UC San Diego
UC San Diego Electronic Theses and Dissertations

Title
Shallow-trench-isolation bounded single-photon avalanche diodes in commercial deep submicron CMOS technologies

Permalink
https://escholarship.org/uc/item/3j73m45f

Author
Finkelstein, Hod

Publication Date
2007

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Shallow-Trench-Isolation Bounded Single-Photon Avalanche Diodes in Commercial Deep Submicron CMOS Technologies

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Electrical Engineering (Photonics)

by

Hod Finkelstein

Committee in charge:

Professor Sadik C. Esener, Chair
Professor Michael J. Heller
Professor Yu-hwa Lo
Professor Robert Mattrey
Professor Shankar Subramanian

2007

Copyright

Hod Finkelstein, 2007

All rights reserved

The dissertation of Hod Finkelstein is approved, and it is acceptable in quality and form for publication on microfilm:

____

____

____

____

____

Chair

University of California, San Diego

2007

To my family.

TABLE OF CONTENTS

Signature Page ... iii

Dedication ... iv

Table of Contents ... v

List of Tables ... x

List of Figures ... xi

Acknowledgements ... xix

Vita and Publications ... xxiii

Abstract ... xxvii

1. Introduction ... 1

   1.1. The Power of CMOS in Single-Photon Detection ... 1

   1.2. Applications of Single-Photon Detectors ... 4

      1.2.1. Ladar and Long-range Three-Dimensional Imaging ... 4

      1.2.2. High-Resolution Three-Dimensional Imaging ... 6

      1.2.3. Single Photon Detection for Molecular Imaging and Spectroscopy ... 7

      1.2.4. Time-Domain Diffuse Optical Tomography ... 19

   1.3. State of the Art in Single-Photon and Low-Light-Level Detection ... 20

      1.3.1. Photomultiplier Tubes and Micro-Channel Plates ... 21

      1.3.2. Superconducting Wire Detector ... 24

      1.3.3. Charge-Coupled Devices ... 26

   1.4. Single-Photon Avalanche Diodes ... 30

1.4.1. Sub-Geiger APDs 30

1.4.2. Geiger-Mode Single-Photon Avalanche Diodes 32

1.5. Performance Comparison 45

1.6. Aim and Challenges 45

1.7. Dissertation Outline 51

2. Physics and Operation of Single Photon Avalanche Diodes 53

2.1. Introduction 53

2.2. Basic Properties of the pn Junction 54

2.3. Breakdown Mechanisms in pn Junctions 59

2.3.1. Thermal Instabilities 59

2.3.2. Tunneling 59

2.3.3. Impact Ionization 61

2.4. Breakdown in Non-Planar pn Junctions 65

2.5. Solid-State Guard Rings for High-Field Devices 70

2.6. Geiger-Mode Single-Photon Avalanche Diodes: Principle of Operation 74

2.7. Avalanche Quenching and Junction Recharge 76

2.7.1. Passive Quenching and Recharging 76

2.7.2. Active Quenching and Recharging 80

2.7.3. Negative-feedback Quenching 82

2.8. Figures of Merit 82

2.8.1. Photon Detection Probability and Spectral Response 83

2.8.2. Dead Time 97

2.8.3. Noise 98

2.8.4. Timing Precision 109

2.8.5. Dynamic Range 111

2.8.6. Cross-talk 113

2.9. Conclusions 114

3. STI-bounded Single Photon Avalanche Diode – Analysis, Modeling and Simulation 117

3.1. Introduction 117

3.2. Device Concept 118

3.3. Device Modeling 122

3.3.1. Physical Modeling 122

3.3.2. Electrical Modeling 125

3.4. Peripheral Circuits Design and Simulation 128

3.4.1. Output Buffers 128

3.4.2. Ultrafast and Compact Active Recharge Circuit 133

3.5. Dual-Color Single-Photon Detection 143

3.5.1. Device Concept 143

3.5.2. Peripheral Circuitry 148

3.6. Conclusion 154

4. Device Characterization 156

4.1. Introduction 156

4.2. Test Devices 157

4.3. Validation of Junction Planarity 158

4.4. Avalanche Pulse 167

4.5. Dead Time 171

4.6. Dark Current 176

4.7. Detection Efficiency and Spectral Response 183

4.8. Optimization Using Active Recharge 190

4.9. Cross-Talk 195

4.10. Performance Comparison with Diffused-Ring SPAD 198

4.11. Dual-Color Single-Photon Detection 203

4.12. Conclusions 209

5. Single-photon Frequency Upconversion via Hot-carrier Electroluminescence 213

5.1. Introduction 213

5.2. Hot-carrier Luminescence in Avalanche Photodiodes 220

5.3. Model for Upconversion Efficiency 223

5.3.1. Upconversion Model 223

5.3.2. Secondary Photon Emission towards the silicon Junction 226

5.3.3. Secondary Photons Self-Absorption Probability 228

5.3.4. Secondary Photons Absorption Probability in Silicon 229

5.3.5. Secondary Avalanche Initiation Probability 230

5.3.6. Numerical Calculations and Design Considerations 231

5.4. Experimental Measurements of Hot-Carrier Electroluminescence 234

5.4.1. Experimental Methodology 234

5.4.2. Optical Response 236

5.4.3. InAlAs/InGaAs Electroluminescence Yield 241

5.4.4. InAlAs Optical Absorption 243

5.5. Conclusions 245

6. Conclusion 247

6.1. Dissertation Summary and Original Contributions 247

6.2. Outlook 249

6.3. Future research 252

References……………………………………………………………………...247

LIST OF TABLES

Table 1.1. Comparison between bulk and surface SPAD performance. 40

Table 1.2. Comparison of detector performance 46

Table 2.1. Summary of SPAD noise sources 108

Table 2.2. Relationship between SPAD figures-of-merit and select single-photon detection applications 115

Table 4.1. Summary of cross-talk experiments. 195

LIST OF FIGURES

Figure 1.1. Optical setup for FCS.. 12

Figure 1.2. FCS simulation scheme and results. 15

Figure 1.3. Dual-color FCCS setup. 17

Figure 1.4. In vitro fluorescence and pH-dependent lifetime-shift image of a stained fibroblast 18

Figure 1.5. Cross-section of a photomultiplier tube 22

Figure 1.6. Operation of a superconducting nanowire photon counter. 25

Figure 1.7. Cross-section of a CCD pixel 26

Figure 1.8. Cross-section of an avalanche photodiode 31

Figure 1.9. Reach-through device diagram 34

Figure 1.10. Slik^{TM} device 35

Figure 1.11. Bulk SPAD used for lidar application. 36

Figure 1.12. Haitz's diode using a diffused guard ring 38

Figure 1.13. Cova's double-epitaxy SPAD 39

Figure 1.14. Cross-section of a triple-well CMOS SPAD. 41

Figure 1.15. Rochas' passively-quenched triple-well CMOS SPAD 44

Figure 1.16. Zappa's actively-recharged triple-well CMOS SPAD 44

Figure 2.1. Doping concentration and electric field around an abrupt pn junction. 55

Figure 2.2. Doping concentration and electric field around an asymmetrical linearly-graded pn junction. 58

Figure 2.3. Energy band diagram illustrating direct band-to-band tunneling. 60

Figure 2.4. Energy band diagram illustrating multiplication by impact ionization...... 61

Figure 2.5. Geometry for calculating the avalanche condition. 62

Figure 2.6. Electron and hole ionization coefficients for silicon 64

Figure 2.7. Effect of junction geometry on breakdown voltage................................. 68

Figure 2.8. Effect of junction curvature on breakdown voltage................................. 68

Figure 2.9. ISE-TCAD simulation of electric field distribution across a curved pn junction 70

Figure 2.10. Beveled junction 71

Figure 2.11. Mesa isolation structure. 72

Figure 2.12. Straddled junction 72

Figure 2.13. Field-limiting ring.. 73

Figure 2.14. Metal overhang guard ring 73

Figure 2.15. Diffused guard ring 74

Figure 2.16. Equivalent circuit for passively-quenched passively-recharged SPAD... 77

Figure 2.17. Photon absorption in an indirect-bandgap semiconductor. 87

Figure 2.18. Silicon absorption coefficient 87

Figure 2.19. Absorption probabilities for photons of various wavelengths impinging from the surface 89

Figure 2.20. Absorption probabilities a shallow and deep junction 90

Figure 2.21. Avalanche initiation probability as a function of primary-pair generation depth in a linearly-graded junction 94

Figure 2.22. Avalanche initiation probability as a function of voltage in a narrow junction ... 95
Figure 2.23. Trap-assisted generation ... 99
Figure 2.24. Band-to-band tunneling ... 102
Figure 3.1. A cross-section diagram of the new device. ... 121
Figure 3.2. Calculated spectral response of the new SPAD ... 122
Figure 3.3. ISE-TCAD simulation of electric field distribution in the SPAD ... 123
Figure 3.4. ISE-TCAD simulation of current in the SPAD. ... 124
Figure 3.5. ISE-TCAD simulation of electric field distribution in diffused-ring SPAD ... 124
Figure 3.6. Nwell and Output waveforms for passively-quenched SPAD with source-follower output stage ... 129
Figure 3.7. Schematics and layout of SPAD with source-follower output ... 131
Figure 3.8. Schematic, layout and simulation output waveforms for SPAD with inverter-chain output stage ... 132
Figure 3.9. Passive and active recharge waveforms. ... 138
Figure 3.10. Schematic diagram of the active-recharge circuit. ... 139
Figure 3.11. Cadence simulation results of an active-recharged SPAD cycle. ... 141
Figure 3.12. Layout of actively-recharged SPAD ... 142
Figure 3.13. Simulation schematic of active-recharge circuit with externally-controlled delay for afterpulse optimization. ... 143

Figure 3.14. Simulation waveforms of active-recharge with externally-controlled delay.. 146
Figure 3.15. Operating principle of dual-junction SPAD 147
Figure 3.16. High-level schematic, simulation model and layout of a dual-color SPAD and readout circuitry 151
Figure 3.17. Electrical simulation results of dual-junction SPAD 153
Figure 4.1. Microscope images of the first and second test chips. 160
Figure 4.2. Microscope image of various SPAD pixels. 161
Figure 4.3. Cross-section of non-planar p+/N-well junction 162
Figure 4.4. Layout of a semi-cylindrical, semi-spherical and STI-bounded diode. 163
Figure 4.5. Socketed test device 164
Figure 4.6. Reverse-bias I-V curves for various junction structures 165
Figure 4.7. Electric field distribution at the edge of a non-planar LDD-bounded pn junction. 166
Figure 4.8. Photocounts as a function of beam position on the device 167
Figure 4.9. Photo and schematic of PCB for measuring avalanche behavior of externally-quenched SPAD 168
Figure 4.10. Single avalanche current pulse 169
Figure 4.11. Avalanche charge vs. overbias for hybrid-quenched SPAD 170
Figure 4.12. Microscope image of SPAD with integrated quenching resistor. 170
Figure 4.13. Oscilloscope image showing hybrid-quenched device dead time and recharge 171

Figure 4.14. Photograph and schematic of dead-time measurement setup using time-histograms 174

Figure 4.15. Time-histogram of dark pulse arrivals for the first test chip pixel......... 175

Figure 4.16. Oscilloscope snapshot of integrated 7 μm SPAD with source follower output 175

Figure 4.17. Avalanche pulses at 2.5V over-bias 176

Figure 4.18. Time-histogram of dark counts processed by time-shifting method...... 177

Figure 4.19. PCB for second test chip characterization and photo of experimental setup for measuring dark pulses. 178

Figure 4.20. Setup for measurement of dark current statistics 179

Figure 4.21. Auto-correlation of SPAD dark pulses with integrated quenching........ 181

Figure 4.22. Histogram of afterpulses with counter triggered by a dark pulse 181

Figure 4.23. Dark count rate dependence on overbias and temperature 182

Figure 4.24. Scope image of a SPAD illuminated by a pulsed laser source 185

Figure 4.25. Schematic and photograph if experimental setup for measuring SPAD detection efficiency and output statistics... 187

Figure 4.26. SPAD counts vs. position with a pulsed laser source for optical alignment 188

Figure 4.27. Time histogram with pulsed-laser excitation at 635nm for various overbias voltages 188

Figure 4.28. Setup for measuring spectral response using a CW source 189

Figure 4.29. Spectral response for STI-SPAD with source-follower output............. 189

Figure 4.30. Micrograph of the new active-recharge circuit 190

Figure 4.31. Oscilloscope image of SPAD output with 3 ns dead time 191

Figure 4.32. Dark count rates for identical diodes with passive- and active-recharging 192

Figure 4.33. Autocorrelation curves of passively- and actively-recharged SPAD..... 193

Figure 4.34. Detection efficiency versus dark count rate for a passively- and an actively-recharged SPAD with 635 nm illumination 194

Figure 4.35. Layout of SPAD array used for cross-talk experiments 196

Figure 4.36. Time histogram of cross-talk experiment 197

Figure 4.37. Cross-section of triple-well SAPD with diffused guard-ring................ 199

Figure 4.38. Spectral responses of passively-quenched STI-bounded and diffused-ring SPADs 200

Figure 4.39. Dark count rate vs. voltage for SPADs with diffused and STI rings..... 201

Figure 4.40. Scope waveform of STI-bounded SPAD with source-follower output . 202

Figure 4.41. Detection efficiency vs. DCR for STI-bounded and diffused-ring SPADs 202

Figure 4.42. Experimental setup for characterizing externally-quenched dual-color SPAD.. 204

Figure 4.43. Scope waveforms for externally-quenched dual-junction SPAD 207

Figure 4.44. Cross-correlation between shallow- and deep-junction dark pulses...... 207

Figure 4.45. Spectral response of shallow and deep junctions................................. 208

Figure 5.1. InP energy band diagram illustrating hot-carrier luminescence. 221

Figure 5.2. Calculated spectrum of electroluminescent photons emitted from an InP pn junction 222
Figure 5.3. Cross-section of the proposed upconversion device. 225
Figure 5.4. Geometrical construction for calculating the percentage of photons emitted from an InP junction plane onto a Si junction plane. 228
Figure 5.5. Numerical analysis of avalanche initiation probabilities as a function of photon absorption depth in a Si SPAD. 232
Figure 5.6. Calculated junction and surface electroluminescence spectral densities. 233
Figure 5.7. Numerical simulation of internal upconversion efficiency as a function of primary SPAD's junction capacitance. 234
Figure 5.8. Cross-section of InGaAs/InAlAs device used for emission measurements. 235
Figure 5.9. Hot-carrier electroluminescence from a diffused-ring silicon SPAD. 237
Figure 5.10. Total emission intensity as a function of avalanche current. 238
Figure 5.11. Silicon electroluminescence spectrum. 239
Figure 5.12. Experimental and published electroluminescence curves for 0.2 μm-deep silicon junction. 240
Figure 5.13. Collection efficiency of optical setup. 240
Figure 5.14. Electroluminescence from an InAlAs/InGaAs SPAD 241
Figure 5.15. Electroluminescence spectrum of InAlAs/InGaAs SPAD. 242
Figure 5.16. Electroluminescence surface emission from InAlAs/InGaAs SPAD. ... 242
Figure 5.17. Absorbance of layer *s1* and *s3* corresponding to lower- and upper-

bounds of junction electroluminescence yields. 244

Figure 5.18 Measured absorption coefficient of InAlAs. 245

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Prof. Sadik Esener, for his mentorship and guidance during my Ph. D. years. Prof. Esener gave me the privilege to find my way by exploring diverse fields, from nanotechnology to chemistry and finally to single-photon detection. He was able to focus me on the most significant contributions of my research, yet keeping in mind the end applications for which they were intended – a skill which I believe will be priceless in my future pursuits. Prof. Esener's vision was the seed for this work, for which I am thankful. I am also grateful for the teamwork Prof. Esener encouraged throughout my Ph. D. studies, which made this work possible.

I am also indebted to Prof. Yu-Hwa Lo, whose patience and technical guidance were invaluable for my single-photon detection work. Prof. Lo's rich experience in single-photon detection enabled me to understand many of the experimental observations, particularly in the context of the upconversion work presented in this dissertation.

I spent one of my most constructive Ph. D. years in Prof. Peter Rentzepis' chemistry laboratory at UC Irvine. Under his close guidance, as well as Dr. Alexander Dvornikov's, I gained important experimental skills and systematic methodologies. I deeply appreciate Prof. Rentzepis' investment in me and will fondly remember my work in his lab.

I owe special thanks to Prof. Michael Heller and Prof. Peter Asbeck. Prof. Heller's good-humored guidance and his enthusiasm for the field of bioengineering, as well as his vision on the importance of interdisciplinary education motivated me to explore the world of biological imaging. Prof. Asbeck, who guided me through my research on nanotechnology during my first year and a half at UCSD, taught me how to conduct deep and thorough analytical research. I am thankful for the long hours he spent with me on technical discussions and for his friendly demeanor.

I would also like to express my appreciation to all committee members – Prof. Esener, Prof. Lo, Prof. Heller, Prof. Subramanian and Prof. Mattrey for taking the time to serve on my committee and for their technical feedback and advice.

Much of the work presented in this dissertation is a direct result of valuable discussions with Dr. Matthias Gross and Dr. Uriel Levy. Through our discussions, Dr. Gross was able to extract me from many dead ends in my research, guided me through the tough physics obstacles, advised me on proper research methodologies and gave me precious advice, for which I am very grateful. Dr. Levy also provided me with professional advice, reviewed some of my papers and was a good friend.

I wish to thank some of my colleagues from the Esener research group. First and foremost I am indebted to Mark Hsu, my collaborator in the latter stages of this research. Many of the experimental results presented here are a result of Mark's support and dedication – I wish him the best of luck in the continuation of CMOS SPAD research. I also closely collaborated with Sanja Zlatanovic on the optical setup of many of the experiments described in this work. Sanja's patience and guidance

were invaluable to this work. Kai Zhao collaborated on the upconversion portions of my research, providing the InGaAs/InAlAs devices as well as helpful technical insights and discussion.

I am indebted to Dr. Anis Husain for focusing and helping me define the upconversion device to maximize end-user benefits. He also provided me with excellent project presentation skills and was crucial and getting funding for the upconversion project. I thank Ms. Priscilla Haase for the administrative support, and for helping get the characterization equipment on time for the various experiments.

Last but not least, I would like to thank my family. My wife, Sharon, whose moral support helped me materialize by long-held dream of pursuing a Ph. D.; my daughter, Shohum, whose lively presence and optimism put other problems in perspective, and my newborn son, Ohad, who slept just enough to let me complete this dissertation. I would also like to thank my parents for their continuous support through the years, and my parents-in-law for their dedication and sacrifice, which allowed me to conduct much of the research presented here.

This research was funded in part by the U.S. Army Research Office under Grant W911NF-05-1-0243. The test chips manufactured in this project were funded by the MOSIS Educational Program.

Chapter 3 and Chapter 4, in part, are a reprint of the materials presented in: H. Finkelstein, M. J. Hsu, S. Zlatanovic and S. Esener, "Performance trade-offs in single-photon avalanche diode miniaturization", submitted to *Optics Letters*; H.

Finkelstein, M. J. Hsu, S. Esener, “STI-bounded single-photon avalanche diode in a deep-submicron CMOS technology”, *IEEE Electron Device Letters*, vol. 27, no. 11, 2006; H. Finkelstein, M. J. Hsu and S. C. Esener, “An ultra-fast Geiger-mode single photon avalanche diode in 0.18 μm CMOS technology,” in Advanced Photon Counting Techniques, W. Becker ed., Proc. SPIE 6372, 63720W1-63720W10, 2006; and H. Finkelstein, M. J. Hsu, S. Esener, “A compact single-photon avalanche diode in a deep-submicron CMOS technology”, *Solid-State Devices and Materials*, Yokohama, Japan, 2006.

Chapter 5 is, in part, a reprint of the material as it appears in H. Finkelstein, M. Gross, YH Lo, and S. Esener, “Analysis of Hot-Carrier Luminescence for Infrared Single-Photon Up-Conversion and Readout”, to appear in *IEEE Journal of Selected Topics in Quantum Electronics – Single Photon Counting: Detectors and Applications*, 2007.

The dissertation author was the primary investigator and first author of these papers.

VITA

1970	Born, New Brunswick, NJ.
1994	B. Sc. in Electrical Engineering, Cornell University, Ithaca, NY.
1998	M. Sc. in Electrical Engineering, Technion – Israel Institute of Technology, Haifa, Israel.
2000	International Executive MBA, Northwestern University, Evanston, Il, and Tel-Aviv University, Tel-Aviv, Israel.
2007	Ph.D., Electrical and Computer Engineering (Photonics), University of California, La Jolla, CA.

PUBLICATIONS

H. Finkelstein, M. Benchimol, K. Zhang, YH Lo, S. Esener, “Experimental demonstration of single-photon frequency upconversion based on hot-carrier recombination,” in preparation for *Applied Physics Letters*.

H. Finkelstein, K. Zhang, YH Lo, S. Esener, “Absolute electroluminescence measurements of InGaAs/InAlAs avalanche photodiodes,” in preparation for *Applied Physics Letters*.

H. Finkelstein, M. J. Hsu and S. C. Esener, “Dual-junction single-photon avalanche diode,” submitted to *Electronics Letters*.

H. Finkelstein, K. Zhao, M. Gross, YH Lo, S. Esener, “Fast and power-efficient infrared single-photon upconversion using hot-carrier luminescence,” to appear in Quantum Communications and Quantum Imaging V, Proc. SPIE, 2007.

H. Finkelstein, M. Gross, YH Lo, and S. Esener, “Analysis of hot-carrier luminescence for infrared single-photon up-conversion and readout”, *IEEE Journal of Selected Topics in Quantum Electronics – Single Photon Counting: Detectors and Applications*, vol. 13, no. 3, 2007.

H. Finkelstein, K. Zhao, M. Gross, YH Lo, and S. Esener, “Fast and power-efficient infrared single-photon upconversion using hot-carrier luminescence”, to appear in Quantum Communications and Quantum Imaging V, Proc. SPIE, 2007.

H. Finkelstein, M. J. Hsu, S. Zlatanovic and S. Esener, “Performance trade-offs in single-photon avalanche diode miniaturization”, submitted to *Optics Letters*.

H. Finkelstein, YH Lo, S. Esener, “Frequency upconversion via hot carrier luminescence with single-photon sensitivity,” UC Docket No. SD2007-156.

H. Finkelstein, M. J. Hsu, S. Esener, “STI-bounded single-photon avalanche diode in a deep-submicron CMOS technology”, *IEEE Electron Device Letters*, vol. 27, no. 11, 2006.

H. Finkelstein, S. Esener, “Shallow-trench-isolation (STI)-bounded single-photon CMOS photodetector,” U.S. Provisional Patent Application, UCSD Attorney Docket No.: 15670-138P01 / SD 2006-228.

H. Finkelstein, M. J. Hsu and S. C. Esener, “An ultra-fast Geiger-mode single photon avalanche diode in 0.18 μm CMOS technology,” in Advanced Photon Counting Techniques, W. Becker ed., Proc. SPIE 6372, 63720W1-63720W10, 2006.

H. Finkelstein, M. J. Hsu, S. Esener, “A compact single-photon avalanche diode in a deep-submicron CMOS technology”, *Solid-State Devices and Materials*, Yokohama, Japan, 2006.

H. Finkelstein, P. M. Asbeck, and S. Esener, “Architecture and Analysis of a self-assembled 3D array of carbon nanotubes and molecular memories”, *IEEE Nano-2003*, San Francisco, 2003.

H. Finkelstein, R. Ginosar, “Front-side bombarded metal plated CMOS electron sensors,” in *SPIE Electronic Imaging Science and Technology*, M. M. Blouke ed., Proc. SPIE vol. 3301, 186-197, 1998.

H. Finkelstein, R. Ginosar, “Novel CMOS electron imaging sensor,” in *Advanced Focal Plane Arrays and Electronic Cameras II,* T. M. Bernard ed. Proc. SPIE Vol. 3410, 21-32 1998.

H. Finkelstein, “Frontside bombarded metal plated electron radiation imaging chip fabricated in CMOS technology”, M. Sc. Dissertation, Technion, Israel, 1998.

ABSTRACT OF THE DISSERTATION

Shallow-Trench-Isolation Bounded Single-Photon Avalanche Diodes in Commercial Deep Submicron CMOS Technologies

by

Hod Finkelstein

Doctor of Philosophy in Electrical Engineering
(Photonics)

University of California, San Diego, 2007

Professor Sadik C. Esener, Chair

This dissertation describes the first single-photon detection device to be manufactured in a commercial deep-submicron CMOS technology. It also describes novel self-timed peripheral circuits which optimize the performance of the new device. An extension of the new device for dual-color single-photon detection is investigated. Finally, an area- and power-efficient method for single-photon frequency upconversion is presented, analyzed, and experimentally examined.

Single-photon avalanche diodes have been used in diverse applications, including three-dimensional laser radar, three-dimensional facial mapping,

fluorescence-correlation techniques and time-domain tomography. Due to the high electric fields which these devices must sustain, they have traditionally been manufactured in custom processes, severely limiting their speed and the ability to integrate them in high-resolution imagers. By utilizing a process module originally designed to enhance the performance of CMOS transistors, we achieve highly planar junctions in an area-efficient manner. This results in SPADs exhibiting high fill factors, small pitch and ultrafast operation. Device miniaturization is accompanied by excessive noise, which was shown to emanate from trapped avalanche charges. Due to the fast recharging of the device, these charges are released in a subsequent charged phase of the device, causing correlated after-pulses. We present electrostatic and electrical simulation results, as well as a comprehensive characterization of the new device. We also show for the first time that by utilizing the two junctions included in the device, we can selectively detect photons of different wavelengths in the same pixel, as is desirable in cross-correlation experiments.

This dissertation also describes an efficient new method for single-photon frequency upconversion. This is desirable for applications including quantum-key distribution and high-resolution near-infrared imaging. The new technique is based on electroluminescence in or near the multiplication region of the device, resulting from hot-carrier recombination. We model a proposed hybrid device and deduce the critical parameters for efficient upconversion. Lastly, we experimentally demonstrate that the electroluminescence yield from an InGaAs/InAlAs avalanche diode is sufficient for highly-efficient upconversion.

1. INTRODUCTION

1.1. The Power of CMOS in Single-Photon Detection

Single-photon detection has gained increased relevance in recent years. It has found uses in diverse areas, including single-molecule dynamics [1], quantum communications [2, 3], military [4] and medical imaging [5], and biometrics [6]. Single photon detection is unique in that, unlike traditional image sensors which provide an aggregate temporal average of impinging photons, it provides information regarding individual photon-arrival events. Consequently, several types of data can be obtained from such detectors, as follows:

- A binary output corresponding to a photon arrival event, within the exposure window of the detector.
- Time density of photon arrival events, corresponding to the instantaneous photon flux.
- Photon times-of-arrival statistics.

Single-photon detection has been achieved using several types of devices, as will be discussed in this chapter. The device that has become mainstream for this purpose is the single-photon avalanche diode (SPAD). However, due to its special operating mechanism, which is based on extremely high electric fields and high instantaneous currents, SPADs have traditionally been manufactured using processes which were specially tuned for this purpose. These processes offered many advantageous benefits, such as high detection efficiencies, low

jitter and desirable spectral response, especially in longer wavelengths. In this dissertation, we argue that the manufacturing of SPADs in leading-edge commercial Complementary Metal-Oxide-Semiconductor (CMOS) technologies offer significant advantages, despite the fact that these technologies are generic in nature, and are not optimized for SPADs.

