSi 2 silicided junctions." *Mat. Sci. Eng.: B,* vol. 124, pp.349-353, 2005.
249. S.-G. Wuu, et al. "High performance 0.25-um CMOS color imager technology with non-silicide source/drain pixel." *Intern. Elect. Dev. Meet.Technical Digest, (IEDM)*, 2000.
250. S.-G. Wuu, et al. "A high performance active pixel sensor with 0.18 um CMOS color imager technology." *Intern. Elect. Dev. Meet. Technical Digest, (IEDM)*, 2001.
251. M. Cohen, et al. "Fully Optimized Cu based process with dedicated cavity etch for 1.75μm and 1.45μm pixel pitch CMOS Image Sensors." *Intern. Elect. Dev. Meet. Technical Digest, (IEDM)*, 2006.
252. R. A. Yotter, et al., "Optimized CMOS photodetector structures for the detection of green luminescent probes in biological applications." *Sens. & Act. B: Chemical*, vol. 103, no. 1 pp. 43-49, 2004.
253. H. M. Jafari, et al., "Nanostructured CMOS Wireless Ultra-Wideband Label-Free PCR-Free DNA Analysis SoC," *IEEE J. S. St. Circ.*, vol.49, no.5, pp.1223-1241, 2014.
254. D. A. Hodges, et al., *Analysis and Design of Digital Integrated Circuits in Deep Submicron Technology, 3rd Ed.,* McGraw-Hill, Boston, 2004.
255. E. A. G. Webster, et al., "Transient Single-Photon Avalanche Diode Operation, Minority Carrier Effects, and Bipolar Latch Up." *IEEE Trans. Elect. Dev.*, vol. 60, no. 3, pp. 1188-1194, 2013.
256. F. Zappa, et al., “SPICE modeling of single photon avalanche diodes,” *Sens. & Act. A*, vol. 153, pp. 197-204, 2009.
257. D. Huang, “SPICE Modeling for Single Photon Avalanche Diode,” *Proc. SPIE*, vol.8908, 8908A, 14 pp., 2013.
258. G. Giustolisi, et al., “Behavioral modeling of statistical phenomena of single-photon avalanche diodes,” *Int. J. Circ. Theor. Appl.*, vol. 40, pg. 661-679, 2011.
259. L. Neri, “Note: Dead time causes and correction method for single photon avalanche diode devices,” *Rev. Sci. Instr.*, vol. 81, 086102, 2010.
260. M. Anti, et al. "Integrated simulator for single photon avalanche diodes." *IEEE Intern. Conf. on Numerical Simulation of Optoelectronic Devices (NUSOD)*, pp. 47-48, 2011.
261. N. V. Loukianova, et al. "Leakage current modeling of test structures for characterization of dark current in CMOS image sensors." *IEEE Trans. Elect. Dev.*, vol. 50, no. 1, pp. 77-83, 2003.
262. A. W. Drake, *Fundamentals of Applied Probability Theory*, New York, NY, USA: McGraw-Hill, 1988, pp. 133–140.
263. A. Yoshizawa, et al., “Quantum efficiency evaluation method for gated-mode single-photon detector,” *Electron. Lett.*, vol. 38, no. 23, pp. 1468–1469, 2002.
264. G. Humer, et al., "A Simple and Robust Method for Estimating Afterpulsing in Single Photon Detectors," *J. of Lightwave Technology*, vol.33, no.14, pp.3098-3107, 2015
265. R. K. Henderson, et al., "A gate Modulated avalanche bipolar transistor in 130nm CMOS technology," *Proc. Eur. Sol. St. Dev. Res. Conf. (ESSDERC)*, pp.226-229, 2012
266. C. Veerappan, et al., “Characterization of large-scale non-uniformities in a 20k TDC/SPAD array integrated in a 130nm CMOS process”. *Proc. Eur. Sol. St. Dev. Res. Conf. (ESSDERC)*, pp. 331-334, 2011.