The argument for the power of CMOS for SPAD devices relies on a general trend towards generic semiconductor technologies, which has been taking place during the past three decades. During the early days of the semiconductor industry, each semiconductor device manufacturer required its own fab, with a process specifically tuned to its products [7]. When TSMC introduced the concept of the silicon foundry, many device manufacturers streamlined their designs to the available process parameters, even if those were not optimized for their specific requirements. The benefits for doing so were the immense cost savings of not maintaining an in-house fab, economies of scale and the quality benefits resulting in high-volume production. During the 1990's, many applications requiring "analog" processes, such as high-speed analog-to-digital converters and phase-locked loops (PLLs) have also migrated to standard digital CMOS, compensating for the sub-optimal processes by using innovative advanced circuit techniques.

This trend has continued in recent years, when the benefits of high-speed III-V processes (such as SiGe) have often been challenged by the cost savings of RFCMOS processes [8]. Although the latter are not optimized for RF

performance, the availability of ultra-dense CMOS circuitry to augment the analog device performance, has often proven decisive in the choice for RFCMOS.

In the field of image processing, a similar trend has also taken place. Traditionally, the charge-coupled device (CCD), first commercialized by AT&T, Fairchild Semiconductor, RCA and Texas Instruments, held a virtual monopoly on solid-state imaging. These devices offered superb noise, dynamic range and responsivity, due to their optimal design, low leakage and high uniformity. However, since the 1990's, CMOS active-pixel sensors (APS) have gained ground and have become the technology of choice for many imaging applications, despite their initial inferior performance [9]. This was due to the availability of CMOS transistors, which allowed one to compensate for the physical deficiencies of the device, and to take advantage of miniaturization and reduced power, which were not available in CCD lines. By riding the cost-reduction bandwagon which was driven by the overall microelectronics industry, APS were able to offer a more compelling cost-performance trade-off compared with the custom CCD technologies.

In this dissertation we will demonstrate that a similar trend is desirable for single-photon detectors. Once an appropriate SPAD device (such as the one proposed herein) can be implemented in an advanced CMOS technology, unprecedented performance and new applications can be achieved, and the technology can be commercialized successfully. This, despite the inherently

inferior performance of the basic device in CMOS compared to implementation in a custom-designed process. If successful, fully-integrated CMOS SPAD imagers can be the technology of choice for high-resolution fluorescent-microscopy systems, for tomography systems which do not pose a health risk to patients, in 3D cameras, and in highly parallel secure optical communication systems.

In this chapter, we will describe some of the exciting applications using single-photon detectors and we will summarize the state-of-the-art in single-photon detection. With a grasp of the requirements posed by these applications as well as an understanding of the existing detectors, we will be able to set the groundwork and formulate the motivation for this work.

1.2. Applications of Single-Photon Detectors

1.2.1. Ladar and Long-range Three-Dimensional Imaging

Optical ranging offers significant advantages over traditional radio-frequency ranging. Due to the shorter optical wavelengths, laser radars (ladars) can resolve fine features of its targets, rather than just detect them [10, 11]. A ladar system consists of a coherent light source which is either pulsed or modulated, illuminating a target with a narrow beam. Usually, a scanning system focuses the returning photons from different directions onto a single SPAD pixel in rapid succession, measuring the

photons' time of flight. Alternately, the returning photons are focused onto a focal-plane array of SPADs [10, 11]. Several closely related techniques allow for the collection of photons arriving within a limited time-gate, thus forming an image from otherwise-obscured targets [12].

The main challenges facing single-photon detectors for ladar applications are:

- Maximizing detection efficiency: since targets are typically tens of kilometers away and may even be located in space, the number of photons returning to the detectors is very small.
- Operating in daylight with an extremely low signal-to-background: several methods have been demonstrated whereby detection in daylight was made possible through time-correlation and frequency-multiplexing [13].
- Attaining sufficient range precision: range precision is attained by minimization of the time-uncertainty (jitter) in the time-of-arrival measurement. As will be explained in Section 2.8.3, there is a physical trade-off between detection efficiency and jitter.
- Processing the time-of-arrival in real-time: assuming a 10 cm precision is desired within a distance range of 1 km (from minimum to maximum expected target range), 14 bits of information are generated per pixel per measurement. Assuming a 640 x 480 pixel array, a 30 Hz frame rate and a probability of 1% to detect a returning photon for each transmitted pulse, 12.9 gigabits (Gb) of information must be processed every second.

1.2.2. High-resolution Three-Dimensional Imaging

Three-dimensional non-stereoscopic imaging has been demonstrated using active-illumination and an array of SPADs [6]. Unlike ladars, where moderate range resolution is sufficient, three-dimensional imaging required sub-millimeter range precision. In order to attain a 0.8 mm depth resolution, the time uncertainty must be reduced to 5.3 ps. This figure includes the jitter of the pulsed light source (specifically, of its rising edge), that of the SPAD, and the uncertainty introduced by the time-to-digital converter.

Since present detectors cannot achieve such precision, one must over-sample the target. Assuming that the jitter has a Gaussian distribution, over-sampling will reduce the variance by the square-root of the number of samples. In one implementation, 10,000 exposures were required to achieve the desired precision [6]. For a 1 megapixel imager, with a depth-of-field of 25 cm, this translates into 90 Gb of data which must be processed per image.

Due to the large amount of data that must be processed, such applications are amenable to implementations in deep submicron commercial CMOS technologies, such that processing can be performed on chip. Other challenges lie in the choice of a light source, which must have a fast rise time, on the order of a few picoseconds, and a fast repetition rate, on the order of a few nanoseconds, to enable realistic data acquisition times.

1.2.3. Single Photon Detection for Molecular Imaging and Spectroscopy

Single-molecule spectroscopy allows us to observe the behavior of individual molecules which are surrounded by a multitude of surrounding molecules, by using tunable optical radiation [14]. To probe the molecule, a light beam, typically from a laser source, is used to pump an electronic transition specific to the target molecule. The resulting optical transition is detected. The indirect detection of this absorption, via fluorescence, is of special interest in this work. This detection must be accomplished in the presence of billions to trillions of solvent or other surrounding molecules, and in the presence of noise from the measurement setup itself. In the past decade, the field of single molecule spectroscopy has become a powerful technique for exploring the behavior of molecules in complex local environments, such as their transport across cellular barriers [15, 16], inter-molecular dynamics [17] and protein folding [18-20].

Unlike classical measurement techniques, which yield observations on ensemble averages of a population, single molecules spectroscopy gives access to the full distribution of an observable, and specifically to the dynamic heterogeneities of a sample. Consider a population of molecules undergoing asynchronous changes, such as the folding of a protein molecule. In order to characterize this folding behavior, the following must be ensured [1]:

(a) Only a single molecule (or at most a few molecules) is observed at a given instant.

(b) The total observation time of this single molecule is substantially longer than the dynamic of interest.

(c) The measurement's time resolution is shorter than the typical rate of change.

These three conditions are required to ensure the maximization of two important parameters – Signal-to-Background ratio (SBR) and Signal-to-Noise ratio. The signal received from a fluorescent molecule can be calculated by modifying the expression developed by [1] to:

$$s = \frac{P}{A \cdot h\nu} \times \sigma Q \times E \times \frac{\tau}{\tau + t_d}$$

Equation 1.1

where P is the exciting beam's power; A its area, $h\nu$ the photon energy; σ the molecule's optical cross-section; Q the emission quantum yield; E the global collection efficiency of emitted photons, which includes optical collection losses and detector efficiency; τ the exposure time and t_d the device dead time. The first terms accounts for the impinging photons, the second terms describes the probability of photon absorption and fluorescence by the molecule, the third term considers the probability of detecting the fluorescence signal, and the last term takes into account the probability of a collected photon impinging during the device dead time.

The environment of the single molecule often adds a background photon-flux contribution which can be modeled by a rate b per unit-volume and unit-power. This can occur, for example, as a result of Rayleigh scattering or due to substrate auto-

fluorescence. The signal-to-background ratio (SBR) for an observation through an excitation-detection intersection volume V can be calculated as:

$$SBR = \frac{E\sigma Q}{bV \cdot A \cdot h\nu}$$

Equation 1.2

The detector will also contribute a dark count rate d. The statistical behavior of this dark count depends on its physical mechanism and will be discussed in the next chapter. For our present discussion, we shall assume that both the dark count rate and the background flux are Poisson-distributed processes, and as such their mean is also their variance [21]. This assumption does not hold if afterpulsing becomes dominant. If the noise sources are independent (i.e., afterpulsing is negligible) then the signal-to-noise ratio can be expressed as:

$$SNR = \frac{s}{\sqrt{s + bVP + d}} \times \sqrt{\frac{\tau}{\tau + t_d}}$$

Equation 1.3

Equation 1.1 is significant because it quantifies the trade-offs between required signal intensity, the device dead time and the dark count rate.

1.2.3.1. Fluorescence Correlation Spectroscopy

The technique of fluorescence correlation spectroscopy (FCS) was developed to measure chemical kinetics of molecular diffusability by analysis of samples of very-

low-concentration solutions, and has become a powerful tool in biophysics [22, 23]. The technique makes it possible to measure diffusion coefficients, chemical rate constants, molecular aggregation and rotational dynamics of one or more species [24]. It has also been used to measure intracellular pH, ligand-receptor and peptide-liposome interaction and for single-molecule DNA sequencing [25, 26]. Furthermore, FCS can quantify the temporal evolution of concentration fluctuations via statistical monitoring of the fluorescence signal.

FCS is based on two related physical concepts, known as the Onsager Hypothesis [27] and the Fluctuation-Dissipation theory [23]. Onsager's Hypothesis states that "*the regression of microscopic thermal fluctuations at equilibrium follows the macroscopic law of relaxation of small non-equilibrium disturbances.*" This can be explained as a general observation that the behavior of a macroscopic system in response to a force can be inferred from the response of its microscopic components to instantaneous perturbations from equilibrium.

FCS utilizes this principle by observing the semi-random motion of particles in solution (Figure 1.1). By calculating the auto-correlation of fluorescence signals from molecules entering and leaving a small excitation-observation volume, it is possible to infer some important parameters of the species of interest.

If the fluorescence signal at time t, $F(t)$, has a mean value $\langle F(t) \rangle$, then the fluorescence fluctuation is:

$$\delta F(t) = F(t) - \langle F(t) \rangle$$

Equation 1.4

The normalized autocorrelation of this observation is:

$$G(\tau) = \frac{\langle \delta F(t) \delta F(t+\tau) \rangle}{\langle F(t) \rangle^2}$$

Equation 1.5

As an example of the data that can be extracted from FCS, consider a confocal scheme, where the observation volume is an ellipsoid. In this case [23]:

$$G_D(\tau) = \frac{1}{\left[N\left(1 + \tau/\tau_D\right)\left(1 + \tau/\omega^2\tau_D\right)^{0.5} \right]}$$

Equation 1.6

Here, τ_D is the characteristic diffusion time during which a molecule resides in the observation volume with an axial z_0 to lateral r_0 dimension ratio $\omega_0 = z_0/r_0$. As τ is minimized, N can be determined, and from there τ_D. This is true for the simple case of a single species, but can easily be extended to multiple species.

Regardless of the number of species, it should be noted that for FCS to be effective, a number of conditions must be ensured:

- A very low sample concentration, such that the probability of detecting a fluorescent signal is very low. This ensures that the statistical signal fluctuation is

not drowned in the background-averaged signal. Typical setups target a detection probability of approximately 0.1% per detection time window.

- Maximal detection efficiency in the relevant wavelength.
- Minimal time between measurements, so that the limit of $\tau/\tau_D << 1$ can be attained.
- Time uncertainty that is much less than τ_D.

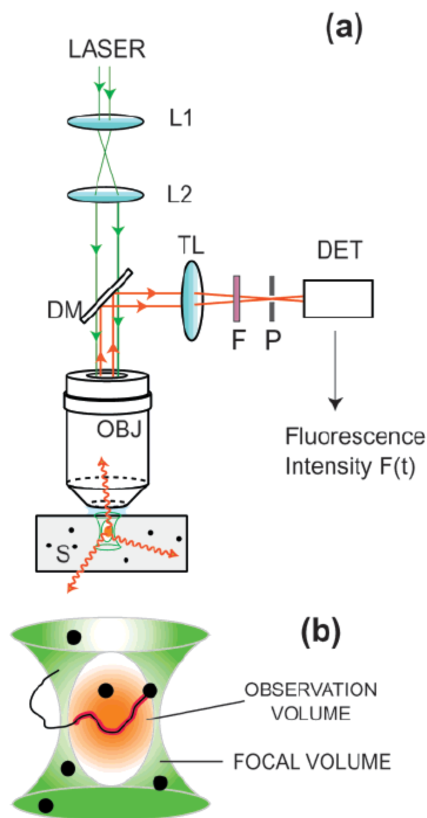


Figure 1.1: (a) Optical setup for FCS. A Laser illuminates a sample volume. The observation volume is imaged onto a detector. (b) an illustration of the observation volume from which fluorescent particles are imaged [23].

Clearly, there is a trade-off between the inter-exposure time (dead time) and the detection efficiency. If photons are emitted at short time intervals but the detector,

even while having a high detection efficiency, suffers from a long dead time, then many of the emitted photons will be lost. Conversely, a fast detector with a short dead time but low detection efficiency, will also lose many of the photons.

This trade-off is of special importance in this work and has been numerically simulated using Matlab. For the simulation, a circular two-dimensional space has been taken to be the observation volume (Figure 1.2 (a)). Diffusion times were varied by changing the number of random steps the particle carries out between samples. Detection efficiencies were accounted for by using a random weight for the probability that a measurement is recorded. Autocorrelation computation was performed on the binary data corresponding to the detector output (Figure 1.2 (b)).

Simulation results (Figure 1.2 (c)) indicate that with 100% detection efficiency, diffusion velocities of 5, 10 and 20 can be easily discerned through their autocorrelation plots. When the detection efficiency is reduced to 10%, correlation is lost quickly and the resolution deteriorates significantly. However, if we assume that the detector has a 10% detection efficiency but is ten times faster than the previous cases, and that the photon flux is sufficiently high (yellow plot), it is again possible to attain a smooth autocorrelation curve.

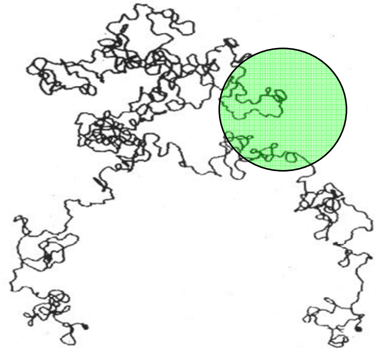
These results agree with analytical results, which derive a figure of merit *F*, which compares the signal-to-noise ratio of the lifetime measured with an ideal recording device to the actually achieved lifetime signal-to-noise ratio for a given number of recorded photons [28]. We can expand the standard definition of *F*, taking into account the detection efficiency η to get:

$$F = \frac{SNR_{ideal}}{SNR_{real}} = \sqrt{1 + r_{ph} \cdot t_{dead}} = \sqrt{1 + \eta \cdot r_{sig} \cdot t_{dead}}$$

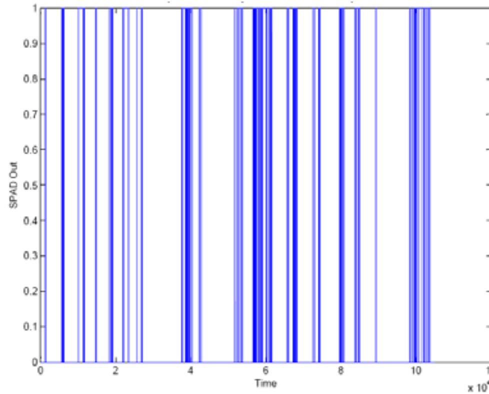
Equation 1.7

Here, r_{sig} is the rate of photon arrivals, r_{ph} is the rate of recorded arrivals of photons by the detector and t_{dead} is the dead time of the detector.

Equation 1.7 captures the importance of reducing dead times for those measurements where the photons' arrival rate is high. In fact, a reduction of t_{dead} is equivalent to an increase in detection efficiency by the same factor.

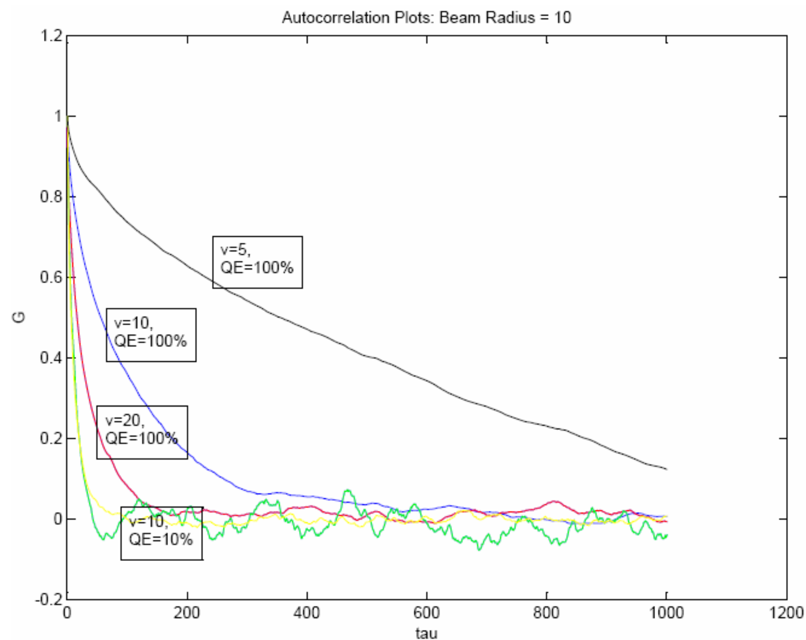


(a)



(b)

Figure 1.2: (a) Illustration of scheme used for FCS simulation. (b) Output of Matlab simulation showing sample signals from the detector.



(c)

Figure 1.2 (cont.): (c) Autocorrelation plots for various particle diffusion velocities and detection efficiencies (QE). The bottom green and yellow plots both correspond to v=10 and QE=10%. The latter assumes ten-times more samples than all other plots.

One area in which FCS has proven to be a powerful technique is the study of protein folding [15, 18]. Folding often occurs in time scales of a few microseconds. Because folding cannot be synchronized and is statistical in nature, protein folding cannot be observed when a large number of molecules occupy the field of view. Therefore, a very dilute solution is used, in the nanomolar range. Förster resonance energy transfer (FRET) between two chromophores attached to the polypeptide chain is then used to investigate the sub-microsecond folding dynamics [29]. Because of the FRET dynamics, an autocorrelation computation of the fluorescence of both the acceptor and donor wavelengths reveals three distinct populations: folded, unfolded

and missing-acceptor. By studying the autocorrelation function, a typical molecular reconfiguration rate can be calculated. Appropriate labeling of different domains in the protein can allow for a 3D mapping of the molecule as it folds.

1.2.3.2. Fluorescence Cross-Correlation Spectroscopy

An important extension of the fluorescence correlation technique is dual-color fluorescence cross-correlation spectroscopy (FCCS) [30]. This technique requires two spectrally-separated fluorescent particles, which label two interacting molecules of interest. The fluorescence emissions from the two dyes are separated by a dichroic mirror and directed onto two detectors (Figure 1.3). In addition to the standard autocorrelation curve obtained for each channel, the cross-correlation between the fluorescence signals is also computed. The cross-correlation between a red and green channel is given by [15]:

$$G_x(\tau) = \frac{\left\langle \delta F_{red}(t)\, \delta F_{green}(t+\tau) \right\rangle}{\left\langle F_{red}(t) \right\rangle \left\langle F_{green}(t) \right\rangle}$$

Equation 1.8

When only one species is present, the value of the above expression approaches 0. When the two fluorophores are fully-bound, the cross-correlation takes the value of the autocorrelation of the bound particle. A partially-bound solution will exhibit a cross-correlation amplitude which depends of the fraction of bound particles.

Therefore, cross-correlation is well suited for the analysis of binding of two differently-labeled molecules.

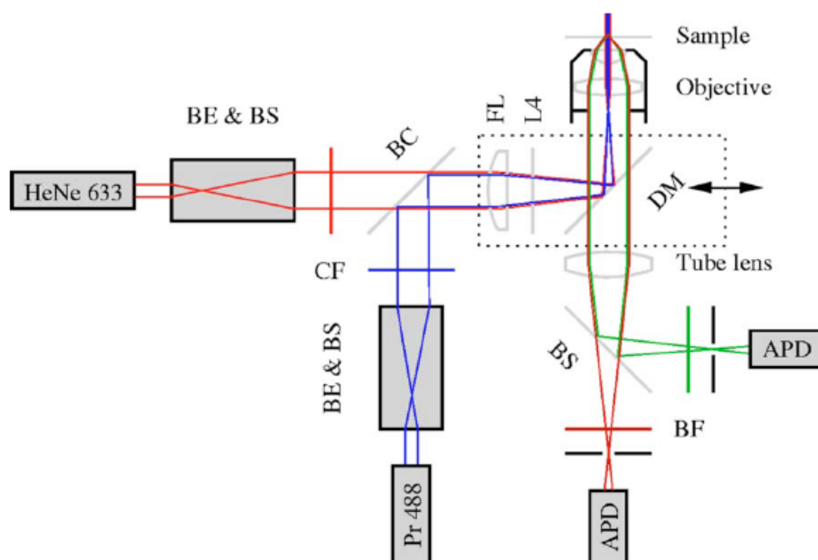


Figure 1.3: Dual-color FCCS setup [31]. Two laser beams are directed through a beam expander (BE), beam stirrer (BS) and cleaner filter (CF) onto a dichroic beam combiner (BC). The combined beams are focused through ha focusing lens (FL) and a quarter-wavelength plate (l4) oriented for maximum fluorescence onto the sample. Fluorescence light is passed through a dual-band dichroic mirror (DM) and separated by a beam splitter (BS) to two channels. Each is detected by an avalanche photodiode.

1.2.3.3. Fluorescence Lifetime Imaging

The acquisition of photon times-of-arrival can be extended to two dimensions using a technique known as Fluorescence Lifetime Imaging Microscopy [32]. By monitoring the fluorescence lifetime in different pixels across a focal-plane array, it is possible to differentiate between background fluorescence and signal fluorescence, map out the local environment of a molecular probe, monitor the presence or absence

of a species by monitoring its effect on the lifetime of a molecular probe, or to differentiate between species [1].

Several FLIM implementations have been demonstrated. In Time-Correlated-Single-Photon-Counting, a sample is raster-scanned with a pulsed laser [28]. One or more detectors record the time-of-arrival of photons as well as the x and y coordinates of each detection event. This information is then processed resulting in decay times per pixels. These are then converted into a color-coded image.

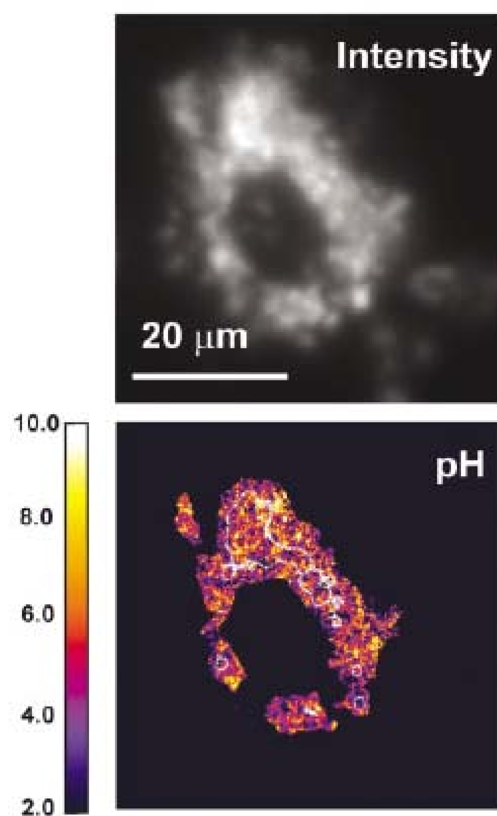


Figure 1.4: In vitro fluorescence (top) and pH-dependent lifetime-shift image (bottom) of a stained fibroblast [33].

Hardware requirements from TCSPC are a major obstacle to obtaining reliable high-resolution images. For a double-exponential fit, 10,000 to several 100,000's of

photons are required per pixel [28]. Each arriving photon's time of arrival must be determined through a time-to-analog converter, and stored, together with its pixel coordinates. This results in acquisition times of up to 30 minutes for even a moderate resolution of 128 x 128 [34].

1.2.4. Time-Domain Diffuse Optical Tomography

Time-domain diffuse optical tomography (DOT) uses near IR light to determine the three-dimensional structure of highly-scattering samples. Some application of DOT include optical mammography [35] and brain imaging [5], though this tool is still in its infancy.

Light is strongly scattered in living tissue [36]. Typical scattering coefficients for visible wavelengths are around 10 cm^{-1}. Consequently there are almost no ballistic (non-scattered) photons in tissues thicker than 1 cm. Instead, photons are said to "diffuse" through the tissue, thereby deteriorating their spatial resolution. NIR absorption in tissue is dominated by oxy-haemoglobin, deoxy-haemoglobin, lipids, and water [37]. These have an absorption minimum in the wavelength band between 650 nm and 900 nm. This band can be used to investigate tissues up to 10 cm thick.

In order to analyze the composition of the sample, we note that the effects of absorption and scattering can be discriminated. Whereas absorption lowers the intensity of the optical signal, scattering smears it in the time domain, due to the wider

distribution of optical paths traversed by the photons. This is the key to analyzing DOT data.

Two DOT setups are commonly used [38]. In a transillumination technique, a source and detector are scanned across opposite sides of the specimen. The resulting series of time and amplitude measurements are used to form a 3D model of the specimen. In the tomographic technique, multiple sources and detectors are placed over the available surface of the tissue, sampling multiple lines-of-sight, resulting in a reconstructed 3D image.

This field, which is rapidly evolving, requires high count rates, low timing jitter and high detection efficiencies for attaining acceptable depth resolutions in a timely manner.

1.3. State of the Art in Single-Photon and Low-Light-Level Detection

The applications reviewed in the previous section have a number of common constraints – they require the sensing of the intensity as well as times-of-arrival of a small number of photons, often within short-duration bursts. There are three strategies for dealing with this challenge [1]:

1. Convert the photons to charge and accumulate this charge over a sufficiently-long time to overcome the various noise sources.
2. Amplify the signal as early as possible in the signal-processing chain in order to maximize the signal-to-noise ratio.

3. Detect each impinging photon individually and record its arrival time.

In the next subsections, we will describe several techniques using these strategies and emphasize their fit to the applications described above.

1.3.1. Photomultiplier Tubes and Micro-Channel Plates

A photomultiplier tube (PMT) is a vacuum tube which is comprised of the following components (Figure 1.5) [39]:

- An input window
- A photocathode
- Focusing electrodes
- An electron multiplier
- An anode

The PMT operation is based on the *external* photoelectric effect, whereby an electron in the valence band absorbs a photon of sufficient energy and is emitted into the vacuum. The photocathode, from which the photoelectrons are emitted, is usually made from an alkali metal or from a III-V compound, such that its work function is lower than the energy of the impinging photons. The work function is defined as the energy difference between the Fermi level of the material and the vacuum level.

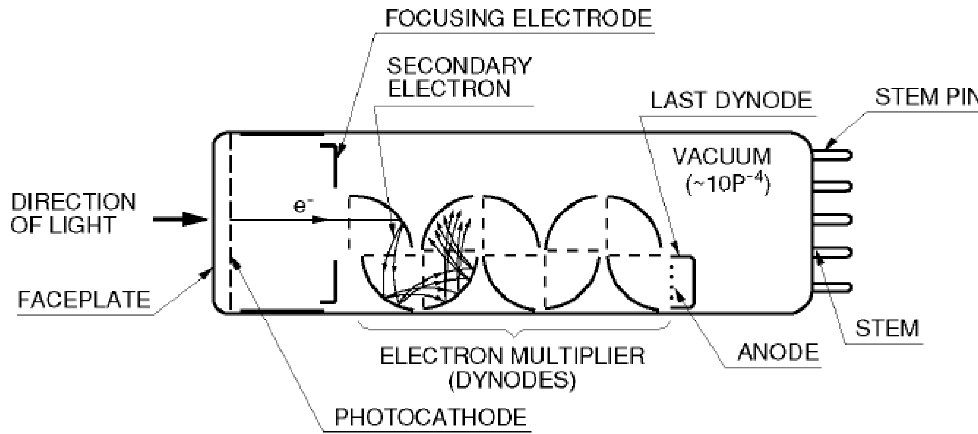


Figure 1.5: Cross-section of a photomultiplier tube [39].

One of the main metrics for evaluating PMT performance is its efficiency in converting impacting photons to detectable electrons. This quantum efficiency is expressed by the following expression [39]:

$$\eta(\nu) = (1-R)\frac{P_\nu}{k} \cdot \left(\frac{1}{1 + \frac{1}{kL}} \right) \cdot P_s$$

Equation 1.9

where ν is the photon frequency; R is the reflection coefficient of impinging photons from the photocathode surface; k is the absorption coefficient of the photons; P_ν is the probability that an absorbed photon will excite an electron to the vacuum level; L is the mean escape length of an excited electron before it is recaptured in the bulk; and P_s is the probability for an electron which has reached the photocathode surface to be released into the vacuum. In an optimized PMT, the limiting parameter will be L. It can be maximized by reducing the number of defects in the photocathode crystal.

Once released into the vacuum, electrons are accelerated in an electric field, impacting a series of positively-charged dynodes (Figure 1.5). When a primary electron with initial energy strikes the surface of a dynode, δ secondary electrons are emitted. For a PMT with n dynode stages, an overall gain of δ^n results. Dynode design determines to a large extent the timing spread of impacting electrons on the anode, and as such is often the limiting factor for PMT jitter.

The last electrode in the PMT is the anode. It converts the electron cloud to an electrical signal. Once charge impacts the anode, it must be quickly removed in order to prevent surface-charge effects. This time will determine the minimum time interval between counts.

In most PMTs, the optimal detection efficiency is approximately 20% in the visible range and falls to less than 1% in the near-IR. With III-V photocathodes such as GaAsP, quantum efficiency peaks at 45% at 600 nm, while with GaAs a maximum detection efficiency of 30% is achieved at 900 nm [40].

Microchannel plates (MCPs) can replace the dynodes in MPT to create MCP-PMTs. MCPs are secondary multipliers consisting of an array of millions of glass capillaries fused to form a thin disk. When an electron enters a channel, it hits the channel wall, producing secondary electrons. These are then accelerated by the electric field and strike the opposite wall to produce additional secondary electrons. As a result of these collisions, gains on the order of 10^4 can be attained for a single-stage MCP [41].

Due to the spatial confinement of each photo-generated electron cloud, MCPs achieve good spatial resolution and excellent timing jitter. Detection of the resulting electron cloud can be through a phosphor screen, sometimes imaged by a CCD or CMOS device, or using a cross-delay line anode [42].

A small amount of current flows in the photomultiplier tube even when operated without the presence of photons. This is termed the dark current. It may arise as a result of the following mechanisms:

- Thermionic emission from the cathode and dynodes.
- Field emission current.
- Ionization current from residual gases.
- Noise current caused by cosmic rays and radiation from glass contaminants.

As the supply voltage is increased, both the quantum efficiency and the PMT noise increase.

Today, PMTs are one of the main technologies of choice for single detection due to their maturity and wide availability. However, their fragility, lack of scalability and weak spectral response in the near IR open the way for new technologies to be used.

1.3.2. Superconducting Wire Detector

In recent years, a new technology was developed for single-photon counting, based on NbN superconducting nanowires [43]. Each detector element consists of a

superconducting nanowire which is cooled to 2-4 K, which is below the critical superconductivity temperature of NbN. The wire is biased with a current density slightly below the critical current density, beyond which its superconductivity breaks down. When a photon is absorbed in the nanowire (Figure 1.6), resulting in a localized heat spot. This spots increases the local resistivity, thereby giving rise to a voltage drop which can be sensed.

This technique does not suffer from many of the problems of PMTs – specifically there is no leakage current associated with the device. Moreover, it can achieve timing precision of 50 ps, which is comparable to the best detection technologies. However, the inter-event interval is limited by the cooling of the hot spot, resulting in a physical limit on the reset time. Moreover, the only known technique for patterning these structures is with an electron beam – a method which is serial in nature and is therefore not amenable to large scale integration.

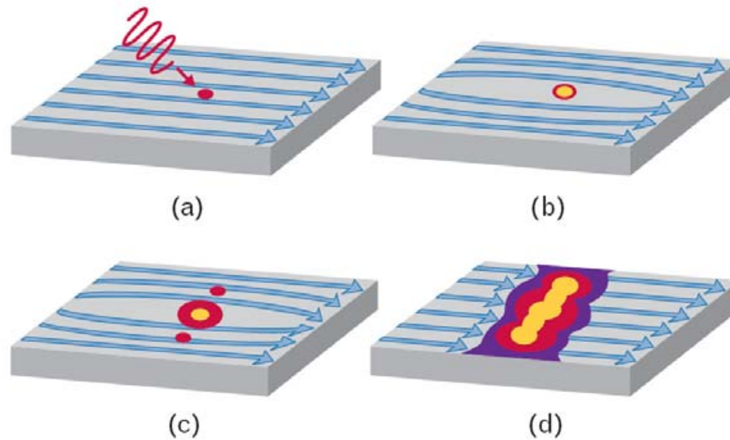


Figure 1.6: Operation of a superconducting nanowire photon counter [43]. (a) A photon impacts a nanowire carrying a direct current near the critical superconducting current. (b) A local hot spot forms, diverting current around it. (c) Current circumvents the hot spot. (d) A voltage develops across the resistive region, and is detected electrically.

1.3.3. Charge-Coupled Devices

1.3.3.1. The Standalone Charge-Coupled Device

Charge-coupled devices (CCDs) are silicon-based integrated circuits consisting of a dense matrix of photodiodes that operate by converting photons into an electronic charge. During an integration phase, photogenerated electrons are stored in a potential well, from which they are subsequently serially transferred across the chip through registers and output to an amplifier.

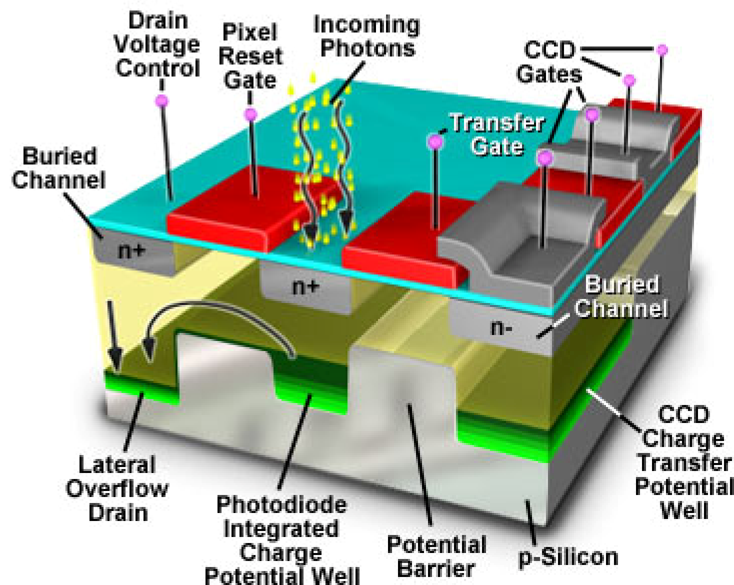


Figure 1.7: Cross-section of a CCD pixel [44].

Referring to Figure 1.7, a pixel reset gate initially empties the photodiode’s potential well. It then disconnects the well from the supply, letting it “float”. Incoming photons generate photoelectrons which are stored in the potential well. A transfer gate applies a voltage such that a potential barrier blocks stored charges from overflowing

to neighboring pixels. Once the exposure, or charge-integration phase, is complete, the transfer gate lowers the potential barrier, and the accumulated charge is shifted serially until it reaches an analog-to-digital converter, where it is measured.

The main noise sources in CCDs are [45]:

- Shot noise (photon noise): Shot noise arises from the quantum nature of light, which results in a Poissonian time-distribution of arriving photons.
- Dark current noise: Dark current arises from thermally-generated charges near the photodiode's depletion region. While the average dark current can be subtracted as a background signal, the random fluctuations of this noise cannot be removed.
- Photoelectron noise: This noise arises from the statistical nature of carrier generation.
- Readout noise: The amplifier noise, which includes quantization granularity as well as electrical noise, becomes dominant for short acquisition times, or when the total collected charge is very small.

The effect of shot noise on the signal-to-noise ratio can be reduced by increasing the integration time as well as the quantum efficiency of the detector. The effect of readout noise can be reduced by increasing integration time and by optimizing the readout electronics. Dark current can only be reduced by using high-quality silicon or by cooling the detector. For this reason, high sensitivity CCDs are cooled, resulting in dark current figures down to a few electrons per second.

A recent enhancement to the traditional CCD is the Electron-Multiplying CCD [46]. This device overcomes the noise introduced by the charge-to-voltage conversion

process by adding a low-noise charge-multiplication stage before this conversion, resulting in noise figures similar to those of intensified CCDs (see below) but without compromising spatial resolution or cost. However, since there is no gain process in the pixel itself, this application is not applicable for collecting single-photon events.

1.3.3.2. Intensified Charge-Coupled Devices

The CCD device in itself has no gain. Some applications such as low-light-level imaging, require real-time data acquisition even when photon flux is low, and therefore require a gain stage. This is accomplished by coupling an image intensifier to the CCD through a phosphor screen and a fiber-optic bundle [47].

The main advantages of the ICCD are:

- Improved sensitivity, down to single photon events, with an improved signal-to-noise ratio compared with the same CCD alone (some high-end cooled CCDs have better signal-to-noise ratios than ICCDs) [47].
- Ability to retrieve event-time information by time-gating the MCP voltage, down to approximately 25 ns [48].

However, these devices are also affected by some of the problems of both the PMT and the CCD, namely:

- MCP noise at high gains
- Reduced MCP sensitivity at near-IR
- Variable MCP sensitivity at short time-gates

- Relatively low spatial resolution, limited by the MCP resolution as well as the number of fibers in the bundle connecting the MCP and CCD
- Fragility, bulkiness and high-voltage requirements

1.3.3.3. Electron-Bombarded CCD

In an electron-bombarded CCD (EBCCD), gain is attained via numerous impact ionizations in the silicon. The device is encased in the same vacuum chamber as a photocathode and in effect serves as the anode in a PMT. The energetic photo-electrons strike the back side of the CCD, creating one electron-hole pair for each 3.6 eV of incident energy [49]. The resulting charges are directed via an electric-field gradient, and collected in the well of a CCD and processed as usual. In order to collect the maximum number of charges, the substrate must be sufficiently thinned, yet remain thick enough for multiplication to take place and for the electrons to lose their energy before causing damage to the sensitive gate-oxide in the front of the device.

The devices can be operated at video frame rate, but have limited gain adjustment range, and exhibit similar disadvantages to the intensified CCD, including reduced quantum efficiency and resolution.

1.4. Single-Photon Avalanche Diodes

An avalanche photodiode absorbs incoming photons and converts them to a cascade of charge-carrier pairs. The gain in these devices is attained by an electrostatic field which accelerates the photogenerated charge carriers, resulting in a chain reaction of impact ionizations. The physical processes responsible for photon absorption and charge multiplication will be discussed in chapter 2. In the next sub-sections we will describe the state-of-the-art in avalanche photodiodes (APDs) and in Geiger-mode single-photon avalanche diodes (GM-SPADs).

1.4.1. Sub-Geiger APDs

When an avalanche photodiode is biased just below its reverse breakdown voltage, it is said to be operating in a sub-Geiger mode. A cross-section of one APD implementation, with a separate absorption and multiplication regions (SAM) is shown in Figure 1.8. Incident photons enter through a passivation layer, used for mechanical and chemical isolation of the device, as well as any anti-reflection coating deposited to improve light coupling into the device. They then traverse through the substrate until some are absorbed by the silicon, most likely in the thick, low-doped π region. The absorbed photons cause the release of an electron-hole pair which are accelerated by the high electric field formed between the anode and cathode. The

number of electrons generated in the p layer continues to increase as they undergo multiple collisions with the crystalline silicon lattice. This avalanche of electrons eventually results in electron multiplication that is analogous to the process occurring in one of the dynodes of a photomultiplier tube. The resulting current is then amplified using a sensitive high-speed amplifier, generating an output signal which is proportional to the number of absorbed photons.

Noise in APDs depends on the same mechanisms discussed in sub-section 1.3.3.1 but is further complicated by the gain mechanism which amplifies the shot noise, photoelectron noise and the dark current noise.

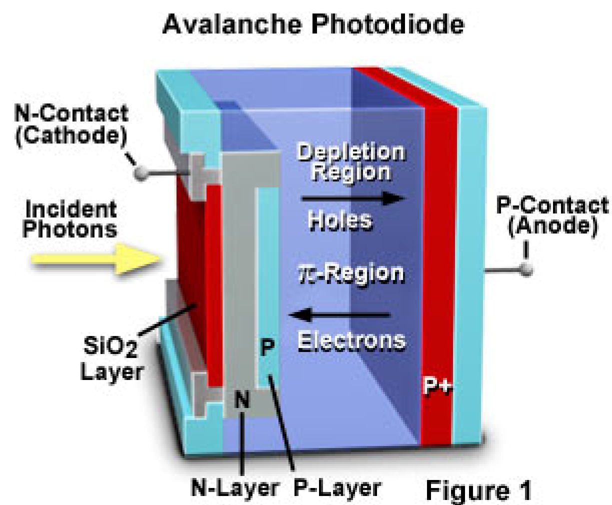


Figure 1.8: Cross-section of an avalanche photodiode [44].

1.4.2. Geiger-Mode Single-Photon Avalanche Diodes

Geiger-mode SPADs were pioneered by Sergio Cova in the 1980's [50] based on research from the early days of semiconductor research at Bell Labs and later at Shockley Laboratories on the avalanche behavior of semiconductor junctions [51-57]. In Geiger mode, the pn junction is biased above its breakdown voltage. At the absence of free carriers nothing happens. However, as soon as a free carrier enters the high-field region, it is accelerated by the electric field, until it collides with the lattice, resulting in impact ionization and the release of an electron-hole pair. These are accelerated in opposite directions, collide with the lattice and release additional carriers in a chain reaction. By definition, above the breakdown voltage, carriers are generated in the junction faster than they are extracted. Therefore, the avalanche must be quickly quenched in order to prevent heating and irreversible damage to the junction.

The interaction between device structure and its performance will be discussed in detail in chapter 2. SPAD devices are designed to optimize the following parameters:

- Detection probability: the probability that an impinging photon will generate an electrical output.
- Spectral response: longer wavelength efficiency is desirable in many applications.
- Noise: noise reduction at room temperature; in addition, minimization of correlated noise (afterpulsing)

- Jitter: reduce the time uncertainty in determining the photon arrival time.
- Count rates: Maximize the number of counts per unit time
- Active area: Some application require a large pixel area to ease optical alignment; others require pixel miniaturization for array integration
- Fill factor: Maximize the percentage of the pixel area which is sensitive to incoming photons.

In the following sections we will describe various implementations of SPADs. These can be broadly categorized to reach-through devices, where the full cross-section of the wafer is used, and surface SPADs, where only a relatively shallow layer is used.

1.4.2.1. Reach-through Devices

In reach-through devices, most of the wafer thickness constitutes the absorption region, thereby maximizing the absorption probability. In one of the first reach-through silicon SPADs, described by McIntyre and Webb [58], a custom technology was used with special ultra-low-doped p-silicon wafers to develop the reach-through avalanche photodiode (Figure 1.10). The devices had a thick depletion layer (from 20 to 100 μm) and a correspondingly high breakdown value (from 100 to 500 V). The active area (diameter from 50 to 500 μm) was defined and edge effects were avoided by a p^+ implanted enrichment and by a reduced thickness of silicon over it, obtained by accurately etching the wafer.

An improvement to the first reach-through device is implemented in the Slik^{TM} device by RCA/Perkin Elmer [58]. This device is also manufactured using an optimized, dedicated process, and has a large active area of 200 µm. The active area is defined by a p^+ implant and a deep diffusion at its center, where the wafer is etched to approximately 30 µm (Figure 1.10).

Due to its relatively wide depletion region, the Slik device exhibits a high quantum efficiency of 30% at 500 nm and 70% at 700 nm. The ultra-clean process used results in a very low dark count of approximately 150 counts per second. However, due to its large diameter, the full-width-half-maximum (FWHM) jitter is greater than 300 ps, which is unacceptable to many applications. In addition, its high operating voltage of 300 V-400 V necessitates specialized circuitry and dissipates excessive power. The dedicated technology required for its manufacturing resulted in an expensive device and in performance which varies between devices.

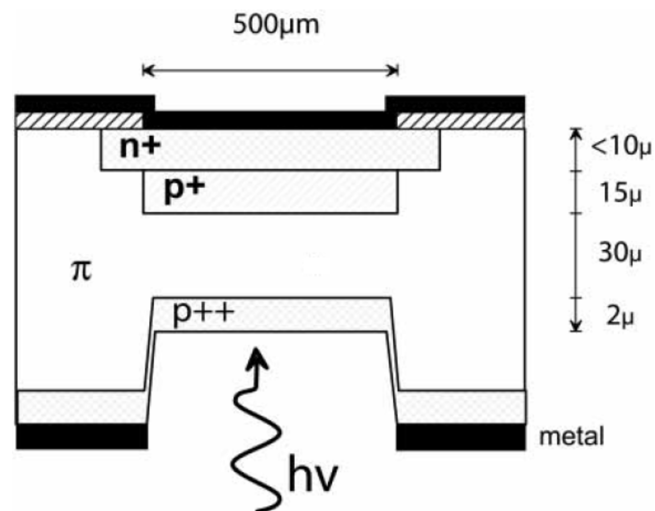


Figure 1.9: Reach-through device: McIntyre [58] [59].

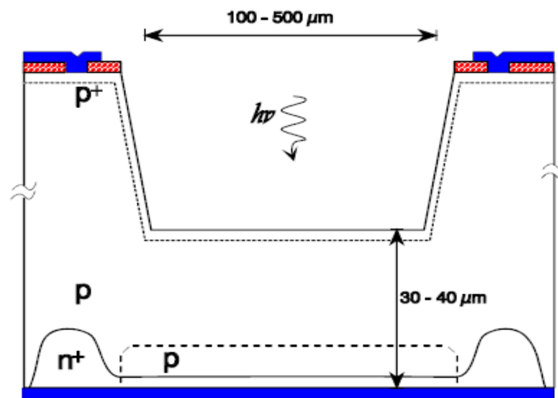


Figure 1.10: Slik^{TM} device [58] [59]

A different approach for fabricating reach-through SPADs has been carries out at Lincoln Laboratory in the framework of the development of a Lidar system [60]. Because the systems had to output images in addition to simple ranging [11, 61], the SPAD device must be scalable.

A heavily-doped p^+ substrate was overgrown by a lightly-diffused p-doped epitaxial layer in a dedicated process (Figure 1.11). The detection structure in this device is known as p-π-p-π-n, where π denotes very-lightly doped. Photons impinge on the back-side of the wafer, and are absorbed in the lower π layer, which, due to its thickness contributes to the high quantum efficiency of the device. The device is biased such that a moderate electric field is established in the absorption region, causing photoelectrons to drift into the upper π layer. The field in this multiplication layer is much stronger, resulting in impact ionizations and an avalanche.

Unlike the previous reach-through devices, electric field confinement in this device does not require etching at the edge of the device. Rather, the deep

implantation structure confines the electric field lines, and directs the generated charges towards the active area.

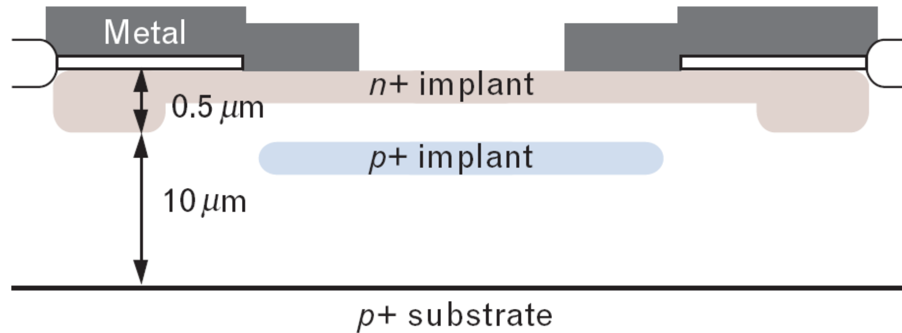


Figure 1.11: Bulk SPAD used for lidar application [60]

32 x 32 arrays of these devices have been manufactured, with 100 μm and 150 μm pitch, corresponding to 30 μm and 50 μm active area diameters. This optimized structure resulted in 60% detection probability in visible wavelengths [4, 61]. Due to the wide absorption region, voltages in excess of 100 V must be applied resulting in high power dissipation. Due to the large absorption volume, the jitter of the devices was estimated to be 150 psec with a dark count rate of approximately 1000 counts per second. Due to the high avalanche charge, 1 μs of dead time is required to release a sufficient fraction of trapped charges. The fill factor of the device was only 5% - substantially higher than competitive structures due to bumping to off-chip processing circuitry, but still very low compared to other imaging technologies.

1.4.2.2. Surface Devices

Surface SPADs have been fabricated since the 1960's by Haitz [62]. They employ processing steps which are similar to those used for the manufacture of mainstream semiconductor structures, such as the Metal-Oxide-Semiconductor (MOS) transistor. The main difference between devices with relatively-shallow structures as opposed to reach-through devices is the reduction in size (diameter and width) of the absorption and multiplication volumes. This has a number of implications for the surface devices:

- Reduction in detection efficiencies due to narrower absorption regions.
- Shift in spectral response towards lower wavelengths.
- Reduction of jitter due to reduced junction diameters.
- Reduction in dead times due to smaller junction capacitances.
- Reduction in power consumption due to lower breakdown voltages.
- Enabling of large-scale integration with processing circuitry on the same die.

Haitz's early implementation of avalanche diodes used a guard-ring structure which is still being used today (Figure 1.12). A shallow n^+ implant forms a junction with a p-doped substrate. Because the implantation and subsequent diffusion steps form a typical rounded junction, where premature breakdown may occur, a lightly-doped deep n^- guard-ring is implanted. Because the resultant depletion region is

narrower than in reach-through devices, much lower voltages, approximately 32V, are required in order to attain junction breakdown.

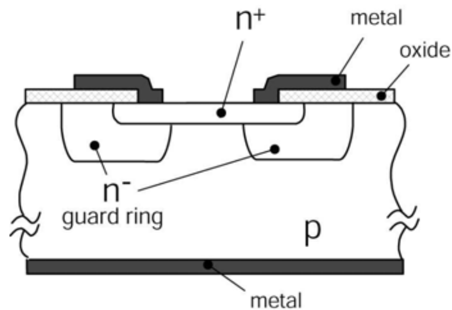


Figure 1.12: Haitz's diode using a diffused guard ring [62, 63]

During the early years of semiconductor fabrication, operation of devices such as the one described above in Geiger mode has been impeded by the poor quality of the silicon substrate. In the late 1980's, wafer quality made the fabrication and commercialization of SPADs possible. In 1989, Cova introduced a double-epitaxy SPAD [64] with an improved jitter response (Figure 1.13). Photons incident upon the top surface are absorbed in the p-type epitaxial layer. They are attracted by a low electric field to the n^+/p junction, where they are multiplied. The novelty in this SPAD is the use of a double-epitaxial structure, forming two diodes. The lower diode prevented photogenerated electrons from the substrate from diffusing into the SPAD junction causing an increased timing uncertainty. The top epitaxial structure was used to form an ultraclean junction for the SPAD. In order to overcome the large resistivity of this layer, a p^+ layer was implanted below the top epitaxial layer. Finally, a p^+

implant at the center of the device reduced edge breakdown effects by focusing the electric field lines to the junction center.

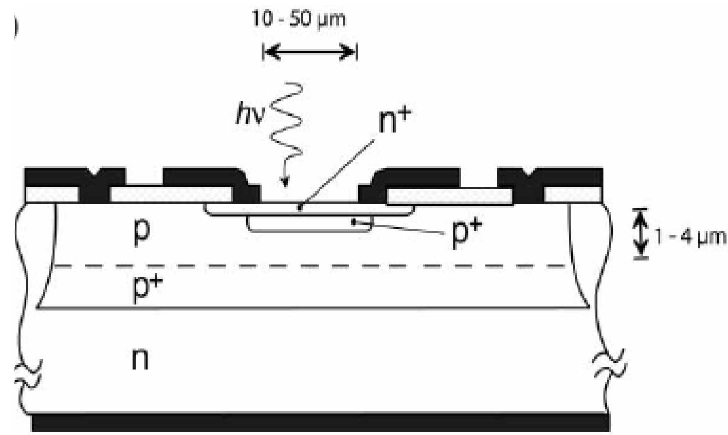


Figure 1.13: Cova's double-epitaxy SPAD [64, 65]

This device exhibited a breakdown voltage of 50 V, good jitter performance of 55 ps FWHM, and an improved dark count rate (compared with the device of Figure 1.12) of 10 kcounts per second. Further improvements were reported over the years to this device, mainly aimed at reducing the jitter of the device for TCSPC applications [66, 67].

A summary of the main differences between reach-through and surface SPADs is shown in Table 1.1.

Table 1.1. Comparison between bulk and surface SPAD performance [60, 63].

	Bulk SPAD's, (Lincoln Labs)	Surface SPADs (Milan Polytechnic)
Quantum efficiency	(~70%)	~40%
Absorption spectrum	Wide (1064nm)	Narrower (800nm)
Timing accuracy	Poor (150ps)	Good (<50ps)
Dark count rate	>1000 / sec	~1000 / sec
Pixel size	100 um (excl. quenching)	50um (incl. quenching)
Integration	Bumping	Possible
Breakdown voltage	~100V	~40V

1.4.2.3. Devices in Commercial High-Voltage CMOS Technologies

In 1992, Rochas demonstrated an implementation of a SPAD similar to Haitz's in a commercial high-voltage CMOS process by Austria Microsystems [68-70]. Rochas' device utilized a triple-well process (Figure 1.14). In order to make the structures compatible with the commonly available p-doped substrates, which are commonly grounded in CMOS circuits, the doping types were reversed compared with Haitz's device. A shallow p^+ source/drain implant forms one terminal of the diode, while an N-well, usually used for PMOS transistors, was used as the second terminal.

The deep-implant diffused guard-ring was implemented using the p-well structure. A deep N-well was used to contact the diode.