267. E. A. G. Webster, R. K. Henderson. "A TCAD and spectroscopy study of dark count mechanisms in single-photon avalanche diodes." *IEEE Trans. Elect. Dev.*, vol. 60, no. 12, pp. 4014-4019, 2013.
268. I. M. Antolovic, et al., "Nonuniformity Analysis of a 65-kpixel CMOS SPAD Imager," *IEEE Trans. Elect. Dev.*, (in press), 2015.
269. M. J. Deen, et al.. "Low frequency noise in polysilicon‐emitter bipolar junction transistors." *J. Appl.Phys.*, vol. 77, no. 12, pp. 6278-6288, 1995.
270. A. Czerwinski, et al. "Activation energy analysis as a tool for extraction and investigation of p–n junction leakage current components." *J. Appl. Phy.*, vol. 94, no. 2, pp. 1218-1221, 2003.
271. E. A. G. Webster, et al. "Per-pixel dark current spectroscopy measurement and analysis in CMOS image sensors." *IEEE Trans. Elect. Dev.*, vol. 57, no. 9, pp. 2176-2182, 2010.
272. J. C. Dunlap, et al., "Interpreting Activation Energies in Digital Image Sensors," *IEEE Trans. Elect. Dev.*, (in press), 2015.
273. G. Vincent, et al., "Electric field effect on the thermal emission of traps in semiconductor junctions," *J. Appl.Phys.*, vol. 50, no. 8, pp. 5484-5487, 1979.
274. H. C. Burstyn, "Afterpulsing effects in photon correlation experiments." *Rev. Sci. Instr.*, vol. 51, no. 10, pp. 1431-1433, 1980.
275. M. Höbel, J. Ricka, “Deadtime and afterpulsing correction in multiphoton timing with nonideal detectors,” *Rev. Sci. Instr.*, vol. 62, no. 7, pp. 2326-2336, 1994.
276. E. Overbeck, et al., “Silicon avalanche photodiodes as detectors for photon correlation experiments,” *Rev. Sci. Instrum.*, vol. 69, p. 3515-3523, 1998.
277. S. Gong, et al. "A 32-channel photon counting module with embedded auto/cross-correlators for real-time parallel fluorescence correlation spectroscopy." *Rev. Sci. Instr.*, vol. 85, no. 10, pp. 103101, 2014.
278. S. Gong, et al. "A simple and flexible FPGA based autocorrelator for afterpulse characterization of single-photon detectors." *11th Intern. IEEE Multi-Conference on Systems, Signals & Devices (SSD)*, 2014.
279. Y. Kang, et al. "Afterpulsing of single-photon avalanche photodetectors." *16th Annual Meeting of the IEEE Lasers and Electro-Optics Society, (LEOS)*, vol. 2, 2003.
280. M. Anti, et al., “Modeling of afterpulsing in Single-Photon Avalanche Diodes,” *Proc. SPIE*, vol.7933, 79331R, 8 pp., 2011.
281. M.A. Itzler, et al., “Power law temporal dependence of InGaAs/InP SPAD afterpulsing,” *J. Mod. Opt.*, vol. 59, no. 17, pp. 1472-1480, 2012.
282. M.A. Itzler, et al., "Dark Count Statistics in Geiger-Mode Avalanche Photodiode Cameras for 3-D Imaging LADAR," *IEEE Sel. Top. Quant. Elec.*, vol.20, no.6, pp.318-328, 2014.

1. <http://www.micro-photon-devices.com/Products/Photon-Counters/PDM-PDF>
2. M. Liu, et al. "Reduce afterpulsing of single photon avalanche diodes using passive quenching with active reset." *IEEE J. Quant. Elect.*, vol. 44, no. 5, pp. 430-434, 2008.
3. H. Chong, et al. "Dynamic range of passive quenching active reset circuit for single photon avalanche diodes." *IEEE J. Quant. Elect.*, vol. 46, no. 1, pp. 35-39, 2010.