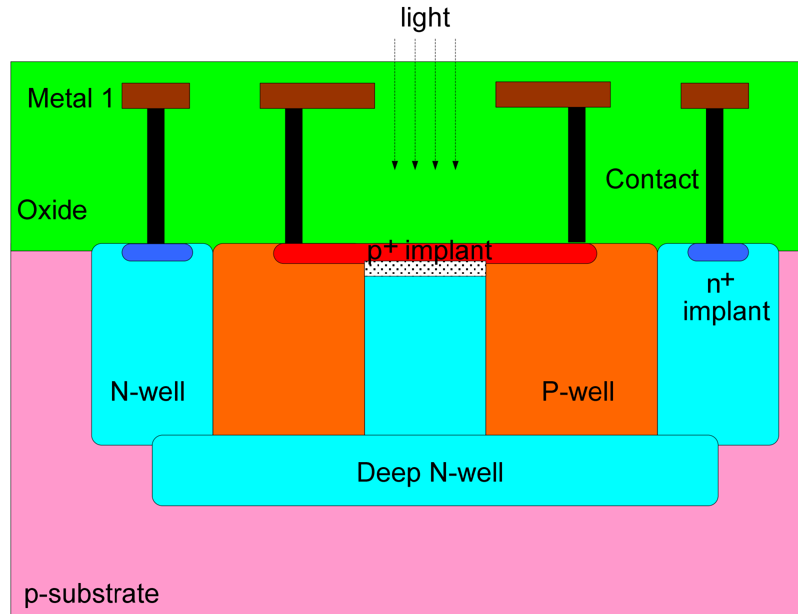


Figure 1.14: Cross-section of a triple-well CMOS SPAD

The device was operated by biasing the p^+ terminal with a high negative voltage, slightly below the breakdown of the junction. The N-well was biased at positive VDD (up to 5V) through an external quenching resistor, which was later implemented as an on-die PMOS transistor.

Device characterization showed a detection peak of 30% at 450 nm. This shift is due to the shallow junction (0.3 μm deep) and narrow depletion region. The dark current rate was only 100 counts per second for a 7 μm active-area diameter. The dead time was 30 ns – 50 ns and the breakdown voltage only 25 V, allowing for a low-power operation. Jitter had a 60 ps FWHM, but a relatively wide tail (FW [M/100]) of 2 ns) due to carrier diffusion (see section 2.8.3 for more details).

Later work by Charbon investigated the possibilities created by this CMOS SPAD in terms of direct porting to a smaller geometry (0.35 μm high-voltage), array integration, and device architectures, and also demonstrated the applicability of the device for 3D facial imaging and biological applications [71-74]. In recent years, a number of other groups manufactured similar CMOS SPADs as Rochas', and integrated them with various circuitry for specific applications [75, 76].

The triple-well CMOS SPAD holds much promise and has exhibited excellent performance. Nevertheless, it has three significant drawbacks:

a) The guard-ring structure, comprised of the p-well ring which is surrounded by an n-well ring, which in turn is surrounded by a p^+ substrate contact, is highly inefficient. It results in fill factors of a few percent (1.7% in a SPAD with 7 μm active area in a 0.8 μm process). This fill factor does not allow for efficient coupling of light, and, in reality, lowers the effective detection efficiency of the device. It also makes it unrealistic to integrate megapixel arrays due to die size and yield limitations.

b) The triple-well SPAD scheme is not scalable to technologies below 0.35 μm. This is due to a shallow-trench isolation (STI) structure which is put automatically by the fab in order to prevent punch through in CMOS circuits. STI is inserted wherever there is no active area and no transistor channels. As such, it would be inserted at the edge of the SPAD's p^+, thereby making the structure impossible to manufacture.

This is a serious drawback because the aim of porting the device to CMOS is increased integration. Only deep-submicron processes can accommodate the required high-frequency circuitry for processing the SPAD data.

c) The diffused guard-ring is relatively wide and experiences a moderate electric field. Therefore, there is a relatively high probability for thermally-generated carriers to drift from this ring area to the multiplication area, thereby causing increased dark current, and possibly a jitter tail (e.g., as a result of captured electroluminescent photons – see section 2.8.6).

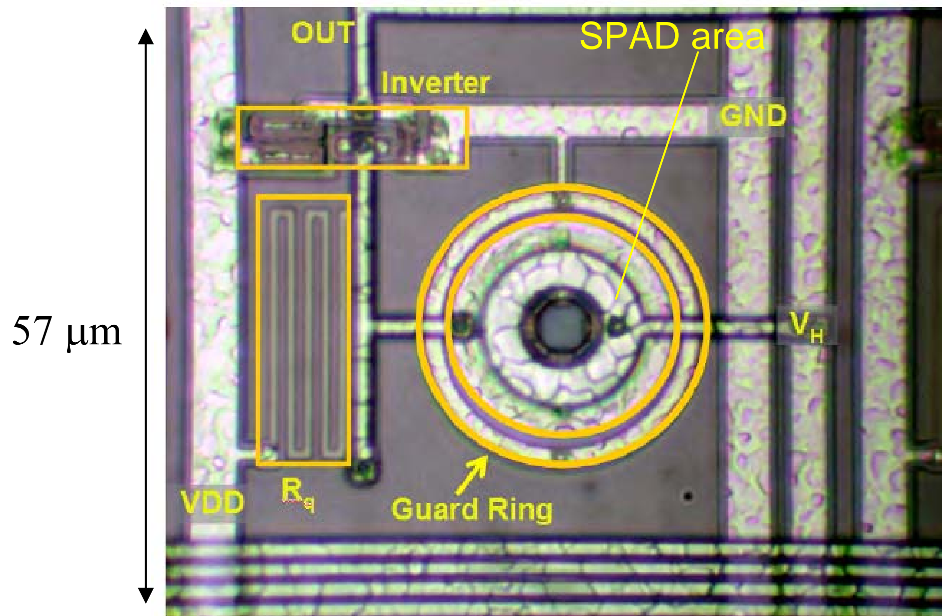


Figure 1.15: Rochas' passively-quenched triple-well CMOS SPAD [77]

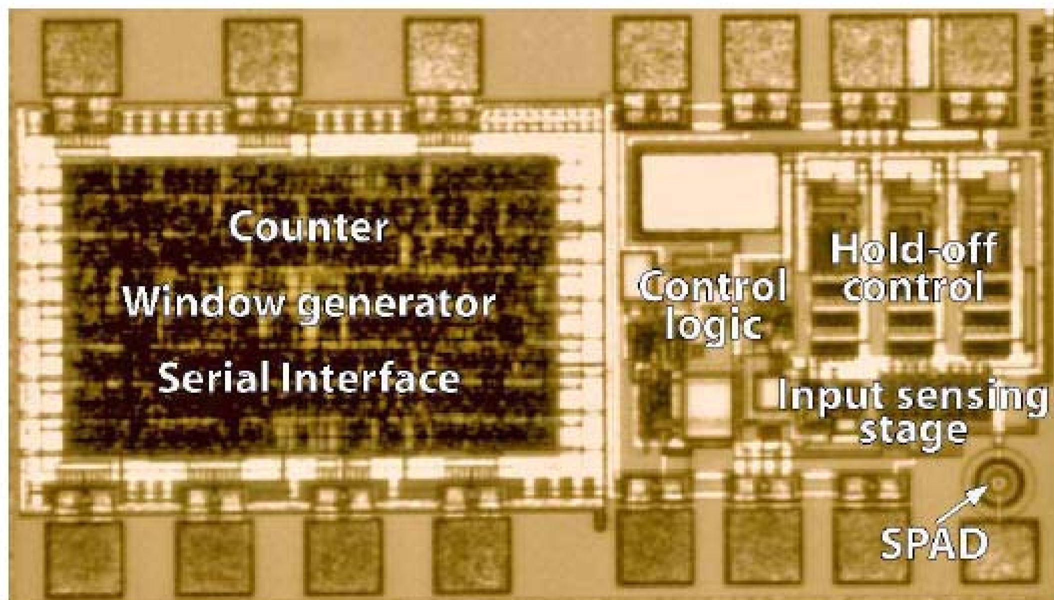


Figure 1.16: Zappa's actively-recharged triple-well CMOS SPAD in a High-Voltage 0.8 µm double-poly, double-metal CMOS process [75]

1.5. Performance Comparison

A performance comparison of the various technologies outlined in this chapter is shown in Table 2.1. It can be seen that cooled CCDs offer the best performance when quantum efficiency is concerned. However, for those applications which require high sensitivity in addition to time-of-arrival information, and which require array imaging, SPADs offer the best performance potential.

1.6. Aim and Challenges

The prevalence of single-photon detectors has been demonstrated in this chapter in areas ranging from single-molecule studies to lidar and three-dimensional imaging. Various technologies are available. SPADs offer distinct advantages, as follows:

- Robustness
- Potential for megapixel scalability
- Small size
- Low power requirements
- Excellent timing response
- Potential for cost advantage
- Room temperature operation
- Uniformity in an array and across devices

- Ultra-low noise

Table 1.2. Comparison of detector performance [1, 47, 78, 79]

	Single-photon counters		**Imagers**		
Type	**SPAD**	**PMT/MCP**	**Cooled CCD**	**ICCD**	**EBCCD**
Enclosure	Solid-state	Vacuum	Solid-state + vacuum	Solid-state + vacuum	Solid-state
Photon counting	Yes	Yes	No	No	No
2D Imaging	Not yet	Yes	Yes	Yes	Yes
Gain	"Infinite"	10^6	~1	10^6	1,000
Photocathode	None	Multi-alkali/ GaAs	None	Multi-alkali/ GaAs	GaAs
QE at 600 nm	20%-60%	6%	90%	<20%	90%
Time resolution	35 ps	20 ps	6 ms	100 ms	100 ms
Count rate	10 MHz	10 MHz	1 MHz	1 MHz	1 MHz
Readout speed	1 MHz	1 MHz	30 Hz	30 Hz	30 Hz

We also showed that the range of requirements from the detector varies widely between applications. These requirements include:

- Quantum efficiency
- Spectral response

- Count rate
- Time precision or device jitter
- Signal-to-noise ratio
- Dynamic range
- Scalability to large arrays
- Power dissipation
- Mechanical robustness and enclosure

The custom processes required for both the reach-through and the surface SPADs described in this chapter impose significant limitations on scaling and integration of the devices. Although dedicated processes offer optimized device structure and results in excellent device performance, the quenching and recharge circuitry, as well as the signal processing, must be accomplished off-chip. This has several implications:

- Large junction capacitance, resulting in excessive afterpulsing and long dead times.
- Complex and expensive interfaces.

One implementation of a SPAD in a CMOS process has been demonstrated by several groups but is still lacking in terms of fill factor, jitter tail and scaling.

This dissertation investigates a new SPAD device, which has been designed with the following aims:

- Utilize generic commercial deep-submicron CMOS technologies.

- Increase the fill factor significantly over the state-of-the-art.
- Decrease dead times.
- Reduce device jitter.
- Reduce power dissipation.
- Decrease pixel-to-pixel pitch.

The potential benefits of such a device include:

- Integration of SPAD array with quenching, recharge and processing circuitry on the same die.
- Improved light coupling into the device thanks to increased fill factor.
- For the first time, megapixel SPAD arrays could be manufactured, in terms of both die size and power dissipation.
- Significant reduction in acquisition times for such applications as FLIM, and increased data throughput in quantum-key distribution, when the SPAD is used as part of a frequency upconversion receiver (see chapter 5).

There are a number of significant challenges in the implementation and characterization of the proposed device. These can be categorized as follows:

A. Physical challenges

- Deep-submicron technologies utilize high doping levels and shallow junctions. These parameters increase the probability for tunneling through the junction, as well as for field-enhanced thermal generation through the Poole-Frenkel effect, as explained in the next chapter. Implications may be

an increased dark current rate. The challenge is to construct a device which can function within these constraints.

- A new, more efficient guard ring must be designed, which is both compatible with deep-submicron-CMOS design rules, and which can contain the high electric fields with which the SPAD operates.
- Shallow junctions have small effective radii of curvature, and as such premature breakdown is a major concern. An area-efficient structure must be found to planarize the junction.
- Deep-submicron technologies are designed for low-voltage operation. The benefit of a non-high-voltage process is in its ability to utilize more third-part design intellectual–property, and in its wider availability and reduced defect density. A way must be found to operate the device with voltages much in excess of the maximum these processes are specified and characterized for, in a reliable manner.

B. Design challenges

- Generic CMOS technologies are optimized and characterized for transistor-based design. The process parameters which are relevant to Geiger-mode operation are usually either not available or not furnished by the fab. They must therefore be extracted indirectly from other process parameters, or characterized using custom designed test chips.
- Similarly, commercially available CAD tools simulate the electrical performance of circuits, based on device models provided by the fab. A

CMOS SPAD's operation involves: a) optics - light absorption in the device, b) electrostatics – electron-hole generation and the build-up of an avalanche, c) ultrafast electronics – due to the psec build-up of the avalanche and its fast quenching, and c) analog electronics – the sensing, processing and output of the electrical signal. Therefore, a simulation tool set must be developed, to analyze the performance of this signal propagation chain with maximal precision.

C. Characterization challenges:

- Precise photon-number generation at different wavelengths: the characterization of SPADs requires knowledge of the precise photon flux impinging on the detector active area. This flux must remain acceptably stable over the time of the experiment. Since photon arrivals at low flux follow a Poissonian distribution, single-photon detection calculations must take this statistical distribution into account. Moreover, in order to characterize the spectral response of the device, the precise photon number must be known for a range of wavelengths.
- Collection of large amounts of data in real-time over short periods of time. Because the statistical behavior of the arriving pulse trains is of utmost importance, data cannot be sampled, and must be collected in real-time, over sufficiently long periods of time. Since the devices are designed to operate at high speeds (short dead times), the collection of large amounts of data in real time becomes a technological challenge.

In addition to the development and characterization of the STI-bounded SPAD, this work describes a novel method for frequency upconversion, which uses the new STI-bound CMOS SPAD. This new method aims to improve present upconversion techniques [2, 80-84] in the following aspects:

- Achieve single-photon upconversion with high efficiency (>70%).
- Enable scalability to large-format arrays.
- Perform upconversion with low power dissipation.
- Facilitate low-cost production of the upconversion device.

1.7. Dissertation Outline

Following the motivation and overview of the state-of-the-art in single-photon detection presented in this chapter, we provide an overview of the physics and figures-of-merit of Geiger-mode SPADs in Chapter 2. This chapter also describes the various quenching and recharge schemes which are required for operating these devices. The new STI-bounded SPAD is introduced in Chapter 3, which also describes the modeling and simulation of the device. In addition, this chapter is concerned with the design and simulation of various circuits used for the operation, optimization and characterization of the detector. In Chapter 4, we describe the characterization setup used to test the SPAD and provide experimental results relating to the performance of the device. The concept of single-photon upconversion via hot-carrier luminescence is introduced in Chapter 5, where we develop the theory, calculate theoretical

upconversion efficiencies and describe our experimental results. We conclude this dissertation with a summary and an outlook on future research directions stemming from the work described here.

2. PHYSICS AND OPERATION OF SINGLE-PHOTON AVALANCHE DIODES

2.1. Introduction

Detecting single photons with single photon avalanche diodes is akin to detecting a spark using a room filled with gasoline vapor. A pn junction is biased higher than its breakdown voltage but is vacated of free charge carriers. As soon as a charge carrier enters this region, for example, as a result of photogeneration, the whole junction collapses into a plasma. The challenge in designing and operating these devices lies in the control and quenching of this charge avalanche before irreversible damage takes place, as well as in the fast resetting of the junction so that subsequent photons may be detected.

In this chapter, we will examine the avalanche process in semiconductor pn junctions. We will quantitatively understand the effect of curvature on the breakdown behavior of these junctions and summarize existing guard ring structures, which are used to contain avalanche breakdown. We will then be able to understand the principle of operation of the Geiger-Mode Single Photon Avalanche Detector. Such an understanding requires examination of the different quenching and recharging schemes. Next, we will examine the various figures of merit of Geiger-Mode Single Photon Avalanche Detectors and compare various device implementations. This will

lay the groundwork for an understanding of the STI-bounded device presented in this dissertation.

2.2. Basic Properties of the pn Junction

The pn junction lies at the heart of the SPAD device. Traditionally, two types of junctions are analyzed – an abrupt junction and linearly-graded one [85]. When the impurity concentration in a semiconductor changes abruptly from a constant acceptor level, N_A, to a constant donor level, N_D, an abrupt junction is formed (Figure 2.1). Such junctions may form when two materials are fused together. An approximately abrupt junction forms when an implant with a slowly-varying doping concentration encounters the substrate.

At steady-state, when the junction is at thermal equilibrium, no bias is applied across it, and no current flows through it, a voltage forms across the junction in order to maintain a constant Fermi energy throughout the structure. This voltage is termed the built-in voltage. In order to account for the vanishing of the electric field away from the junction, the charges on either side of the junction must cancel each other out:

$$N_A x_p = N_D x_n ,$$

Equation 2.1

We employ Poisson’s equation to obtain the electric field distribution around the junction:

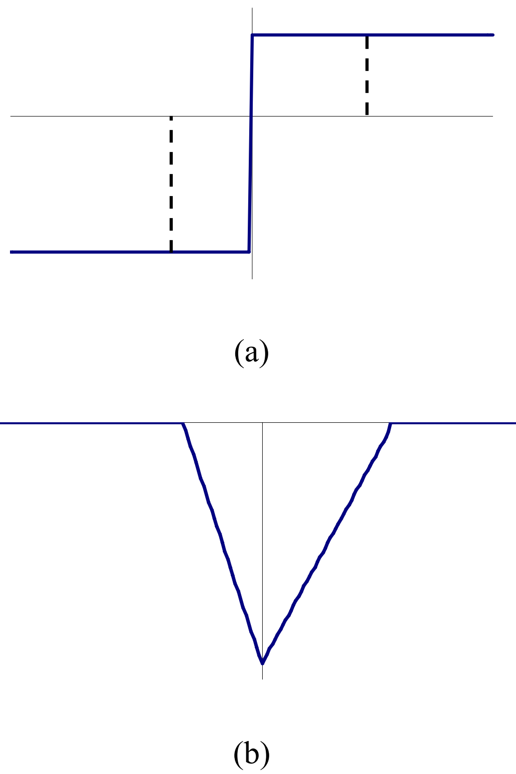


Figure 2.1: (a) Doping concentration and (b) electric field around an abrupt pn junction. Depletion region boundaries are shown as broken lines.

$$-\frac{\partial^2 V}{\partial x^2} \equiv \frac{\partial \boldsymbol{\mathcal{E}}}{\partial x} = \frac{\rho(x)}{\varepsilon_s}$$

or,

$$-\frac{\partial^2 V}{\partial x^2} \approx -\frac{q}{\varepsilon_s} N_A \qquad \text{for } -x_p \leq x < 0$$

$$-\frac{\partial^2 V}{\partial x^2} \approx \frac{q}{\varepsilon_s} N_D \qquad \text{for } 0 < x \leq x_n$$

Equation 2.2

Integration of Equation 2.2 results in an expression for the electric field:

$$\mathcal{E}(x) = -\frac{qN_A\left(x + x_p\right)}{\varepsilon_s} \qquad \text{for } -x_p \le x < 0$$

$$\mathcal{E}(x) = \frac{qN_D\left(x - x_n\right)}{\varepsilon_s} \qquad \text{for } 0 < x \le x_n$$

Equation 2.3

with a maximum field at x=0:

$$\left|\mathcal{E}_m\right| = \frac{qN_D x_n}{\varepsilon_s} = \frac{qN_A x_p}{\varepsilon_s}$$

Equation 2.4

Integrating Equation 2.2 again results in an expression for the voltage:

$$V(x) = \mathcal{E}_m\left(x - \frac{x^2}{2W}\right);\ W = x_n + x_p$$

Equation 2.5

If the doping profile is highly asymmetrical, i.e., one side is much more heavily doped than the other, the junction is said to be one-sided. In this case, Equation 2.5 is reduced to an expression for the depletion region's width:

$$W = \sqrt{\frac{2\varepsilon_s V_{bi}}{qN_B}}$$

Equation 2.6

where N_B is the doping of the lower-doped region.

Most junctions found in semiconductor devices are not abrupt. For example, the source/drain implant of a MOSFET transistor has a peak doping concentration, $N_{B,max}$ close to the surface. At the junction of this implant with the substrate (or well), at a depth x_j, the net doping concentration is zero. We can often approximate the doping gradient of such structures as linear, with an effective grading coefficient $a = \frac{N_{B,\max}}{x_j}$.

The relevant parameters of such junctions can be calculated in a similar fashion to those of abrupt junctions (Figure 2.2). The Poisson equation in a one-sided linearly-graded junction becomes

$$-\frac{\partial^2 V}{\partial x^2} \equiv \frac{\partial \mathcal{E}}{\partial x^2} = \frac{\rho(x)}{\varepsilon_s} \approx \frac{q}{\varepsilon_s} ax \qquad 0 \le x \le W$$

Equation 2.7

The electric field now decreases quadratically with distance from the junction:

$$\mathcal{E}(x) = -\frac{qa}{\varepsilon_s} \cdot \frac{W^2 - x^2}{2}$$

Equation 2.8

The peak electric field is positioned at the junction itself:

$$|\mathcal{E}_m| = \frac{qaW^2}{2\varepsilon_s}$$

Equation 2.9

The depletion width in this case becomes:

$$W = \left(\frac{3\varepsilon_s V_{bi}}{2qa} \right)^{1/3}$$

Equation 2.10

The junction capacitance per unit area can be calculated for both junction types as:

$$C = \frac{\varepsilon_s}{W}$$

Equation 2.11

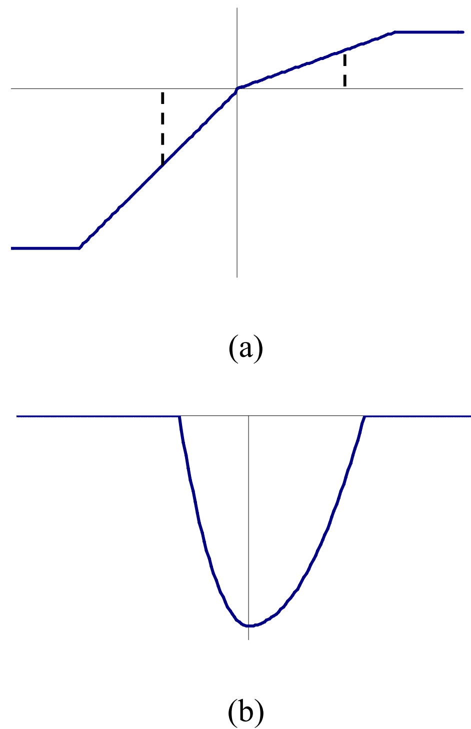


Figure 2.2: (a) Doping concentration and (b) electric field around an asymmetrical linearly-graded pn junction. Depletion region boundaries are shown as broken lines.

2.3. Breakdown Mechanisms in pn Junctions

When the reverse-bias across a pn junction increases, so does the electric field. As the electric field rises, several processes kick into gear. At sufficiently high fields, these processes accelerate into a runaway regime which is called junction breakdown. We will describe three types of high-field effects: thermal instabilities, tunneling and impact ionization [85].

2.3.1. Thermal Instabilities

When a high reverse-bias is applied across a pn junction, a leakage current flows through it. This current causes heating in the junction region which, in turn, increases the current [85]. If a current-limiter is not placed in series with the device, this process can accelerate, resulting in irreversible damage. Because the current temperature-dependence is proportional to $exp(-E_g/KT)$ this effect is more dominant in narrow-bandgap materials.

2.3.2. Tunneling

As the electric field in the depletion region approaches 10^6 V/cm, the band-to-band tunneling increases significantly [85]. This process is illustrated in Figure 2.3.

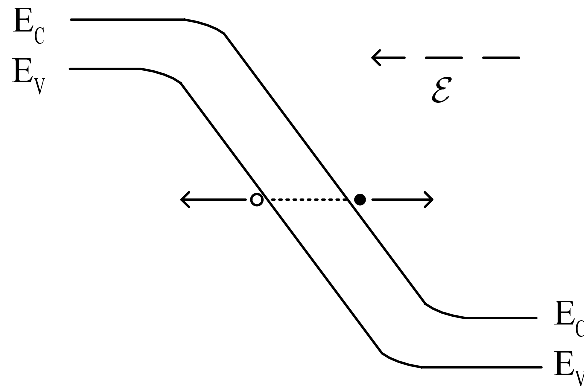


Figure 2.3: Energy band diagram illustrating direct band-to-band tunneling.

The tunneling current is often modeled using the expression [86] [87]:

$$J = cqV_j F^{\sigma} \exp\left(-{F_0}/{F}\right)$$

Equation 2.12

where V_j is the applied junction voltage, F is the electric field at the junction and σ is a numerical constant ($\sigma = 1$ for direct transitions). F_0 and c are both material-dependent constants which depend on the energy bandgap and the effective mass. As a rule of thumb, tunneling becomes the dominant breakdown mechanisms for junctions whose breakdown voltage satisfies $V_{Breakdown} < 4E_g/q$. This translates to silicon junction with a breakdown voltage lower than approximately 5V. As we will show below, this is the case in processes with high doping concentrations, specifically, deep-submicron ones. We will return to the tunneling current in Section 2.8.3 where we discuss noise sources in SPADs.

2.3.3. Impact Ionization

When a free electron experiences a strong electric field, the energy it receives from the electric field may become so large that it can reach an energy larger than the band gap. If this electron impacts the lattice, it can cause impact ionization and create an electron-hole pair (see Figure 2.4). The primary and secondary electrons, and the secondary hole, will accelerate in opposite directions, and may cause additional ionizations. An avalanche breakdown may then develop.

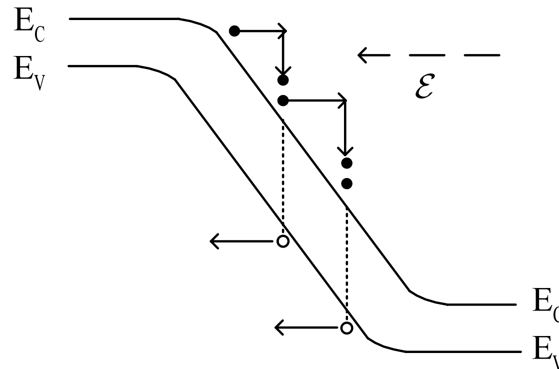


Figure 2.4: Energy band diagram illustrating multiplication by impact ionization. Holes are depicted by empty circles and electrons by filled circles.

The breakdown behavior of a pn junction has been investigated since the early days of semiconductor research [51]-[53, 55], [88]. It is convenient to define an electron ionization rate, α, as the number of electron-hole pairs produced by an electron per distance traveled in the direction of the field. β is the analogous quantity for holes. Referring to Figure 2.5, we assume that at steady state, a hole current I_{p0} enters the junction at x = 0 and an electron current I_{n0} enters the junction at x = W. The

incremental hole current generated per second by impact ionizations of holes and electrons in the interval dx inside the depletion region is:

$$\frac{dn_p}{dx} = n_n \alpha dx + n_p \beta dx$$

Equation 2.13

Equation 2.13 can be solved with boundary condition $n_p(W) = M_p n_{p0}$ resulting in [52]:

$$1 - \frac{1}{M_p} = \int_0^W \beta \exp\left[-\int_0^x (\beta - \alpha) dx' \right] dx$$

Equation 2.14

An avalanche breakdown occurs when the multiplication factor approaches infinity, i.e.,

$$\int_0^W \beta \exp\left[-\int_0^x (\beta - \alpha) dx' \right] dx = 1$$

Equation 2.15

When the initiating carriers are electrons, β and α are exchanged in Equation 2.15.

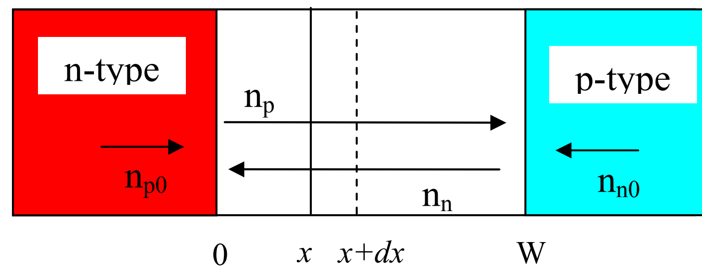


Figure 2.5: Geometry for calculating the avalanche condition.

When analyzing and designing pn junctions for operation near or above the junction breakdown, it is necessary to estimate the junction breakdown voltage, the electric field at breakdown, and the width of the depletion region. This is possible by solving Equation 2.15, assuming the ionization rates are only field-dependent. α and β for silicon have been extensively measured, although results are not always in agreement [53, 55, 89], [90]. It is widely accepted that the ionization rates follow a relationship:

$$\alpha, \beta = A \exp\left(-\frac{b}{E(x)}\right)^m$$

Equation 2.16

In this work we follow Sze's [85] and Grant's [89] numbers, and approximate them as (Figure 2.6):

$$\alpha = 7.16 \times 10^4 \exp\left(-1 \times 10^6 / E\right)$$
$$\beta = 3 \times 10^6 \exp\left(-2 \times 10^6 / E\right)$$

Equation 2.17

The breakdown voltages can be expressed as:

$$V_B = \left(\frac{\varepsilon E_m^{\ 2}}{2q}\right)\left(\frac{1}{N_B}\right)$$ for an abrupt junction and

$$V_B = \frac{4 E_m^{3/2}}{3} \sqrt{\frac{2\varepsilon}{q}} \left(\frac{1}{\sqrt{a}}\right)$$ for a linearly-graded one.

Equation 2.18

where E_m is the electric field at breakdown [90]. Various approximations have been used to evaluate the breakdown field.

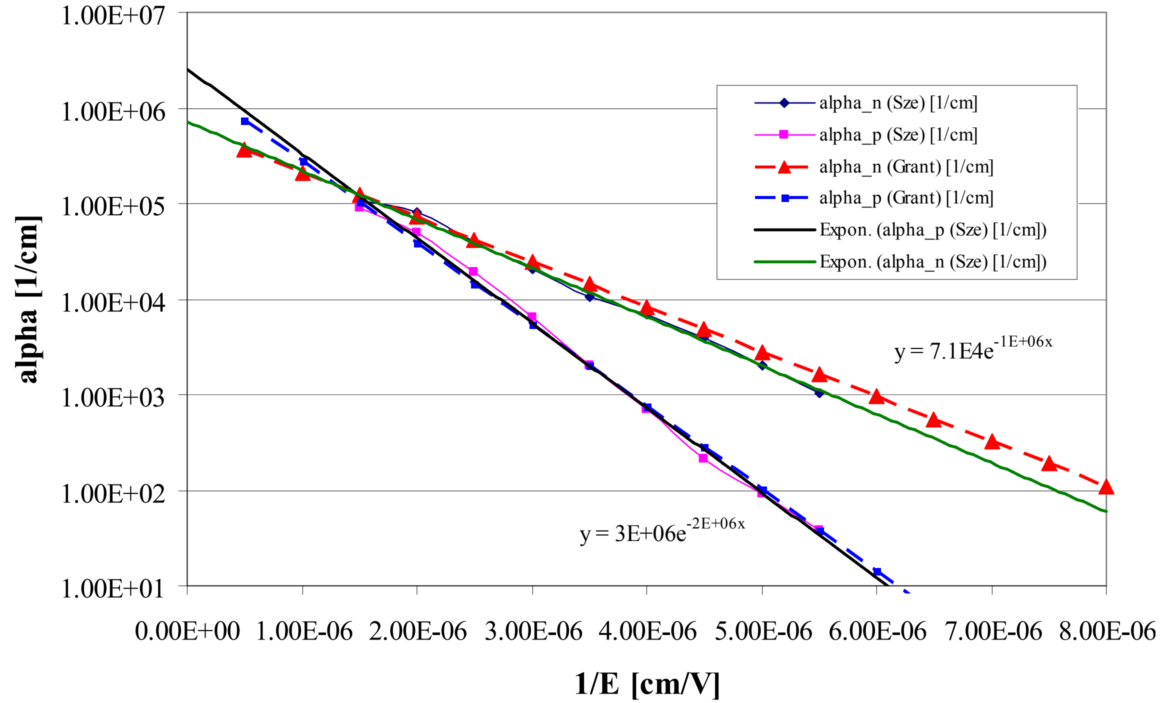


Figure 2.6: Electron and hole ionization coefficients for silicon [[89], [85]].

We follow Sze to get:

$$V_B = 60\left(\frac{E_g}{1.1}\right)^{3/2}\left(\frac{N_B}{10^{16}}\right)^{-3/4}$$ for an abrupt junction, and

$$V_B = 60\left(\frac{E_g}{1.1}\right)^{6/5}\left(\frac{a}{3}\times 10^{-20}\right)^{-2/5}$$ for the linearly-graded one.

Equation 2.19

Equation 2.19 indicates that, as the doping intensity increases, the breakdown voltage decreases. This conclusion can also be explained intuitively: as doping concentrations increase, depletion widths decrease. As a result, lower voltages are required to achieve the breakdown field. This effect is exacerbated by the bandgap narrowing at high doping concentrations.

2.4. Breakdown in Non-Planar pn Junctions

The foregoing discussion in Section 2.3 assumed that the junction is planar. In reality, pn junctions are usually formed by implanting a dopant at a known concentration with a known energy into a substrate with an opposite-type dopant [91]. This implantation step is usually followed by annealing at a given temperature and for a given time. The resulting profile is usually pear shaped. It is therefore important to understand the behavior of non-planar junctions under high field conditions.

The theory of breakdown in curved junctions has been developed by Sze and Gibbons [56] and is summarized here. We analyze two types of curved junctions:

a) A cylindrical junction forms along the sides of a rectangular doping region. Its doping profile varies along one radial axis, r_1:

$$C(r) = C_i(r) - N_B$$

Equation 2.20

where $C_i(r)$ is the distribution of diffused impurities and N_B is the background doping concentration.

b) A spherical junction forms along the corners of a rectangular doping region. Its doping profile varies along two orthogonal radial axes, r_1 and r_2:

$$C\left(r_1, r_2\right) = C_i(r_1, r_2) - N_B$$

Equation 2.21

As in 2.4, we consider an abrupt junction and a linearly-graded one. For an abrupt junction, the charge of ionized donor or acceptor atoms in the space-charge region is:

$$\rho\left(r\right) = -qN_B \qquad r_j \leq r \leq r_d$$

Equation 2.22

r_j and r_d are the junction's radius of curvature and the extent of the depletion region. For a linearly-graded junction,

$$\rho\left(r\right) = -aq\left(r - r_j\right) \qquad r_1 \leq r \leq r_2$$

Equation 2.23

Poisson's equation in this case is:

$$E\left(r\right) = \frac{1}{\varepsilon r^n} \int_{r_1}^{r_2} r^n \rho\left(r\right) dr + \frac{c}{r^n}$$

Equation 2.24

where n = 1 for a cylindrical junction and n = 2 for a spherical junction, and c is a constant.

Following a similar approach as in Section 2.3.3, the breakdown voltage is determined:

$$V_B = 60\left(\frac{E_g}{1.1}\right)^{3/2}\left(\frac{N_B}{10^{16}}\right)^{-3/4} \times \left\{\left[\left(n+1+\gamma\right)\gamma^n\right]^{1/(n+1)} - \gamma\right\}$$

for an abrupt junction, and

$$V_B = 60\left(\frac{E_g}{1.1}\right)^{6/5}\left(\frac{a}{3}\times 10^{20}\right)^{-2/5}$$

for the linearly-graded one.

Equation 2.25

Here, $\gamma = r_j/W$, where W is the depletion layer width, which is a function of the maximal field the curved junction can withstand before breakdown:

$$W = \left(\frac{10^{16}}{N_B}\right)\left(\frac{E_m}{1.6\times 10^5}\right)$$

Equation 2.26

E_m must be determined numerically from Equation 2.24. Note that in the case of a symmetrical linearly-graded junction, the breakdown voltage is independent of the radius of curvature. For reasons of symmetry, this junction is equivalent to a planar junction.

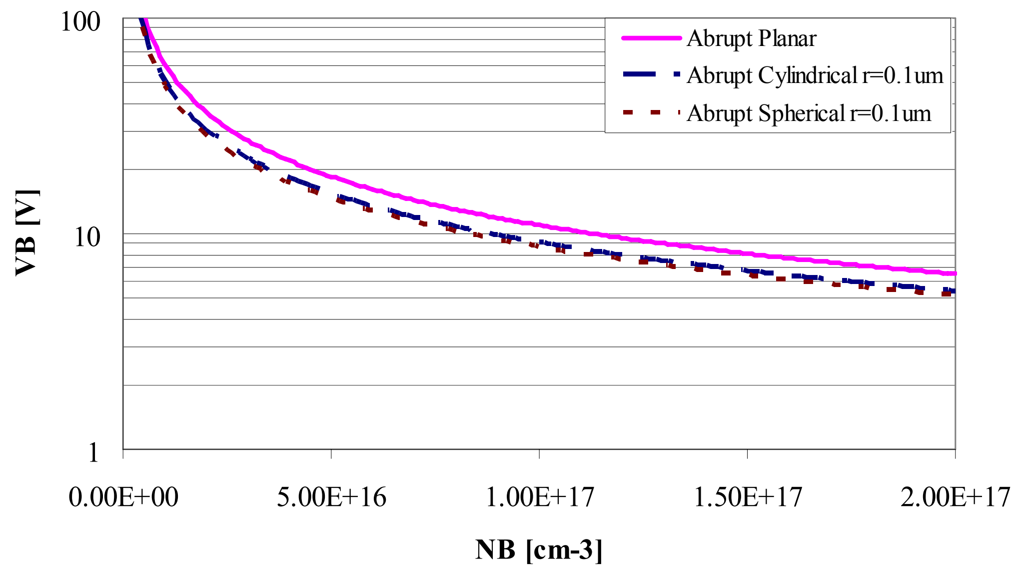


Figure 2.7: Effect of junction geometry on breakdown voltage.

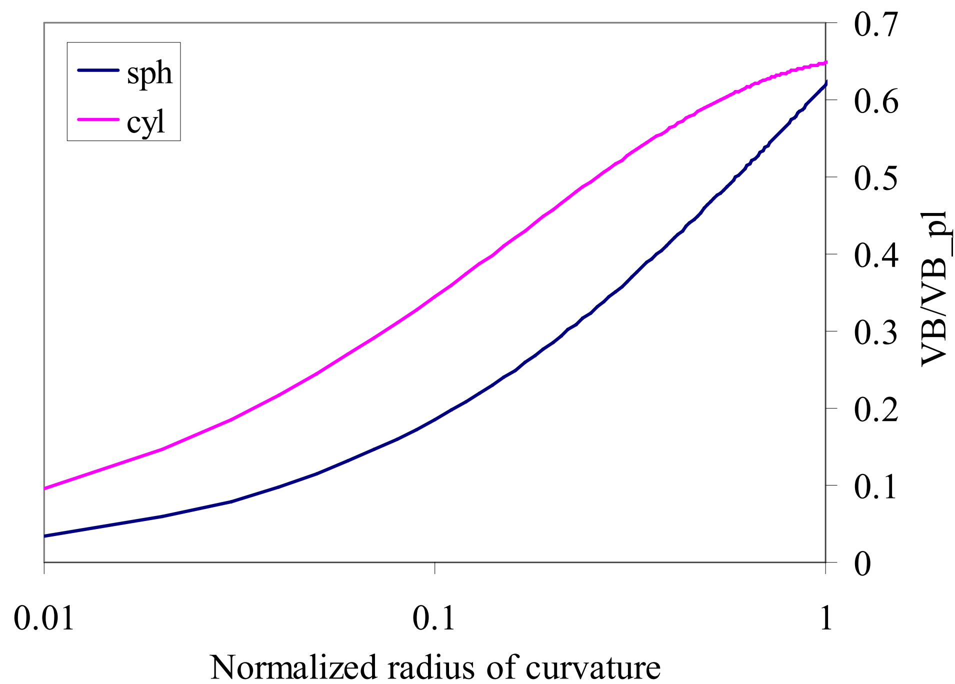


Figure 2.8: Effect of junction curvature (normalized to planar junction width) on breakdown voltage (relative to planar breakdown) for spherical and cylindrical junctions.

Figure 2.7 demonstrates the effect of junction curvature on the breakdown voltage. As the background doping concentration increases, the reduction in breakdown voltage becomes more pronounced. As can be seen from Figure 2.8, as the radius of curvature decreases, so does the breakdown voltage. This is due to electric field enhancement. We should note that shallow junctions necessarily have small radii of curvature, with dimension approximately equal to the junction depth.

There are a number of important conclusions that can be drawn from the foregoing discussion:

- Non-planar junctions will experience breakdown at different voltages in separate regions of the junction (planar region, sides and corners). This is illustrated in Figure 2.9 where a high electric field develops in the curved junction regions. This effect is undesirable because it can lead to excessive current densities in areas breaking down prematurely.
- Non-planarities can lead to localized breakdown (microplasmas). In advanced processes which have high doping concentrations as well as shallow junctions, non-planarities are more significant than in older technologies.

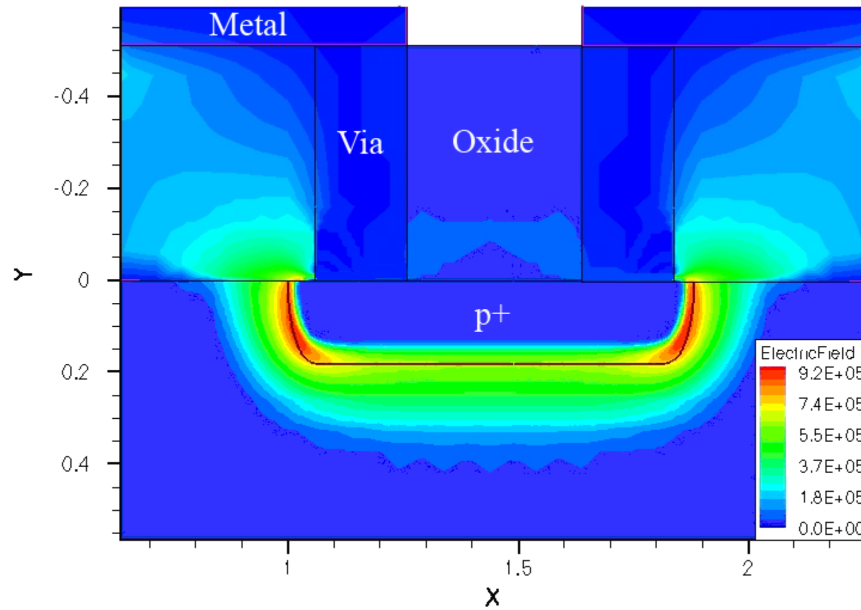


Figure 2.9: ISE-TCAD simulation of electric field distribution across a pn junction formed by consecutive implantation and diffusion steps. A uniform field exists in the planar junction region but the field is significantly higher in the curved regions, resulting in premature breakdown and in a higher avalanche probability in these areas. Field strengths are in V/cm and coordinates are in microns.

2.5. Solid-State Guard Rings for High-Field Devices

A robust and reliable Geiger-mode SPAD device must withstand the high electric fields associated with avalanche breakdown. Moreover, it must be able to sustain instantaneously high current densities repeatedly without performance degradation. Lastly, when integrated in arrays or with their supporting circuitry on the same die, these avalanches must not interfere with the operation of neighboring pixels, or with the analog or digital peripheral circuits. Specially-design guard-rings have been designed to meet these challenges. Another critical role for SPAD guard rings is

the elimination or reduction of the effects of curved junctions, which arise out of standard device processing steps, such as implantation and diffusion. It is the guard ring structure that mostly distinguishes between SPAD implementations, determines their fill factors and often their noise performance.

In this section we will describe several guard rings that have been implemented either in SPADs or in other high-field devices.

One of the first junction-planarization methods used in high-field devices was the beveled junction (Figure 2.10), whereby the junction was cleaved so that only the planar junction surface remains [92], [93]. This technique is not suitable for mass production or for arrays of devices.

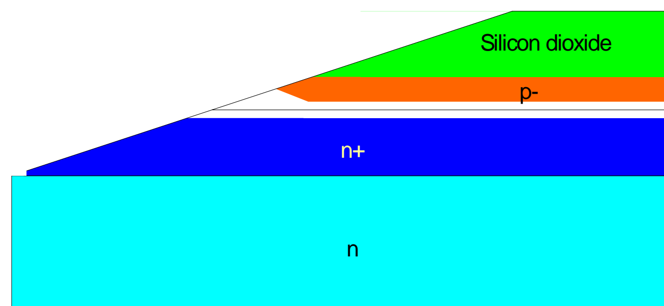


Figure 2.10: Beveled junction [93]. Depletion region is shown in white.

An extension of this guard-ring structure is the mesa structure, whereby etching and subsequent filling with a dielectric physically planarizes and isolates adjacent junctions (Figure 2.11) [94]. While arrays of devices using this technique have been demonstrated, the fill factor and pitch are adversely impacted.

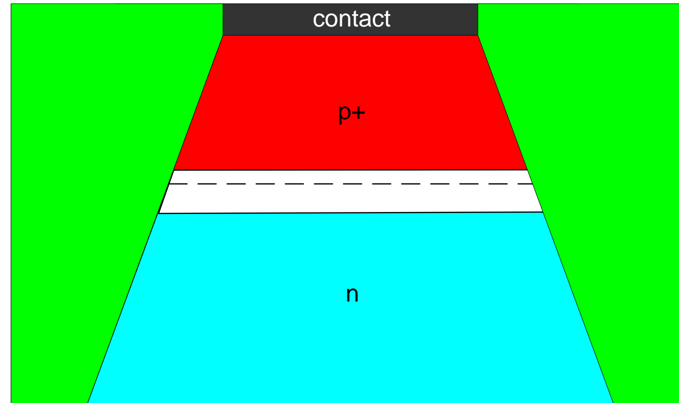


Figure 2.11: Mesa isolation structure.

The straddled junction [65] assures junction planarity by extending the high-doped p^+ region beyond a deep n-doped region. Due to the confinement of the electric field, breakdown only occurs at the planar interface.

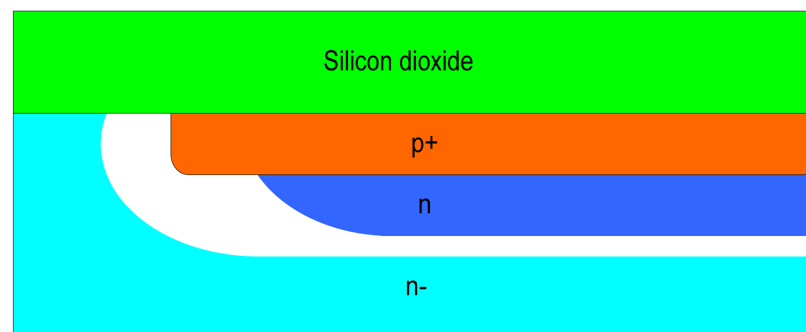


Figure 2.12: Straddled junction (doping polarities exchanged for compatibility with other structures shown here) [65].

A field-limiting ring (Figure 2.13) has been used in high-voltage devices to prevent premature breakdown in the curved junction regions [95]. This is accomplished by extending the depletion region of the device at the junction edges and thereby reducing the localized electric field.

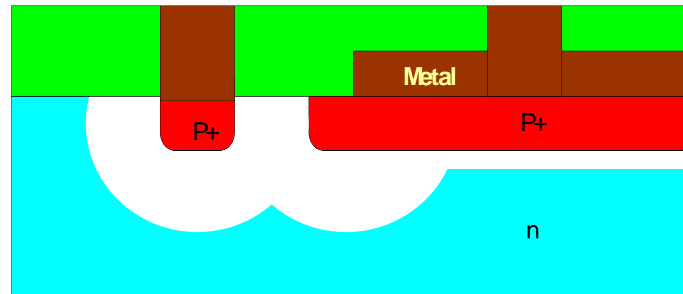


Figure 2.13: Field-limiting ring [95].

Silicon micro-strip detectors often employ a metal overhang guard ring [96]. Similarly to the field-limiting ring, the aim of this scheme is to extend the depletion region at the curved junction edges, thereby reducing the high electric field. This is done by applying a bias above the curved edge through a thin dielectric layer (Figure 2.14).

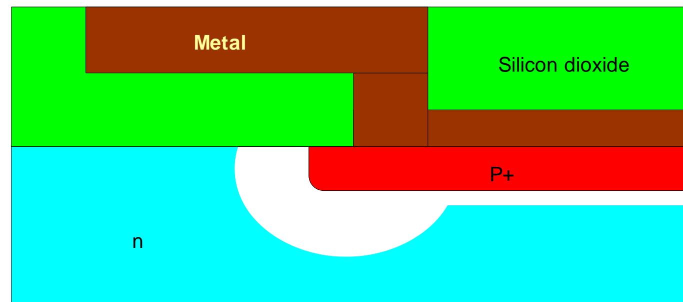


Figure 2.14: Metal overhang guard ring [96].

Avalanche diodes have been using diffused-ring structures since the 1960's (Figure 2.15) [97]. This is the most common guard-ring in SPADs. Edge breakdown is prevented by implanting a low-impurity ring at the edges of the junction. This extends the region across which the electric field develops, thereby decreasing it at the edges.

Diffused guard-rings are compatible with standard processing steps. However, they are inefficient in area, and result in decreased fill factors.

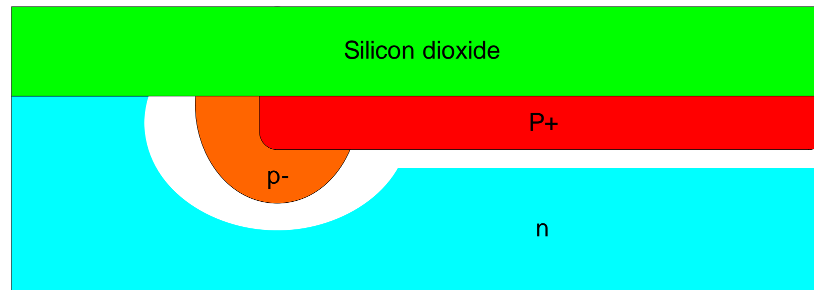


Figure 2.15: Diffused guard ring [70].

2.6. Geiger-Mode Single-Photon Avalanche Diodes: Principle of Operation

When a pn junction is reverse-biased beyond its breakdown voltage, a high electric field develops across its depletion region. As long as no free carriers enter the high-field area, no current flows. However, if a photon of sufficient energy is absorbed inside or in proximity to the depletion region, generating an electron-hole pair, a self-sustaining avalanche may develop through a series of impact ionizations, as discussed in the previous sections. This avalanche can be sensed electrically, generating an output signal corresponding to the photon absorption event.

There are several important distinctions between sub-Geiger device operation (APD mode) and Geiger-mode operation:

- In sub-Geiger-mode, the device operates as a linear amplifier. However, at high gains, the statistical variations in the gain introduce a multiplication noise.

SPADs, on the other hand, operate as bistable devices. In this case the positive-feedback amplification process is exploited to quickly generate an avalanche pulse.

- APDs can, within the limits of their sensitivity, resolve the number of impinging photons. SPADs produce a single avalanche pulse regardless of the number of absorbed photons.
- The build-up of an avalanche in APDs is relatively slow. In SPADs, where the high electric field quickly accelerates the free charges, avalanche build-up is fast – less than a nanosecond. Therefore, precise timing information can be extracted from these devices.
- Whereas APDs may be operated in continuous mode, and their current can be assumed to be proportional to photon flux, SPADs must be operated with a quenching mechanism to shut-off the avalanche and prepare the device for a subsequent photon. Therefore, the output of a GM-SPAD is a train of pulses corresponding to the arrivals of individual photons.
- Due to the high number of charge carriers crossing the junction, new noise mechanisms come into play. The physics and statistics of the resulting afterpulsing will be discussed in sub-section 2.8.3.

In the following section we will review mechanisms for quenching the avalanche and for recharging the junction.

2.7. Avalanche Quenching and Junction Recharge

Unlike APDs, which can operate standalone, whereby their current is proportional to the incoming photon flux. GM-SPADs require peripheral circuitry in order to quench the avalanche and recharge the device in preparation for the arrival of subsequent photons. Various methods have been used to carry out these tasks as described next.

2.7.1. Passive Quenching and Recharging

The simplest method to quench an avalanche is connecting a series quenching resistor, R_q, to one of its terminals, in a scheme known as passive quenching. The equivalent circuit in this case is shown in Figure 2.16 [98, 99]. The depletion region's capacitance, C_d, together with any parasitic capacitance, $C_{parasitics}$, is initially charged to a voltage V_{HIGH}, which is higher than the breakdown voltage of the junction, V_B. The diode is connected to the supplies by the quenching resistor, R_q, and by the series resistance, R_s. Typically $R_s << R_q$. The onset of an avalanche is modeled by closing the switch. At this phase the current through the diode (through R_d) is:

$$i_d(t) = \frac{V_d(t) - V_B}{R_d}$$

Equation 2.27

$V_d(t)$ decrease exponentially, as the junction and parasitic capacitances are discharged, until a steady-state is achieved, when:

$$I_f = \frac{V_{HIGH} - V_B}{R_q + R_d} \cong \frac{V_E}{R_q}$$

Equation 2.28

and

$$V_f = V_B + R_d I_f$$

Equation 2.29

Here, V_E is the excess voltage above the breakdown voltage. The approximation of Equation 2.28 is made based on the discussion in section 2.8.1.4.

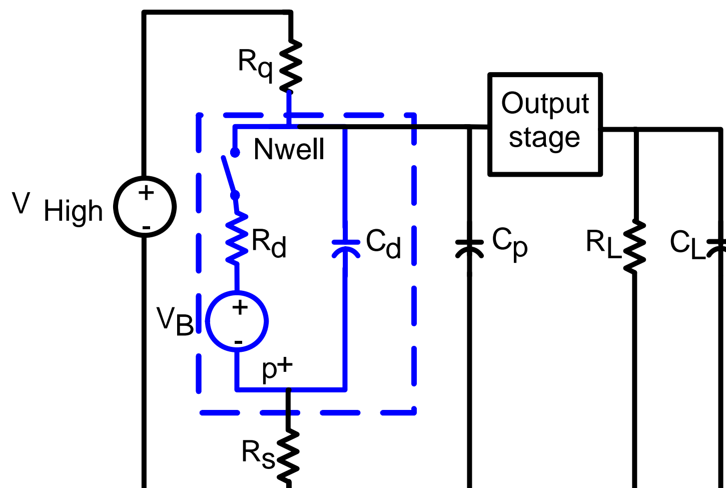


Figure 2.16: Equivalent circuit for passively-quenched passively-recharged SPAD.

We can also calculate the quenching time, which is determined by the network comprised of R_d, C_d, R_q and $C_{parasitics}$:

$$T_q = \left(C_d + C_{parasitics}\right)\frac{R_q R_d}{R_q + R_d} \cong \left(C_d + C_{parasitics}\right) R_d$$

Equation 2.30

Assuming R_d = 1 kΩ, and the total capacitance is 1 pF, the quenching time constant is 1 ns. In integrated devices, the capacitance can be on the order of 50 fF resulting in very short quenching times, on the order of 20 ps.

Equation 2.28 is used to determine the required resistance of R_q. In order to assure that the avalanche is quenched, we must ensure that none of the carriers entering the high field region may impact ionize. A good rule-of-thumb for a latching I_f which ensures avalanche termination, is 100 μA or less [58, 98]. If, for example, the SPAD is operated at V_E = 5V, R_q > 50 kΩ should be used. Larger resistors will avoid statistical quenching fluctuations, but will also result in larger parasitic capacitances. Typical quenching-resistor values are approximately 120 kΩ, resulting in a quenching voltage, V_q, slightly higher than the breakdown voltage.

The total charge flowing during a passively-quenched avalanche can now be determined:

$$Q_{pc} = \frac{V_{HIGH} - V_B}{C_d + C_{parasitics}} \cong \frac{V_E}{C_d + C_{parasitics}}$$

Equation 2.31

Equation 2.31 is very important because it captures some crucial relationships in the design of a SPAD. As discussed later in this chapter, avalanche charge minimization id desirable not only to reduce the power dissipated during each avalanche, but also to

suppress some noise mechanisms. For that reason $C_{parasitics}$ must be reduced to a minimum. The excess voltage, which is related to the quantum efficiency of the device, also cannot be made arbitrarily high for the same reasons.