4. G. Bonanno, et al. "Precision measurements of photon detection efficiency for SiPM detectors." *Nucl. Instr. Meth. Phys. Res. A,* vol. 610, pp. 93-97, 2009.
5. M. Stipcevic, et al., "Characterization of a commercially available large area, high detection efficiency single-photon avalanche diode." *J. Lightwave Technology*, vol. 31, no. 23, pp. 3591-3596, 2013.
6. K. Iiyama, et al., "Hole-Injection-Type and Electron-Injection-Type Silicon Avalanche Photodiodes Fabricated by Standard 0.18-m CMOS Process." *IEEE Photon. Tech. Lett.*, vol. 22, no. 12, pp. 932-934, 2010.
7. B. Nakhkoob, et al., "High speed photodiodes in standard nanometer scale CMOS technology: a comparative study." *Opt. Exp.*, vol. 20, no. 10, pp. 11256-11270, 2012.
8. D. Durini, "Solid-State Imaging in Standard CMOS Processes." *PhD thesis*., Universität Duisburg-Essen, 2009.
9. Y. Yamada, S. Okawa. "Diffuse optical tomography: Present status and its future." *Opt. Rev.*, vol. 21, no. 3, pp. 185-205, 2014.
10. R. R. Alfano, et al., "Effect of soap on the fluorescent lifetime and quantum yield of rhodamine 6G in water." *Optics Communications*, vol. 7, no. 3, pp. 191-192, 1973.
11. K. A. Selanger, et al., "Fluorescence lifetime studies of Rhodamine 6G in methanol." *The J. Phys. Chem.*, vol. 81, no. 20 pp. 1960-1963, 1977.
12. M. Ishikawa, et al. "Simultaneous measurement of the fluorescence spectrum and lifetime of rhodamine b in solution with a fluorometer based on streak-camera technologies." *Analytical Chemistry*, vol. 67, no. 3, pp. 511-518, 1995.
13. M. Fischer, J. Georges, "Fluorescence quantum yield of rhodamine 6G in ethanol as a function of concentration using thermal lens spectrometry." *Chemical Physics Letters*, vol. 260, no. 1, pp. 115-118, 1996.
14. D. Magde, et al., “Solvent dependence of the fluorescence lifetimes of xanthene dyes”, *Photochemistry and Photobiology*, vol. 70, pp. 737-744, 1999.
15. F. M. Zehentbauer, et al. "Fluorescence spectroscopy of Rhodamine 6G: Concentration and solvent effects." *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 121, pp. 147-151, 2014.
16. A. Penzkofer, Y. Lu. "Fluorescence quenching of rhodamine 6G in methanol at high concentration." *Chemical physics*, vol. 103, no. 2, pp. 399-405, 1986.
17. J. N. Miller, Ed., *Standards in Fluorescence Spectrometry,* London: Chapman & Hall Ltd, 1981.
18. W. Koechner, *Solid-state Laser Engineering*. Vol. 1. Springer: Berlin, 2013.
19. H. N. Aizawa, et al., "Fabrication of ruby sensor probe for the fiber-optic thermometer using fluorescence decay." *R. Sci. Instr.*, vol. 73, no. 10, pp. 3656-3658, 2002.
20. H. C. Seat, et al., "Single-crystal ruby fiber temperature sensor." *Sens. & Act. A*, vol. 101, no. 1, pp. 24-29, 2002.
21. V. B. Urošević, et al., "Effect of pressure on the ruby fluorescence lifetime." *Chemical Physics Letters*, vol. 155, no. 3, pp. 325-328, 1989.
22. D. E. Chandler, et al., "Ruby crystal for demonstrating time-and frequency-domain methods of fluorescence lifetime measurements." *J. of Fluorescence*, vol. 16, no. 6 pp. 793-807, 2006.
23. R. J. Baker, et al., *CMOS: Circuit Design, Layout and Simulation*, Wiley: New York, 1998.