After the avalanche has depleted the junction of all free carriers and has been quenched, it is recharged. In a passive recharging scheme, this is simply done through the quenching resistor. We can calculate the time required to recharge the junction simply by the corresponding RC time-constant:

$$T_r = \left(C_d + C_{parasitics}\right) R_q$$

Equation 2.32

There are some important implication to Equation 2.32:

- If the interval between photon arrivals is smaller than T_r, based on the discussion in section 2.8.1.4, the avalanche initiation probability, and consequently, the detection efficiency will vary with time. This is highly undesirable.
- In the same scenario, the output pulse heights will also vary. This may result in some pulses being of smaller amplitudes than others. The implications of this are twofold: some pulses may be missed by the sensing circuitry; because most sampling circuits have a constant amplitude threshold, the variable amplitude will introduce timing uncertainty (jitter). This can be resolved by using a constant fraction discriminator, but this option is costly and complex [100, 101].
- Because R_q is determined by the latch-off current and C_d is an artifact of the junction itself, the only way to reduce T_r in a passively-quenched SPAD is by

reducing the parasitic capacitance. The most effective way of doing this is by integrating the quenching resistor on the same die as the SPAD itself [102-104].

Lastly, the choice of the resistor element can affect the device performance, as follows:

- An off-chip resistor can have highly-precise resistance but will add parasitic capacitance. Moreover, it is not suitable for dense array operation.
- On-chip resistors typically have very high temperature coefficients and take up large areas.
- Transistors are a good choice but require a high-quality CMOS process and clean voltage supplies. Moreover, they are limited by the value of VDD of the specific technology. This sets a limit on the allowed excess voltage with which the SPAD can operate.

2.7.2. Active Quenching and Recharging

In order to overcome the difficulties of passive quenching and recharging, more complex circuits, known as active-recharge circuits were introduced [75, 98, 105-109]. In an active quenching scheme, an external circuit senses the onset of an avalanche. For example, a comparator may sense a voltage drop, and quickly trigger a reduction in the SPAD voltage to a level below the breakdown of the device. While active quenching schemes can be effective in large and slow SPADs, their

effectiveness has been shown to be limited if a high overbias is required [58] or in cases where the junction capacitance is relatively small [70].

A more common peripheral circuit is the active recharge circuit. The idea behind the various implementations of this scheme is that while complete quenching can only be achieved via a high resistance, device recharge can be accelerated if R_q in Equation 2.32 is replaced by a smaller resistor. Therefore, two resistors are used – one for quenching and one for recharge. Analog circuitry senses the onset of an avalanche and then disconnects the quenching resistor and connects the smaller recharge resistor. Another circuit senses the completion of the recharge process and quickly disconnects the smaller resistor and re-connects the quenching resistor. The inherent risk here is the arrival of a photon after recharging is fully or almost complete, but before the disconnection of the recharging resistor. This results in an avalanche which is not quenched.

One of the benefits of an actively-recharged SPAD is that the delay between the onset of an avalanche and the time it is recharged can be controlled. This is desirable in terms of device noise, and especially afterpulsing. The main disadvantage of present active-recharge scheme is their large size and high power consumption, both of which are prohibitive to large-scale SPAD arrays.

2.7.3. Negative-feedback Quenching

A third quenching scheme [110, 111] integrates a resistive layer on top of the SPAD structure. This provides a negative feedback in the avalanche area. As the avalanche progresses, charges start accumulating across the interface layer between the silicon and the high-resistivity material. An electric field results with an opposite polarity to that of the junction field, thereby screening the original field. The avalanche process is decelerated and quickly terminated. This negative-feedback scheme has not been extensively explored and may provide significant benefits in large arrays of SPADs with low power consumption.

2.8. Figures of Merit

The performance of SPADs can be measured by various metrics. Clearly, it is impossible to optimize all simultaneously, so a clear understanding is required of the various figures of merit, their physical manifestation, and the prioritized performance parameters of the target application. In the next sub-sections we will explore these various figures of merit.

2.8.1. Photon Detection Probability and Spectral Response

In most applications, the most important figure of merit is the photon detection probability. This is the probability that a photon impinging upon the detector will generate a detectable electrical output. In order for a photon to be detected, all of the following events must take place:

a) A photon impinging onto the surface of the detector will be transmitted into the body of the detector and not reflected back.

b) The photon will either be focused onto (if optics are part of the detector) or will directly hit the absorbing (active) region of the pixel, rather than the peripheral circuitry and interconnect areas.

c) The photon will be absorbed in or near the depletion region of the SPAD.

d) A detectable avalanche will be initiated.

Next, we will examine each of these conditions analytically.

2.8.1.1. Transmittance from air to silicon through dielectric layers

As part of standard processing, a dielectric insulation layer, usually comprised of SiO_2 is deposited on top of the SPAD device in order to maintain planarization and thus reduce mechanical stresses, and to protect the sensitive circuitry. Moreover, some processes also include a silicon nitride passivation layer for added mechanical strength and to prevent ionic contamination arriving to the substrate. Through reflections, these

dielectric layers significantly affect the probability that an impinging photon be transmitted onto the junction.

The transmittance of light across an air - silicon dioxide - silicon interface has been studied for a long time (e.g., [112]). Here we use Wolffenbuttel's expression to calculate the wavelength-dependent transmittance from air to silicon through a thin silicon-dioxide layer, assuming normal incidence [113]:

$$T \propto \frac{4n_{si}}{\left[(n_{si}+1)\cos\delta - \frac{k_{si}}{n_{ox}}\sin\delta\right]^2 + \left[\frac{n_{si}+n_{ox}}{n_{ox}}\sin\delta - k_{si}\cos\delta\right]^2}$$

Equation 2.33

where n_{ox} is the refractive index of the oxide layer, n_{si} and k_{si} are the real and imaginary parts of the silicon index of refraction, and δ is the phase shift across the oxide layer:

$$\delta = \frac{2\pi}{\lambda} n_{ox} d$$

Equation 2.34

with d the thickness of the silicon dioxide film.

Three important conclusion arise from Equation 2.33 and Equation 2.34:

- Surface reflections can account for up to 40% loss in detection probability [70].
- They can also create large swings in spectral response, with up to 30% change in response within a 50 nm spectral range.

- The thickness of the silicon-dioxide inter-layer dielectric layers in deep-submicron CMOS technologies is determined by Chemical-Mechanical-Polishing (CMP) [91], which is used to planarize the wafer. This results in relatively large thickness tolerances, on the order of ±2 μm in a 6-metal process [114]. This variation corresponds to a large variation in the transmittance of the same wavelength in dies manufactured on different wafers, and constitutes a problem in commercial CMOS implementations.

These deficiencies can be addressed by using an anti-reflection coating which will prevent reflections at the air-SiO_2 interface.

2.8.1.2. Device fill factor

We define the pixel fill factor as the percentage of the pixel area which can detect a photon if one impinges upon it:

$$FillFactor = \frac{ActiveArea}{PixelArea}$$

Equation 2.35

Fill factors in SPADs are mainly determined by: a) the guard ring structures, b) the active area, and c) the interface and quenching/recharging scheme. Because the guard ring width and area are approximately constant for a given field intensity, regardless of the active area, as the latter increases, fill factors increase. Passive quenching usually has the advantage of higher fill factors compared with the more

complex active quenching. If bumps must be inserted in each pixel to interface the SPAD with the peripheral circuitry, fill factors are adversely affected.

It is noteworthy that usually reported detection efficiencies disregard the SPAD fill factor. This is understandable in the case of single-pixel devices or when a very small number of pixels are integrated. However, when large arrays are concerned, there is to date no proven method to fully compensate for the low fill factors of approximately 1%, and, consequently, fill factors should be taken into account when considering single-photon detection efficiencies.

2.8.1.3. Absorption probability

In order for a photon, which has been transmitted into the junction, to be absorbed and generate an electron-hole pair, it must have an energy greater than the bandgap of the absorbing material, or a wavelength than the bandgap or cut-off wavelength [45]:

$$\lambda_g = \frac{1.24}{E_g}$$

Equation 2.36

where the wavelength is in microns and the bandgap energy, E_g, is in eV. For silicon, E_g=1.12 eV and λ_g=1107 nm.

In addition to the minimum-energy requirement, both energy and momentum must be conserved (Figure 2.17). Even though silicon is an indirect-bandgap material,

it is an efficient detector, because the excitation and subsequent thermal relaxation occur in sequence rather than simultaneously.

The absorption probability per length is described by the absorption coefficient, α (Figure 2.18). Below 360 nm (3.4 eV) no vibrational energy is required to assure the transition. This explains the increase in the absorption coefficient below this wavelength.

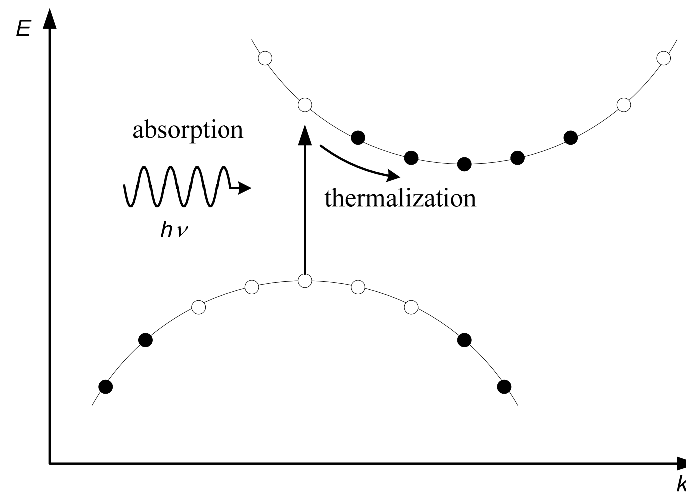


Figure 2.17: Photon absorption in an indirect-bandgap semiconductor.

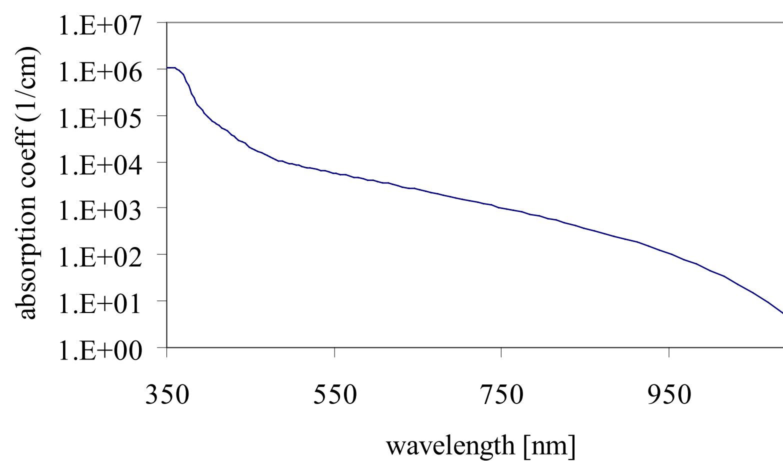


Figure 2.18: Silicon absorption coefficient [115].

The probability for a photon of wavelength λ to be absorbed after traversing a distance z in the silicon substrate is given by:

$$P_{abs}(\lambda, z) = 1 - \exp(-\alpha(\lambda) z)$$

Equation 2.37

This function is plotted for various wavelengths in Figure 2.19. The probability for the photon to be absorbed in a depth region between z and $z+\Delta z$, shown as broken lines in the figure, is simply:

$$P_{abs}(\lambda, z \rightarrow z + \Delta z) = \exp(-\alpha(\lambda) z) - \exp(-\alpha(\lambda)(z + \Delta z))$$

Equation 2.38

Figure 2.19 shows that shallow junctions with a narrow depletion region are expected to have peak sensitivity in shorter wavelengths. In order to detect long wavelengths, absorption regions of one or more microns are required.

To illustrate the effect of junction depth and width on spectral response, we calculate the absorption probability for two junctions (Figure 2.19). One is a shallow junction, 0.2 μm deep with a 0.2 μm-wide depletion region (these are typical parameters for the p^+/N-well SPAD described later in this work). The other is a deeper junction, 0.8 μm deep with a 1 μm-wide depletion region (an example of such a junction is the N-well/substrate junction in the SPAD device described in the next chapter). Typically, shallow junctions require higher doping, resulting in narrower depletion regions.

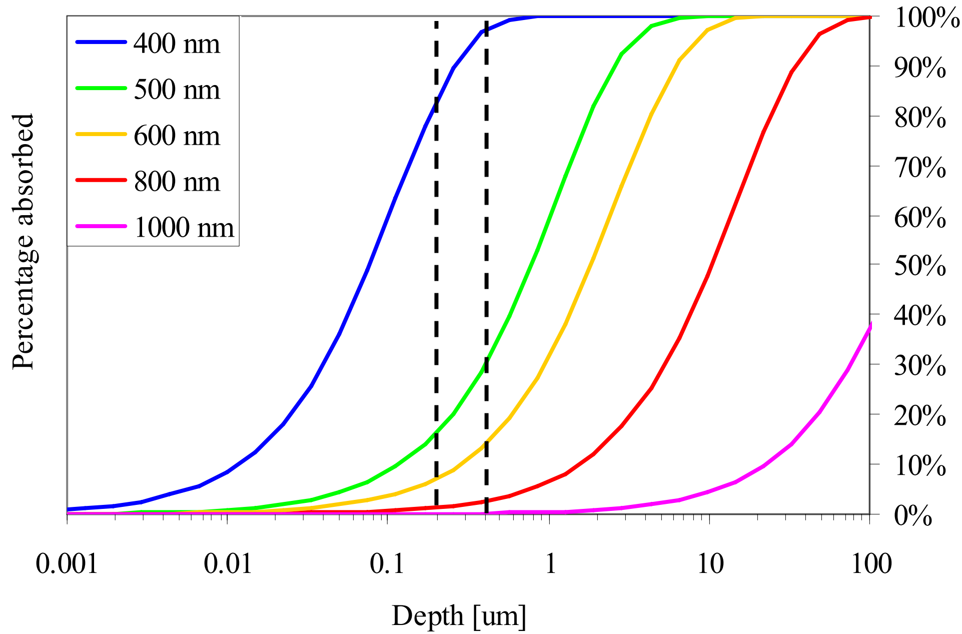


Figure 2.19: Absorption probabilities for photons of various wavelengths impinging from the surface.

We see that the shallow junction's response peaks at 430 nm while the deeper junction peaks at 510 nm, as discussed above. We also see that the absorption probability for the deeper junction is higher than for the shallow one. This is due to the wider depletion region, allowing for a longer interaction length between photons and the lattice.

In order to increase the absorption width in a given junction, a higher reverse-bias must be applied. For a one-sided linearly-graded junction, reverse-biased at $-V$, we use Equation 2.10 and Equation 2.38:

$$P_{abs}\left(\lambda, x_j \rightarrow x_j + w_d, V\right) \approx \exp\left(-\alpha\left(\lambda\right)x_j\right)\left[1-\exp\left\{-\alpha\left(\lambda\right)\left(\frac{3\varepsilon_s V}{2qa}\right)^{1/3}\right\}\right]$$

Equation 2.39

In addition to its effect of the depletion region's width, the applied voltage determines the probability of a photogenerated charge to initiate an avalanche. This is discussed in the next sub-section.

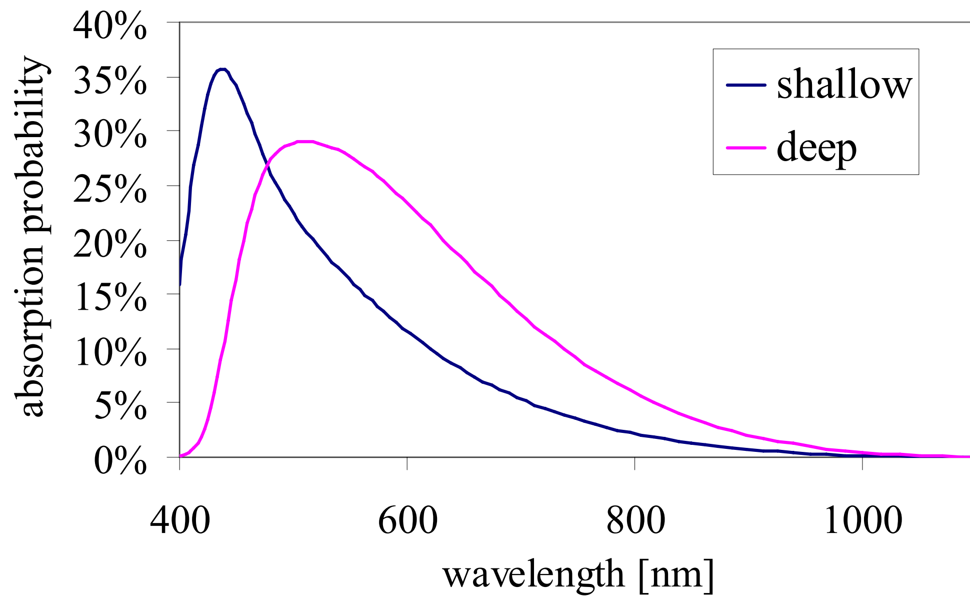


Figure 2.20: Absorption probabilities a shallow and deep junction

2.8.1.4. Avalanche Initiation and Propagation

The foregoing discussion, and specifically Equation 2.13, assumed a steady-state condition. For single-photon detection, the avalanche initiation process is of

utmost interest since it determines the photon detection probability. Once a photon has been absorbed in or near the depletion region of a junction which is biased beyond its breakdown voltage, the primary photogenerated charges may create additional "descendent" charges through a series of impact ionizations. Alternately, the primary charges may recombine. Even if secondary charges are generated, the avalanche may extinguish itself before being detected. The probability for a photogenerated electron-hole pair to initiate a full avalanche initiation is the topic of this section.

We follow the formalism of avalanche initiation probability which has been developed by Oldham [99]. We define the function $P_{be}(x)$ as the probability that an electron starting in position x in the depletion layer will have an infinite number of descendents, i.e., will trigger an avalanche (Figure 2.5). $P_{bh}(x)$ is the analogous function for holes. If the primary electron-hole pair is generated in position x, then the probability that neither charge causes an avalanche is given by:

$$1-P_{bp}=\left(1-P_{be}\right)\left(1-P_{bh}\right)$$

Equation 2.40

and the probability that either the electron or hole triggers an avalanche is:

$$P_{bp}=1-\left(1-P_{be}\right)\left(1-P_{bh}\right)$$

Equation 2.41

The probability $P_e(x+\Delta x)$ that an electron starting in position $x+\Delta x$ triggers an avalanche can be computer from three terms:

(i) The probability $P_e(x)$ that it reaches position x and triggers an avalanche there; and

(ii) The probability that in the transit between $x+\Delta x$ and x it creates an electron-hole pair via impact ionization and this pair initiates an avalanche; minus

(iii) The joint probability of the two events coinciding.

We can now write the expression:

$$\begin{aligned} P_{be}(x+\Delta x) = P_{be}(x) + \alpha\Delta x\left[P_{be}(x) + P_{bh}(x) - P_{be}(x)\cdot P_{bh}(x)\right] \\ -P_{be}(x)\alpha\Delta x\left[P_{be}(x) + P_{bh}(x) - P_{be}(x)\cdot P_{bh}(x)\right] \end{aligned}$$

Equation 2.42

When the electron ionization coefficient is much higher than the hole ionization coefficient, most of the ionizations are achieved by electrons. The ionization coefficient ratio (ionization ratio) should therefore be as small (or as large) as possible. Although a small ionization ratio (no hole ionizations) decreases the gain because holes do not take part in the avalanche, it has several significant benefits [45]:

- It reduces the avalanche build-up time.
- It reduces the randomness of avalanche build-up and therefore decreases the jitter.
- It reduces the probability for avalanche extinction before the avalanche is detected.

The ionization ratios for silicon, InGaAs/InP and Ge avalanche photodiodes have been measured to be 0.02, 0.35 and 0.7 [116]. Therefore, silicon is an excellent choice as an APD multiplication material.

Equation 2.22 can be more conveniently written in differential form:

$$\frac{dP_{be}}{dz} = (1 - P_{be})\alpha P_{bp}$$

$$\frac{dP_{bh}}{dz} = (1 - P_{bh})\beta P_{bp}$$

Equation 2.43

These coupled differential equations can be solved with the boundary conditions:

$$P_{be}(0) = 0$$
$$P_{bh}(W) = 0$$

Equation 2.44

In order to calculate the avalanche initiation probability in a one-sided linearly-graded junction, we plug Equation 2.8 and Equation 2.18 in Equation 2.43:

$$\frac{dP_{be}}{dx} = 7.16\times10^{4}\exp\left(-1\times10^{6}\frac{{w_d}^{2}}{V(w_d - x)}\right)(1 - P_{be})(P_{be} + P_{bh} - P_{be}P_{bh})$$

$$\frac{dP_{bh}}{dx} = -3\times10^{6}\exp\left(-2\times10^{6}\frac{{w_d}^{2}}{V(w_d - x)}\right)(1 - P_{bh})(P_{be} + P_{bh} - P_{be}P_{bh})$$

Equation 2.45

These equations can be solved numerically, e.g., using Matlab, to solve for the avalanche initiation probability for electrons, holes, and overall as a function of primary pair generation (Figure 2.21).

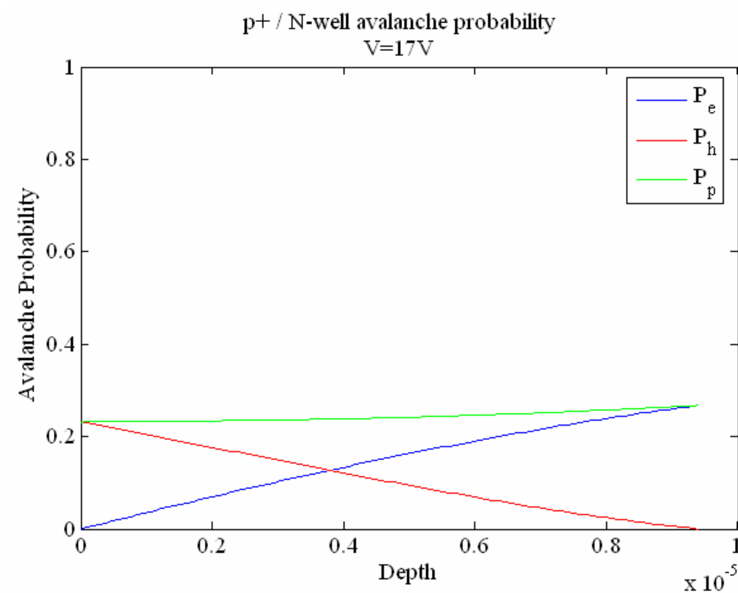


Figure 2.21: Avalanche initiation probability as a function of primary-pair generation depth in a linearly-graded junction.

If we assume a uniform primary-pair generation probability across the junction (as is the case in narrow junctions), we can average the calculated avalanche initiation probabilities calculated above. Figure 2.22 shows the effect of junction bias on this average probability. As can be seen, although we previously assumed the breakdown voltage to be a discrete point, in reality, the probability of initiating a sustainable avalanche is a continuous function.

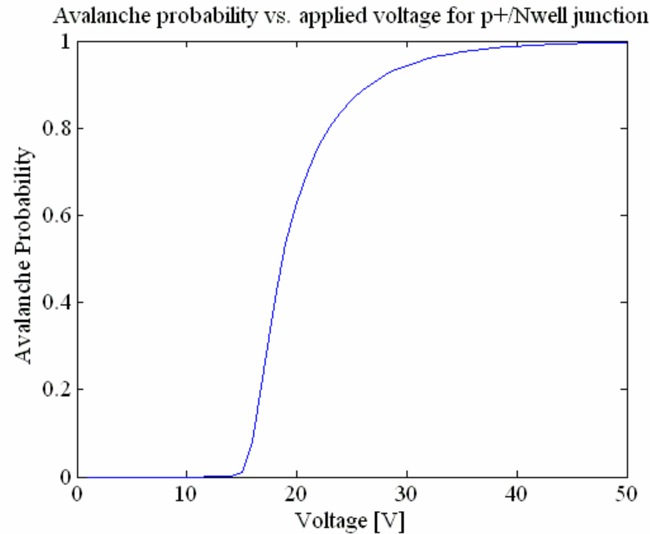


Figure 2.22: Avalanche initiation probability as a function of voltage in a narrow junction.

As impact ionizations build up the avalanche current, the current is limited by the SPAD resistance. This resistance is comprised of three elements:

a) the ohmic resistance from the contact to the neutral region of the junction,

b) the resistance of the neutral region, and

c) the space-charge resistance, which is due to charge crowding during the avalanche.

Referring to Figure 1.14, the first element can be readily extracted from resistivity data of the N-well material, which is usually provided by the foundry. This resistance increases with contact-to-junction distance and decreases with doping concentration. The second and third elements increase with smaller diode area. The latter is given by Sze and Shockley [117] as:

$$R_d=1/GN;\ N=A/W^2$$

Equation 2.46

where *N* is the number of unit cubes in the depletion region with cube edge, $\sqrt{A}$, equal to the depletion width, *W*, and *G*= 40 μmhos for silicon.

2.8.1.5. Surface recombination

Finally, we must consider surface recombination. Since the lattice structure is broken at the surface (and at any silicon-SiO_2 interface), the incomplete bonds may cause free carriers to recombine, and become unavailable for avalanche generation. This effect is dominant in short wavelengths which are absorbed close to the surface and thus reduces their detection efficiency [85]. Quantitatively, it has been shown that for shallow junctions, where the depletion region depth z_0 is much shallower than the electron diffusion length, the probability for an electron generated at depth $z < z_0$ in the p^+ anode of a SPAD to diffuse to the depletion region is [70]:

$$P_{diff,e}\left(z,z_0\right)=\frac{z}{z_0}$$

Equation 2.47

We do not need to take into account holes generated at the other side of the depletion region because they are blocked by n-type concentration gradient from diffusing to the depletion region.

The overall detection probability can now be expressed as:

$$DP(\lambda,V)=T(\lambda)\left[P_e(z_0,V)\int_0^{z_0}\frac{z}{z_0}\alpha(\lambda)\exp(-\alpha(\lambda)z)dz+\int_{z_0}^{z_w(V)}\alpha(\lambda)\exp(-\alpha(\lambda)z)P_p(z,V)dz\right]$$

Equation 2.48

The first term corresponds to a photon being absorbed in the anode region and diffusing to the depletion region. The second term corresponds to photo-absorption in the depletion region and a resulting hole-initiated avalanche (hole ionization coefficient is higher in silicon).

2.8.2. Dead Time

Geiger-mode SPADs switch between their active operational phase, when the device is charged above the breakdown voltage, and a dead time phase, when, following an avalanche, they are biased below breakdown, and an impinging photon will not produce an avalanche. Device dead time has a number of implications in terms of device performance:

- Maximal photon flux: Assume n_{imp} photons impinge upon the SPAD active area per second, that their detection probability is *D.P.*, and that the dead time is t_{dead}. Some of the photons' arrivals will coincide with the dead time of the SPAD, so that only $n_{det} < n_{imp} \times D.P.$ photons will be detected. The total dead time per second will be $n_{det} \times t_{dead}$ so the percentage of time the SPAD is "dead" is $(n_{det} \times t_{dead}) / 1$. We can then write (neglecting noise):

$$n_{\det} = \left(1 - n_{\det} \times t_{dead}\right) n_{imp} \times D.P.$$

Equation 2.49

and

$$n_{\det} = \frac{n_{imp} \times D.P.}{1 + n_{imp} \times D.P. \times t_{dead}}$$

Equation 2.50

- Uniformity of detection probability: In passively-recharged devices, the dead time actually consists of a gradual recharging of the junction. As discussed in Section 2.7.1, this results in a non-uniform detection probability.
- Noise: A long dead time makes it possible to release charges which were trapped during the avalanche without incurring false avalanches (afterpulses). Conversely, SPADs with very short dead times may exhibit self-sustaining afterpulsing, resulting in excess noise (see Section 4.8).

2.8.3. Noise

GM-SPADs exhibit noise which arises out of its unique operating conditions and is therefore inherently different than the noise found in other imagers, such as CCDs, APDs and CMOS imagers. Because the output of SPADs is binary in format, noise in these devices appears as “false” pulses. The mechanisms responsible for their generation are described below.

2.8.3.1. Thermal generation

Because of the relatively large bandgap of silicon, direct thermally-activated transitions of electrons from the valence to the conduction band, is very unlikely. However, due to defects and contaminations, electronic levels called traps do exist within the bandgap (Figure 2.23). These make it possible for electrons to become thermally excited to the conduction band. If this occurs in or near the depletion region, an avalanche may be triggered.

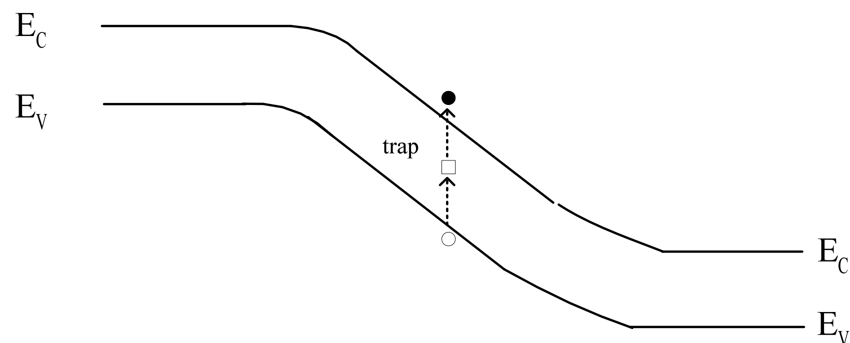


Figure 2.23: Trap-assisted generation.

The thermally-activated generation has been analyzed via the Shockley-Read-Hall (SRH) theory [118]. The model calculates the probabilities of four types of events: hole capture by and release from a trap, and the equivalent scenarios for an electron. Due to the high electric field in the depletion region, SRH generation is enhanced by the Poole-Frenkel effect, which effectively lowers the trap energy at the presence of an electric field [64, 119]. The derivation of the trap-assisted generation rate is outside the scope of this work. It is given by [120]:

$$DCR_{SRH} = S\int_{z_0}^{z_w} \frac{n_i}{\tau_g} P_{bp}\left(z\right) dz$$

Equation 2.51

Here, S is the surface area of the junction, n_i the intrinsic doping of silicon, and τ_g is the generation lifetime:

$$\tau_g = \frac{1}{1+\Gamma}\left[\tau_{e0} exp\left(-\frac{E_t - E_i}{kT}\right) + \tau_{h0} exp\left(\frac{E_t - E_i}{kT}\right)\right]$$

Equation 2.52

in which τ_{e0} and τ_{h0} are the electron and hole recombination lifetimes, E_t is the energy of the dominant recombination center and E_i is the intrinsic energy level. Γ is the field effect function of the model [85]:

$$\Gamma = 2\sqrt{3\pi}\frac{\left|\xi\right|}{\xi_\Gamma}\exp\left(\left(\frac{\xi}{\xi_\Gamma}\right)^2\right)$$

Equation 2.53

in which ξ is the local electric field and

$$\xi_\Gamma = \frac{\sqrt{24m^*\left(kT\right)^3}}{q\hbar}$$

Equation 2.54

in which m^* is the effective mass for tunneling (0.25 m_0), m_0 is the free electron mass, k is the Boltzmann constant, T is the absolute temperature and $\hbar$ is the Dirac constant. n_i is temperature-dependent [85]:

$$n_i \propto T^{3/2} \exp\left(-\frac{\left[1.17 - \frac{4.73\times10^{-4} T^2}{(T+636)} \right]}{2kT} \right)$$

Equation 2.55

2.8.3.2. Band-to-band tunneling

The phenomenon of tunneling has been described in Section 2.3.2. It is especially prevalent in highly-doped junctions, where the depletion region is narrow and the electric field high (Figure 2.24). In the context of SPAD noise, Hurkx derived an expression for the dark count rate by band-to-band tunneling [119]:

$$DCR_{tunn} = -B|\xi|^{5/2} D\left(\xi, E, E_{fn}, E_{fp}\right) \exp\left(-\frac{\xi_0}{|\xi|} \right)$$

Equation 2.56

where ξ is the local electric field, B = 4 x 10^{14} $cm^{-1/2} \cdot V^{-5/2} \cdot s^{-1}$ is a temperature-independent coefficient,

$$\xi_0 \propto E_g^{3/2} = 1.9\times10^7\ V/cm$$

Equation 2.57

and

$$E_g(T) = E_g(0) - \frac{\alpha T^2}{T+\beta}$$

Equation 2.58

with α and β fitting parameters. D is 1 inside the depletion region and 0 elsewhere.

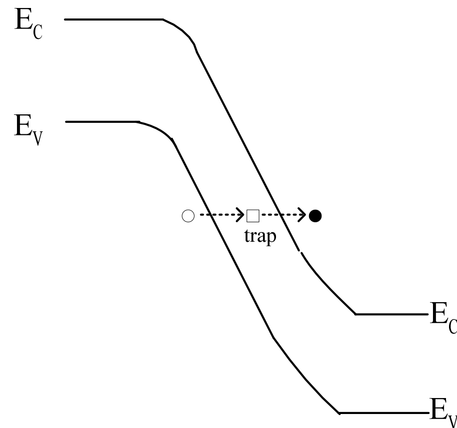


Figure 2.24: Band-to-band tunneling.

2.8.3.3. Afterpulsing

Afterpulsing is a unique noise source to GM-SPADs. It is also unique in the fact that it is correlated to previous avalanche. This has far-reaching consequences in terms of its implications in terms of device performance.

Afterpulsing results from the release of avalanche charges which get trapped in deep levels during a subsequent avalanche. This release is a random event with an exponentially decreasing probability [121]. The filled trap population at time *t*, after a seed avalanche at time t_s can be expressed as:

$$N_{traps}(t_s, t) = \sum_{i=1}^{n} A_i \exp\left(\frac{-[t - t_s]}{\tau_i}\right)$$

Equation 2.59

where A_i are the average number of traps of type *i*, which are filled immediately after an avalanche, and τ_i are their respective lifetimes. If the trap is released when the junction in charged above its breakdown voltage, a false count may result.

Trap lifetimes can be expressed as a function of temperature, *T*, using the Arrhenius equation [122]:

$$\tau_i(T) = \frac{1}{\sigma_i v(T) N(T)} \exp\left(E_{a,i} / kT \right)$$

Equation 2.60

where E_a is the trap activation energy (in the presence of the high electric field which decreases the effective activation energy compared to the low-field situation [121]), σ is the trap cross-section, *v(T)* is the average thermal velocity of the carriers, and *N(T)* is the effective density of states of the relevant band. $v \propto T^{1/2}$ and $N \propto T^{3/2}$ so $\sigma v N \propto T^2$.

A_i has been found to be approximately linear with the total avalanche charge [122]. This means that in typical cases, only a small percentage of the available traps are filled during an avalanche.

In order to find the number of filled traps at time *t*, we must add all the charges which have filled traps during subsequent avalanches, and account for their emissions:

$$N_{traps}(t) = \int_{t_s=-\infty}^{t} N_{traps}(t_s, t)\, dt_s$$

Equation 2.61

A SPAD with N_{traps} filled traps, will have an instantaneous avalanche probability related to the rate of trap emission:

$$p_{afterpulse}(t) = -\frac{dN_{traps}(t)}{dt} p_{bp} = -p_{bp} \frac{d}{dt} \int_{t_s=-\infty}^{t} \sum_{i=1}^{n} A_i \exp\left(\frac{-[t-t_s]}{\tau_i}\right) dt_s$$

Equation 2.62

where p_{bp} is the avalanche initiation probability, which we derived in Section .

In the integral above, the integral is in fact a summation over all prior avalanches and is proportional to the sum of charges flowing in these avalanches. These can result from the noise mechanisms we discussed above or from detected photons. Assuming an average Poisson-distributed photon arrival rate, r_{ph}, an average SRH rate $DCR_{SRH,}$ and an average tunneling rate DRC_{tunn}, because these events are independent Poisson processes, the average avalanche rate, excluding afterpulses, will be:

$$r_{counts} = p_{bp}(r_{ph} + DCR_{SRH} + DRC_{tunn})$$

Equation 2.63

Equation 2.62 can now be re-written:

$$p_{afterpulse}(t) = -p_{bp} \frac{d}{dt} \sum_{j=-\infty}^{0} \sum_{i=1}^{n} A_i \exp\left(\frac{-[t \cdot r_{counts} - j]}{\tau_i r_{counts}}\right) = -\frac{d}{dt} p_{bp} \sum_{i=1}^{n} \exp\left(\frac{-t}{\tau_i}\right) \sum_{j=0}^{\infty} \exp\left(\frac{-j}{\tau_i r_{counts}}\right)$$

Equation 2.64

which can be evaluated is the sum of an infinite geometrical series:

$$p_{afterpulse}(t) = \sum_{i=1}^{n} \frac{A_i \frac{p_{bp}}{\tau_i} \exp\left(-\frac{t}{\tau_i}\right)}{1-\exp\left(-\frac{1}{\tau_i r_{counts}}\right)}$$

Equation 2.65

Until now, we neglected the effect of afterpulses prior to time t. With their contribution, Equation 2.65 becomes:

$$p_{afterpulse}(t) = \sum_{i=1}^{n} \frac{A_i \frac{p_{bp}}{\tau_i} \exp\left(-\frac{t}{\tau_i}\right)}{1-\exp\left(-\frac{1}{\tau_i r_{counts}}\right)} + p_{bp} \int_{-\infty}^{t} p_{afterpulse}(t') \sum_{i=1}^{n} A_i \exp\left(-\frac{t-t'}{\tau_i}\right) dt'$$

Equation 2.66

This equation can be solved numerically. It assumes a very short dead time compared with the dark count rate and the trap lifetimes. It also assumes that following the dead time, the device is instantaneously charged to its final voltage. In reality, afterpulses may occur before completion of the recharge. This results in an effective reduction to the number of charges flowing and traps filled in an avalanche (A_i) but also in a reduced detection efficiency.

In many scenarios, such as when a pulsed laser is used as the excitation source, photon arrivals are expected periodically within a certain time window, Δt. For a detector with a dead time $t_{dead} > t_{rep}$, where the latter is the pulse repetition interval, an expression for the total afterpulsing probability has been derived by Ghioni [123]:

$$p_{afterpulse,T} = p_{av} \sum_{i=1}^{n} A_i \tau_i \exp\left(-\frac{t_{dead}}{\tau_i}\right) \frac{1-\exp\left(-\Delta t/\tau_i\right)}{1-\exp\left(-t_{rep}/\tau_i\right)}$$

Equation 2.67

Afterpulsing has several implications for SPAD performance:

- Amplification of noise: As was shown in Equation 2.66, afterpulsing results from a "seed" avalanche, regardless of its origin. Therefore, for every seed dark count, many others will follow, increasing the total number of dark counts. At certain conditions, afterpulsing may be a self-sustaining process, becoming the dominant noise source (see section 4.8).
- Reduction in detection efficiency: Two mechanisms contribute to reduction in sensitivity as a result of afterpulsing. In the case of a self-sustaining process with a passively-recharged SPAD, the junction may never be completely re-charged before an afterpulse appears. This corresponds to a variable and lower overbias than the DC overbias, and consequently to a lower detection efficiency. Secondly, if an afterpulse arrives within a dead time before the arrival of a photon (in either a passively- or actively-recharged device), the latter will not be detected because the junction would not have recharged.
- Limitation of device bandwidth and dynamic range: Because afterpulses are much more probable immediately after an avalanche, a hold-off time must be imposed on the SPAD. During this hold-off, implemented using active recharge circuits, the junction is biased below its breakdown voltage, to allow for the

release of the majority of the traps without initiating an avalanche. This hold-off time sets the inter-exposure interval, thus limiting the total number of counts per unit time.

Afterpulsing can be reduced in the following ways:

- Reduction of avalanche charge: this can be done by decreasing the junction volume, lowering the overbias or by using a fast quenching scheme. The former may result in optical setup difficulties while the latter reduces the detection efficiency.
- Utilizing clean processes where defect densities are low.
- Enforcing a hold-off time following an avalanche: Active recharge circuits can impose artificial delays before junction recharge. Such delays can ensure that an acceptable percentage of the trapped charges have been released following. Active-recharge circuits are typical expensive in silicon real-estate. Increasing the hold-off limits the dynamic range of the SPAD and lowers the number of counts per second it can detect.
- Operating with inter-photon arrival times which are much longer than the trap lifetimes, so that afterpulses can die out before the arrival of a subsequent photon.
- Implementing time-gating: If the expected time-of-arrival of the incoming photons is known to within a range, the SPAD can only be biased above breakdown for the expected time window. In this gated mode, the afterpulsing probability is reduced (because the junction is only biased for short durations).

2.8.3.4. SPAD Noise: Summary

SPAD noise sources are summarized in Table 2.1. It shows that it is possible to differentiate between the noise sources. Afterpulsing can be distinguished from the other sources by looking at the noise autocorrelation function. Heating the junction can isolate the dominant noise source, because the different noise mechanisms exhibit unique temperature dependence. It is important to note that, unlike other detectors, cooling the SPAD does not guarantee a reduction in noise. At lower temperatures, trap lifetimes increase, so afterpulsing becomes dominant. Different materials have specific optimal working temperatures which minimize the sum of all sources, depending on their bandgap energy, trap levels and defect densities. As it turns out, for silicon, the optimal temperature is approximately room temperature.

Table 2.1. Summary of SPAD noise sources

	SRH	**Tunneling**	**Afterpulsing**
Correlation	Uncorrelated	Uncorrelated	Correlated
Distribution	Poisson	Poisson	Exponential
Voltage dependence	Strong dependence	Very strong dependence	Strong dependence
Temperature dependence	Strong dependence	Weak dependence	Inverse dependence

Lastly, we note that SPAD noise also reduces the detection efficiency of the device. As discussed above, a necessary condition for photon detection is that no photons have arrived at least for a time t_{dead} before the photon's arrival. The corresponding detection probability reduction due to such a scenario is given by:

$$D.P._{with_noise} = \left(1 - \left[p_{afterpulse}\left(t_{dead}\right) + p_{SRH}\left(t_{dead}\right) + p_{tunn}\left(t_{dead}\right) \right]\right) \times D.P._{no_noise}$$

Equation 2.68

where the three probability events refer to the cumulative probability of a noise event during a duration t_{dead}.

2.8.4. Timing Precision

The time-spread of delays between the arrival of a photon and the clocking of the output electrical signal is known as the jitter of the SPAD. Jitter is important in TCSPC experiments as well as in applications where time-gating is used to reduce the effects of noise. The uncertainty in the time-of-arrival consists of several components:

- Unknown position of photon absorption within the depletion region [124-126]: The position of the photon's absorption (if within the depletion region) will determine the position of the primary electron-hole pair. It is from this point that the avalanche will start spreading until it produces sufficient current to be detected electronically. The dynamics of this spreading depends on this initial position – an avalanche will spread faster if generated at the center of the

depletion region than at the edge. This uncertainty follows a roughly Gaussian distribution, usually characterized by the Full-Width Half-Maximum (FWHM). This mechanism is dominant in large-area SPADs.

- Unknown diffusion times of the primary carrier to the depletion region: Carriers may be photo-generated outside the depletion region and, if within the diffusion length, may still reach the depletion region via a random walk, generating an avalanche. This scenario is less likely than absorption in the depletion region, but results in a "diffusion tail" in the device jitter. It is this element that often limits SPAD jitter performance.
- Uncertainty in sense amplifier's threshold: Ideally, all avalanches are expected to have the same amplitude and the threshold for timing the threshold should remain constant. In reality, avalanche amplitudes may vary, either due to incomplete recharging, or due to secondary effects, such as temperature variations. As with all electronic circuitry, the amplifier threshold will have a certain uncertainty. These two effects result in additional uncertainty in the timing of the photon arrival event.
- Limited resolution of time-to-digital converter (TDC): Following the conversion of the avalanche spike to a binary pulse, the precise time of the leading edge of this pulse must be determined. A sampling clock must be used to clock this edge. Since the clock frequency is limited, an additional jitter component is added. The jitter in high-precision SPADs, which is in the range of tens of picoseconds, is often limited by the resolution of the TDC.

In order to reduce the device jitter, the following steps may be taken:

- Reduce the volume of the depletion region, thereby minimizing the uncertainty in the primary carrier's position.
- Design the SPAD such that the electric field is highly localized. In addition, structures, such as retrograde doping, can prevent charges generated far from the depletion region, from diffusing to the depletion region.
- Use active recharge schemes so that avalanche amplitudes are identical. Also, utilize electrical isolation to reduce noise in the sense amplifier. Specialized circuitry has been designed, whereby the avalanche pulse is high-pass filtered to improve precision [127].
- Perform time-to-digital conversion on-chip to prevent jitter from inter-chip interconnects, such as wire-bonds. This requires the SPAD to be manufactured in a CMOS technology.
- Utilize leading-edge CMOS technologies capable of multi-GHz clocking, so that TDC resolution can be improved.

2.8.5. Dynamic Range

The lower-bound on the dynamic range of imagers is commonly determined by the noise-equivalent power (NEP). The sensitivity of imaging SPADs is the signal level for which their signal-to-noise level (SNR) is one, assuming that the noise is

Poisson distributed (i.e., that afterpulsing is negligible). The saturation level in SPADs, n_{max}, depends on the percentage of photons which can be afforded to be lost, i.e., the portion of photons which may impinge during the dead time. If this percentage is c, then:

$$1-c=\frac{n_{\det}}{n_{\max}\times D.P.}=\frac{1}{1+\left(n_{\max}\times D.P.\times t_{dead}\right)}$$

Equation 2.69

where n_{det} photons are detected by the SPAD, and the maximum allowed photon flux is:

$$n_{\max}=\frac{c}{\left(1-c\right)\times D.P.\times t_{dead}}$$

Equation 2.70

The dynamic range can now be calculated (neglecting the effect of afterpulsing):

$$D.R.=\frac{c}{\left(1-c\right)\times D.P.\times t_{dead}\times DCR}$$

Equation 2.71

The calculations above assumed no knowledge on the expected photons' times-of-arrival or their statistics. In many single-photon applications, such knowledge is available, making it possible to use time-correlated measurements. In this case, the dynamic range is a less useful metric, and a quality factor, F, is used instead [28, 128]

$$F=\sqrt{1+\frac{t_{dead}}{n_{photons}\times D.P.}}$$

Equation 2.72

where $n_{photons}$ impact the detector per unit time, and where ideally F is minimized to 1.

2.8.6. Cross-talk

When integrating SPADs in arrays, inter-pixel cross-talk must be minimized. Cross-talk exists in many devices, and may result from imperfect electrical isolation, from charge overflow or other mechanisms. SPADs are susceptible to two types of cross talk – optical and electrical.

Optical cross-talk results from electroluminescence which is a byproduct of hot-carrier recombination [129, 130]. We discuss the physical mechanism responsible for this effect in section 5.2. In brief, the number of photons emitted via this electroluminescence, is proportional to the total avalanche charge. The spectral component of the emitted photons is concentrated near the bandgap of the multiplication region, but has a non-negligible tail at shorter wavelengths [131, 132]. The electroluminescent photons are emitted isotropically and may be absorbed by adjacent SPADs.

Optical cross-talk may be minimized by reducing the avalanche charge, spacing pixels farther apart, or physically separating them, e.g., using a trench [133].

Electrical cross-talk may occur if the capacitance of the power lines is too small. In such a scenario, an avalanche will draw current from the common supply, causing a voltage spike in adjacent pixels. This spike may either increase or decrease

the overbias on these adjacent pixels, and thus alter their avalanche probability. Measurements from such a scenario are described in section 4.9.

2.9. Conclusions

This chapter provided a physical basis for understanding the operation of Geiger-mode single-photon avalanche diodes. We explained the mechanism of junction breakdown and elucidated the effects of junction curvature on junction breakdown. In order to prevent premature breakdown, various guard ring structures were studied, each with its advantages and drawbacks. We also developed a model for avalanche initiation once an electron-hole pair has been generated.

The second part of this chapter developed the figures-of-merit for SPADs. We described the device parameters which determine the detection probability and spectral response of the device. The significance of dead time, jitter and cross-talk were described, in the context of device physical attributes. The relevance of these metrics to some of the applications described in Chapter 1, are summarized in Table 2.2. Lastly, we developed expressions for the unique noise sources in SPADs, and emphasized their differences, both in terms of statistical distributions and in terms of their voltage and temperature behavior. These observations will allow us to better understand the experimental results described in Chapter 4.

Table 2.2. Relationship between SPAD figures-of-merit and select single-photon detection applications

Performance Metric	**Physical Attributes**	**Application Figure of Merit**	**Applications Most Affected**
Detection probability	Absorption layer thickness	Dynamic range (sensitivity)	Lidar, FCS, TCSPC
Spectral response	Absorption layer depth and thickness, material	Dynamic range (sensitivity) vs. wavelength	Fluorescence imaging
Dark current	Process quality	Dynamic range (sensitivity), false alarm probability	Lidar
Recharge time	Junction + load capacitance, recharge mechanism	Dynamic range (saturation), image acquisition time	FCS, 3D imaging
Timing spread	Junction diameter, readout electronics	Time resolution	Lidar, 3D imaging, TCSPC

Table 2.2 (cont.). Relationship between SPAD figures-of-merit and select single-photon detection applications

Afterpulsing	Junction + load capacitance, quenching mechanism	Dynamic range (saturation), image acquisition time	FCS, 3D imaging
Detection area	Process quality	Ease of optical alignment	Microscopy
Pixel pitch	Pixel structure, isolation scheme, process geometry	Image resolution	Array imaging

3. STI-BOUNDED SINGLE PHOTON AVALANCHE DIODE – ANALYSIS, MODELING AND SIMULATION

3.1. Introduction

In the previous chapters, we demonstrated that SPAD devices are designed with various performance trade-offs, depending on the target application. For example, in order to increase the detection probability and spectral response, wider depletion regions must be used. This increase in depletion region volume must come at the expense of timing accuracy and dead time. Another trade-off comes when a reduction in afterpulsing results in smaller device bandwidth, due to an increased hold-off. Lastly, easing optical alignment by increasing the active diameter, results in increased power consumption and higher jitter.

The present work chooses to optimize a number of operational parameters which are especially important in biological applications, such as FLIM, and in quantum key distribution. These applications require short dead times for faster acquisitions, as well as large-scale array integration. Moreover, these applications would greatly benefit from a SPAD implementation in a standard, commercial technology, such that data processing can be performed on-chip, thereby reducing the huge amounts of data that would otherwise need to be output for external processing.

The main novelty that makes it possible to achieve faster device performance in a standard technology is a new guard ring. In the next chapters we will describe this

new structure and examine its performance. As in all devices, the reduction in dead time comes at the expense of other performance parameters. We will examine this trade-off and quantify it.

3.2. Device Concept

In most deep-submicron CMOS processes, a shallow-trench isolation (STI) structure is implemented early in the manufacturing process [91]. Traditionally, this isolation scheme has been used to prevent punch-through and latch-up in CMOS circuits. The trench is almost perpendicular to the surface and is filled by silicon dioxide. STI is automatically created in all regions not functionalized by source or drain implants. When a dopant is implanted and is subsequently diffused, lateral diffusion is inhibited by the oxide trench and a nearly planar junction forms (striped region in Figure 3.1). Because the dielectric strength of SiO_2 is 30 times higher than the breakdown field of silicon [85], an STI guard-ring can be made 30 times narrower than a silicon ring. Since it is a dielectric, the shallow trench is also an excellent choice for efficiently electrically isolating the noisy SPAD from the sensing and processing circuitry. Optical isolation must still be achieved by the silicon substrate, but a much narrower ring is sufficient to reduce optical cross-talk to an insignificant level.

A cross-section of the new device is shown in Figure 3.1. A junction is formed between a p+ drain implant and an n-well. Planarization is achieved through the

formation of an STI ring around the junction. Further lateral localization of the absorption region is achieved by metallization (shown in black) above the perimeter of the junction. N-well and p-substrate contacts are made through n^+ and p^+ implants, respectively. Upper layers, including inter-metal dielectric, additional metal layers and passivation have been omitted for clarity. The device is completely compatible with most deep-submicron commercial CMOS technologies.

Layout of the device in a commercial 0.18 μm 1.8V/3.3V CMOS technology was possible without generating any design rule violations. From a device perspective, the size of the active area and the fill factor must be optimized to best suit the end application – a large diode is desired in single-pixel applications for ease of optical alignment, while small pixel pitches are advantageous when a high spatial resolution is needed. The advantage of the new device is in the latitude it allows in the choice of these parameters. On the one hand, the very low defect densities of commercial CMOS processes makes it possible to manufacture large-area diodes with a low probability for a defect. Alternately, because deep-submicron design rules are used, very small pixels with tight pitches can be manufactured, down to an active area of 1 μm per side in a 0.18 μm process.

Small pixels have a number of advantages. Due to the very limited uncertainty of the lateral avalanche initiation position, excellent timing accuracies can be achieved – limited by the quality and accuracy of the electronics. With a decrease in the junction capacitance, down to approximately 50 fF, the total charge flowing during an avalanche decreases as well. As a result, the total charge available for after-pulsing

decreases, and the recharge time drops off, thereby diminishing the dead time. Optical cross-talk, which results from hot-electron luminescence, is also linearly dependent on the total charge flowing and is thus reduced, making it possible to reduce the spacing between pixels. All the benefits outlined above are highly desirable for time-correlated array imagers.

One deficiency of surface SPADs which is exacerbated in deep-submicron devices is their relatively low detection probabilities and their weak infrared spectral response. A SPAD's detection probability is given by

$$\eta = (1-R)\xi P_A F\left[1-\exp\left(-\alpha w_a\right)\right]$$

Equation 3.1

where R is the reflectivity at the surface (which can be significantly reduced with an anti-reflection coating), ξ is the fraction of electron-hole pairs which avoid surface recombination, P_A is the probability that an electron-hole pair will induce an avalanche, F is the effective fill-factor of the device (including compensation by lenslets), α is the wavelength-dependent absorption coefficient of the material, and w_a is the width of the zone from which a photogenerated electron or hole can diffuse to the high-field region of the device. As devices become miniaturized, doping concentrations increase, and as a result the depletion region becomes narrower, thus reducing the detection probability. While this effect is inevitable, the new device's efficient guard ring improves the fill-factor significantly over existing CMOS SPADs for identical active areas, resulting in correspondingly higher detection efficiencies.

With the higher fill-factors, a lenslet array can be added to direct virtually 100% of incoming photons onto the active area.

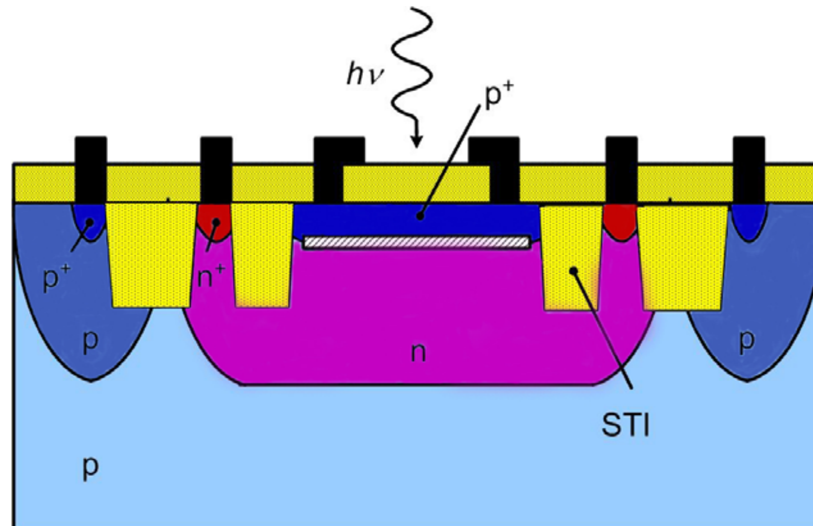


Figure 3.1: A cross-section diagram of the new device.

As device geometries become miniaturized, so do junction depths. This shifts the absorption spectra towards shorter wavelengths. Calculation of the spectral response of a SPAD is possible by taking into account silicon's absorption coefficient, solving McIntyre's equations for avalanche initiation [134], and by considering reflections at layer interfaces, as stipulated above. Figure 3.2 shows the calculated spectral response of the diode, neglecting reflections. While near-infrared response is desirable for fluorescence imaging, the shorter-wavelength response of the SPAD is still useful for many fluorophores which emit at 600-800 nm.

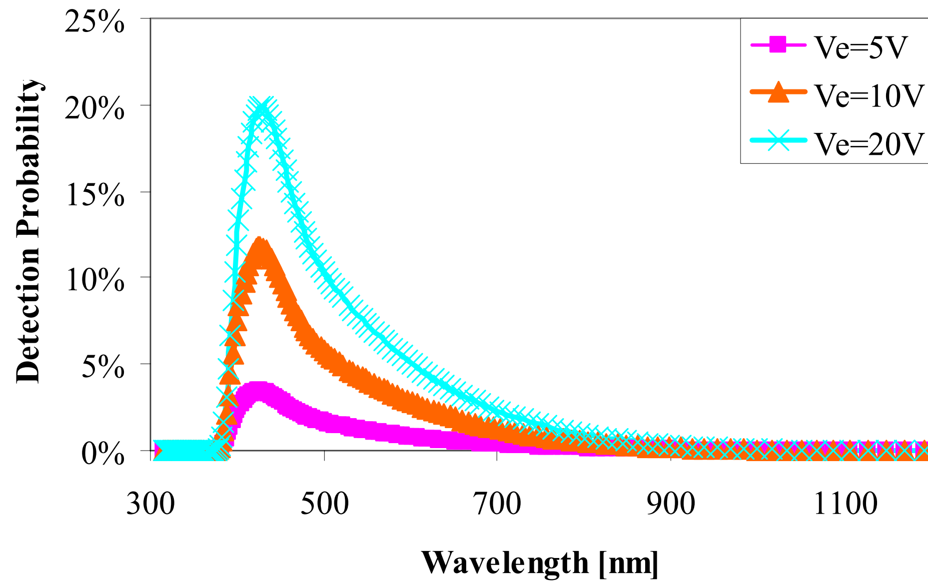


Figure 3.2: Calculated spectral response of the new SPAD, neglecting reflections at the surface and charge diffusion. Increased bias increases the detection probability at the cost of increased dark current.

3.3. Device Modeling

3.3.1. Physical Modeling

In order to validate the design of the new device, the ISE-TCAD device simulator was used to model the SPAD's operation. Process parameters were extracted from the fab's characterization data and design manual. An electrostatic simulation, shown in

Figure 3.3, shows that, as desired, the electric field is confined only to the planar junction region, and that the STI can easily withstand the electric field between the n^+ and p^+ regions. A current density plot, shown in Figure 3.4, shows that although a curvature exists in the current flow around the STI – which will manifest itself as a

space-charge resistance affecting the RC time constant of the SPAD – current is quite uniformly distributed across the N-well and, more importantly, across the junction region. Thus we should not expect local "hot" spots in the junction, which would be a reliability concern.

The electric field confinement of the STI-bounded SPAD was compared to that of the traditional diffused-ring SPAD. The latter was modeled in ISE as a p^+ implant surrounded by a P-well. Electrical connection to the N-well was achieved through a deep-N-well, which is commonly available in analog and RFCMOS processes. Simulation results, shown in Figure 3.5, indicate that, although the highest electric field is concentrated in the planar junction region, additional high-field regions exist in the diffuse p-well to n-well junction. These may result in premature breakdown in the curved regions and in additional junction capacitance.

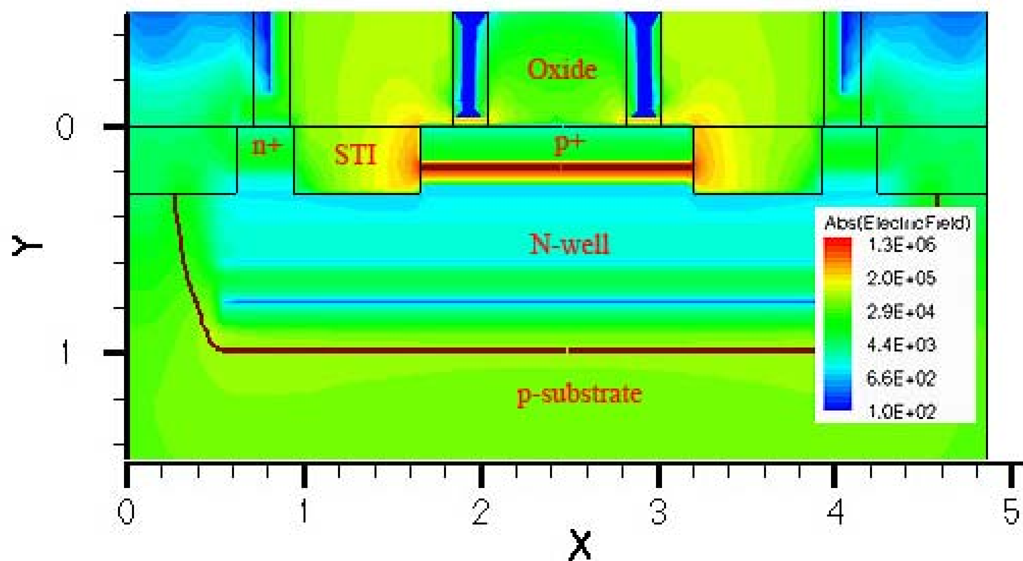


Figure 3.3: ISE-TCAD simulation of electric field distribution in the SPAD when biased above breakdown. As desired, the high field region is confined to the planar junction region. The electric field is given in V/cm and coordinates are in microns.

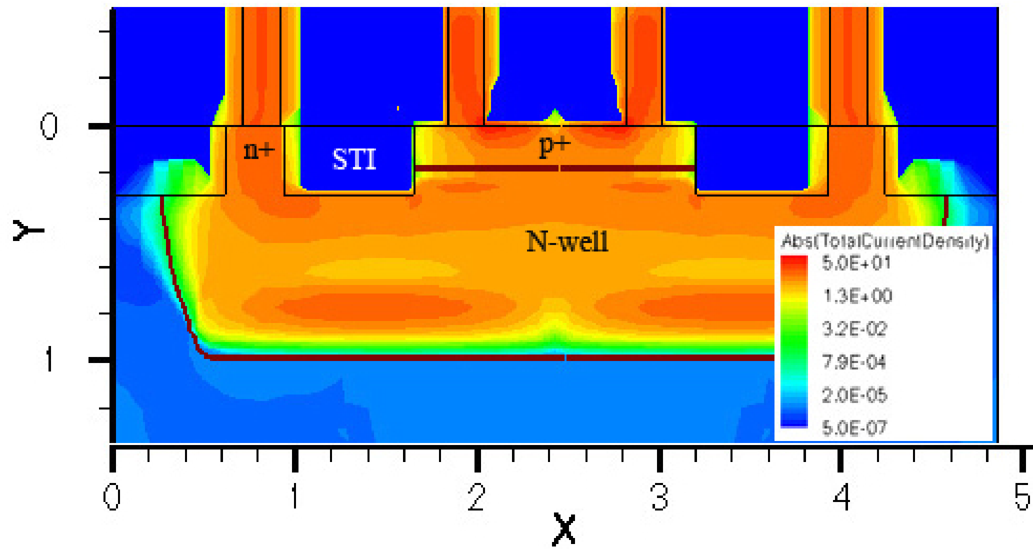


Figure 3.4: ISE-TCAD simulation of current in the SPAD when biased above breakdown. Current curvature around the STI region validates proper operation. The electric field is given in V/cm and coordinates are in microns.

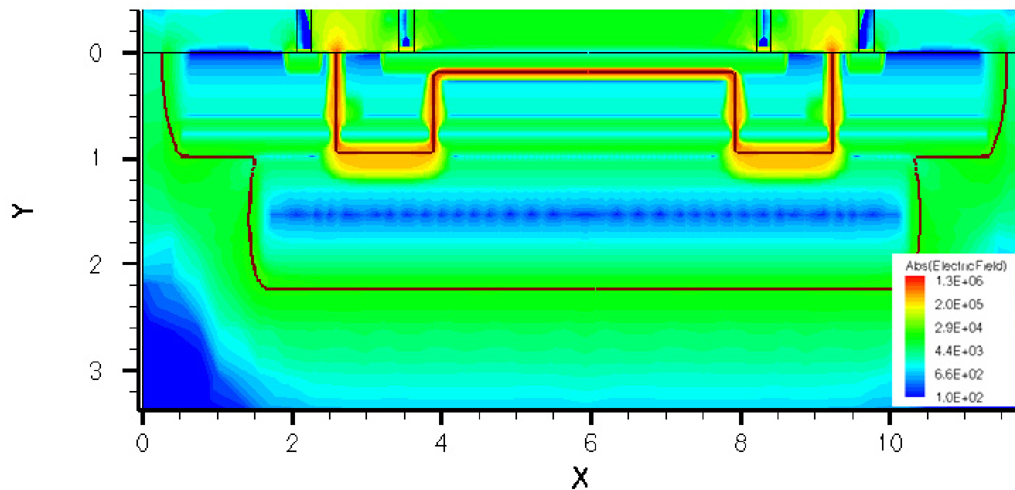


Figure 3.5: ISE-TCAD simulation of electric field distribution in diffused-ring SPAD

3.3.2. Electrical Modeling

The electrical model for simulating a SPAD has been described above and is shown in Figure 2.16. A voltage source V_{HIGH} pre-charges the Nwell/p$^+$ junction, which is enclosed in a rectangle in the drawing. The junction itself is comprised of a resistance R_d and a capacitance C_d. The diode's resistance, dominated by the space-charge resistance, is given by Equation 2.46. Measurements on a 15 μm-diameter triple-well SPAD with a depletion region width of approximately $W = 0.5$ μm, found this total resistance to be approximately 1 kΩ [70]. The value of the space-charge resistance for this structure is $R_s = 1/(40\text{x}10^{-6} \text{ x } 15^2/0.5^2) \approx 28\ \Omega$, i.e., the total resistance is dominated by the series resistance rather than the space-charge resistance.

Because the neutral region in the STI-bounded structure was very thin, the dominant component of the SPAD's resistance must be the ohmic resistance of the N-well. Even for much smaller SPAD pixels, e.g., having a 4 μm^2 active area, where the space-charge resistance becomes more dominant, R_s= 390 Ω is much smaller than the ohmic resistance. Since the path from the electrode to the junction in the STI-bounded SPAD is significantly shorter than that of the triple-well SPAD, we conservatively assume a resistance of 1.5 kΩ.

The diode's capacitance is simply the diode's depletion capacitance. It is given by the fab [135]. The junction capacitance results from the depletion region at the p-n interface. This capacitance varies in inverse proportion to the width of the depletion

region, which is itself a function of the applied voltage V_A. The junction capacitance scales with the area and perimeter of the device:

$$C = C_a \times Area + C_p \times Perimeter$$

Equation 3.2

The voltage (reverse bias is negative voltage) dependence of the area capacitance, C_a, and the perimeter capacitance, C_p, is given by ($i = a$ or p)

$$C = \frac{C_{i0}}{\left(1 - \frac{V_A}{V_{bi}}\right)^{m_i}}$$

Equation 3.3

where V_A is the applied voltage and V_{bi} is the applied voltage. The parameters used in these equations for the capacitance are given by the foundry.

For a 7 µm x 7 µm diode reverse-biased at $V_A = -11$ V, the capacitance is 22 fF and for a 2 µm x 2 µm diode it is 2.34 fF. In both cases, and especially in the smaller diode, the parasitic capacitance, C_p, which must be charged and discharged together with the depletion region's capacitance, must be accounted for. With special attention to routing, this parasitic capacitance is dominated by the gate capacitance of the sensing stage, with a capacitance of:

$$C_{ox} = \varepsilon_0 A / t_{ox} = 1.1 \text{ fF}/\mu\text{m}^2$$

Equation 3.4

The value of the quenching resistor can be determined by considering the requirements for terminating the avalanche process in an abrupt manner. Haitz determined that when the current through a diode falls below 60 μA, the probability of a self-sustaining avalanche drops significantly. A commonly used rule-of-thumb places this final current at 100 μA [98]. In our design, we designed for I_f= 20 μA and V_e = 2.5 V, resulting in R_q = 120 kΩ. This resistor can either be implemented using a resistive element which is part of the process offering, or by an active resistor in the form of a MOSFET, as long as the voltage across any two of its nodes does not exceed VDD = 3.3V. Using an active resistor has three advantages: a) it is more compact in area and b) its temperature coefficient is lower and c) its resistance value can be externally controlled by adjusting its gate voltage.

Traditionally, the MOSFET resistance, R_q is sized such that fast and complete quenching is achieved but cannot be made arbitrarily large so as to unnecessarily lengthen the recharge "dead" time, set by the RC time constant. In the model, the capacitance consisting of the diode's capacitance, C_d, the output stage's input capacitance and any parasitic capacitance, $C_{parasitics}$, is first charged to V_{High}, in excess of the breakdown voltage of the diode, V_B. The avalanche is simulated by closing a switch, discharging the capacitance to V_B. As soon as the avalanche is quenched the switch re-opens, re-charging the SPAD.

3.4. Peripheral Circuits Design and Simulation

The design of a CMOS SPAD chip involves the simulation of the ultrafast avalanche and its interaction with the peripheral quenching, recharging and sensing circuitry. Parasitic effects must be carefully considered because they can dominate the SPAD's behavior. Finally, special care must be paid to the layout of the SPAD because it operates beyond the allowed performance envelope of the process, and because it cannot be verified using the fab-supplied validation tools. Lastly, one must consider the external test environment and its effect on the observed signals. In order to maintain the fast slopes and repetition rates, even when confronted by the large output capacitance, proper buffering must be designed on the chip.

Circuit design, simulation and layout of the test chips and support circuitry were performed using the Cadence environment, with design tools supplied by IBM.

3.4.1. Output Buffers

The electrical simulation model for a SPAD has been described in section 2.7 and is shown in Figure 2.16. The simulated waveforms of a SPAD operating cycle are shown in Figure 3.6. Initially, the junction is charged with an overbias of *VDD* on the N-well node and the switch. An avalanche is simulated by closing the switch, quickly discharging the junction capacitance. The avalanche is quenched through R_q resulting

in the current pulse shown in green. Upon completion of the quenching (defined as reaching a current of 100 μA), the switch is re-opened, allowing for junction recharge. The output is read from the N-well node, resulting in the characteristic waveform for passively-quenched SPADs, comprised of a fast leasing edge and a slow trailing edge. Since the time-of-arrival information is contained in the leading edge of the N-well signal, special care has been paid to maintaining its fast slope, thus reducing its jitter.

We designed two output buffer configurations. The aim of the first one was providing maximal visibility into the waveform of the N-well node of the SPAD, without loading excessive capacitance on it. A source-follower output was the simplest choice, providing an output, which is proportional to the N-well voltage (Figure 3.7).

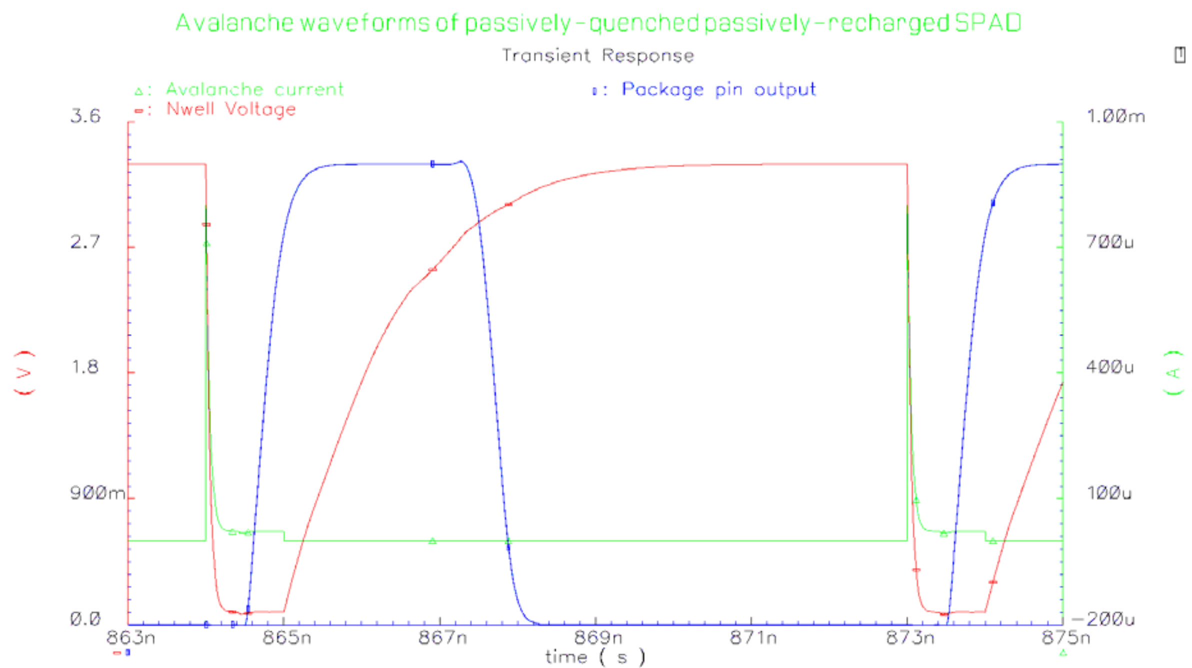
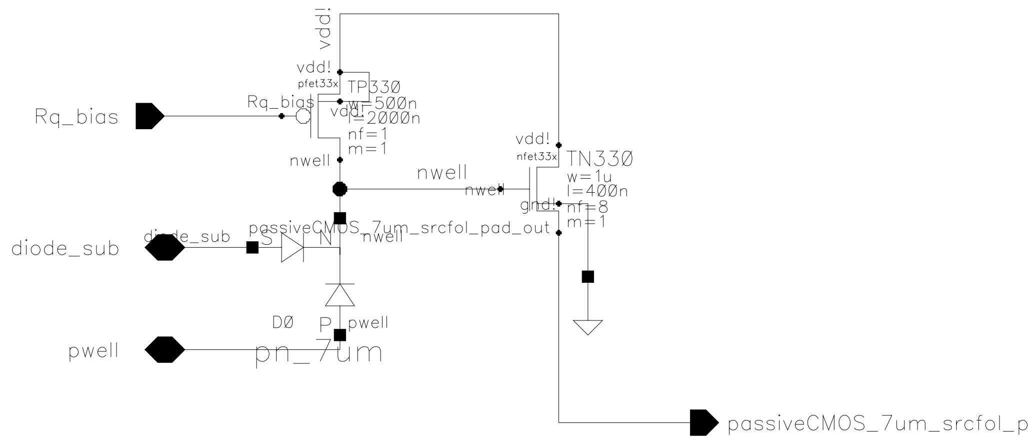
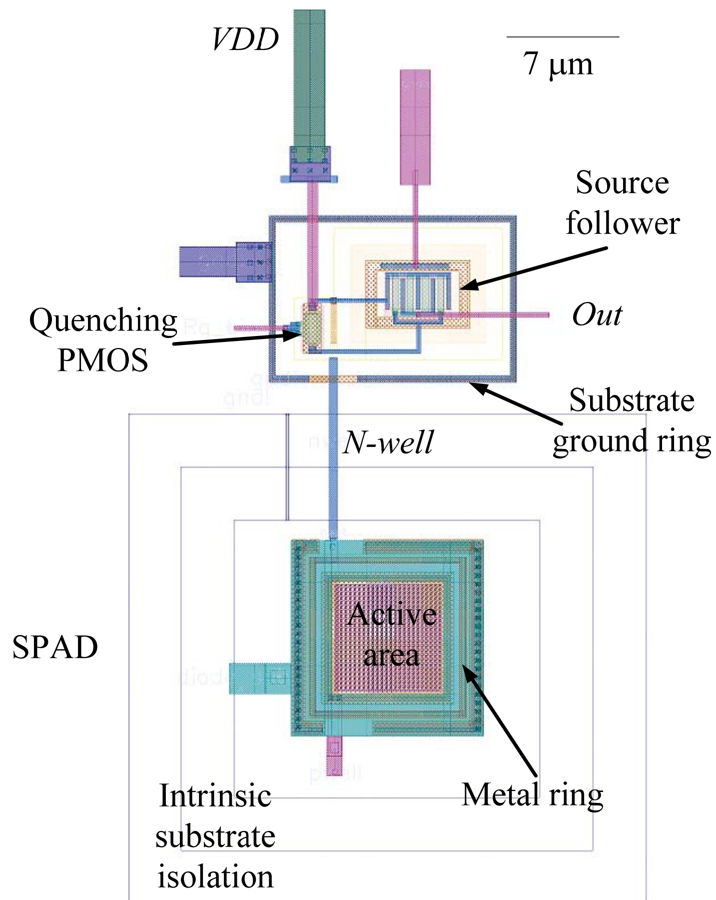


Figure 3.6: Nwell and Output waveforms for passively-quenched SPAD with source-follower output stage.

The drawbacks of the source-follower scheme lies in the small signal amplitude resulting from the current-mode output. Moreover, whereas in the simplest implementation which was used in the present design, threshold detection is performed off-chip, it is, in fact, advantageous from a signal-to-noise perspective, to perform the threshold detection on-chip. This can be done using a comparator, or, more simply, using an inverter with a carefully-chosen triggering voltage. Cova showed that in order to minimize jitter, the comparator threshold should be as low as possible (i.e., at the onset of the avalanche) [127]. Timing information can be maintained by designing the inverter asymmetrically, with a fast slew rate corresponding to the leading edge of the avalanche pulse. A cascade of such inverters can generate an output which can drive the picofarad board capacitance while minimizing jitter [6]. Such an inverter chain, with its corresponding simulated waveforms is shown in Figure 3.8.



(a)



(b)

Figure 3.7: (a) Schematics and (b) layout of SPAD with source-follower output.

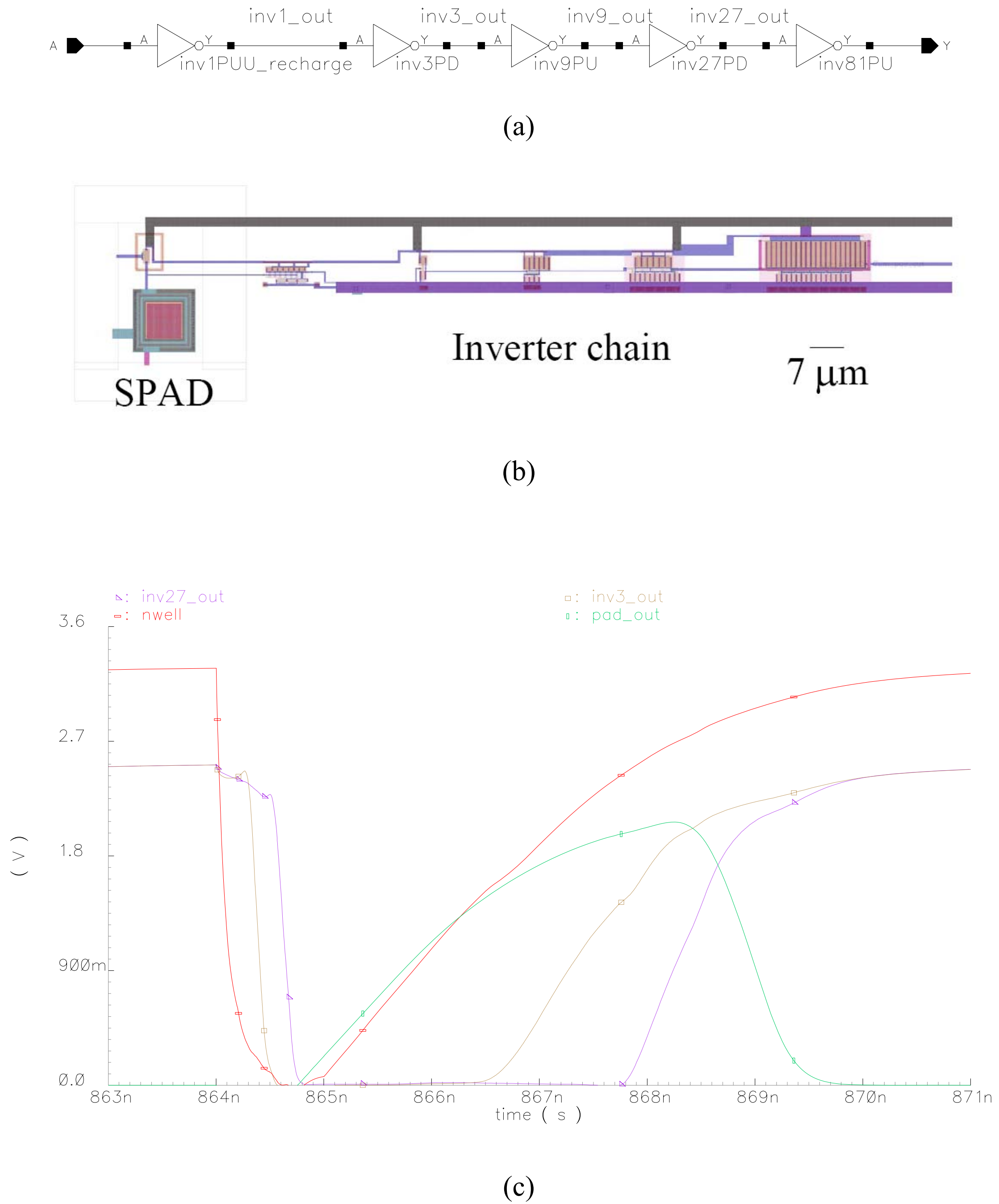


Figure 3.8: (a) Schematic, (b) layout and (c) simulation output waveforms for SPAD with inverter-chain output stage

As can be seen from Figure 3.8, the voltage across the diode varies over the whole recharge time of 4 ns. This is undesirable, because the avalanche probability is strongly dependent on the excess voltage [7] so that with the present scheme we have a non-uniform detection probability during this relatively long interval (if operating at high photon flux). Moreover, because the after-pulsing probability decays exponentially with time, but can be self-sustaining, it is often useful to delay the time between an avalanche and the full recharge of the junction. This is in contrast with the situation in larger SPADs where the aim is to reduce the recharge time, which is determined by the RC time constant of the SPAD and the junction's parasitic capacitance.

3.4.2. Ultrafast and Compact Active Recharge Circuit

3.4.2.1. Motivation

Detection probability is one of the most important figures of merit of single photon avalanche diodes. It is determined by the ability of the SPAD's absorbing layer to absorb photons in the relevant wavelengths and by the capability of the device to separate the photogenerated electron-hole pair and multiply it to a detectable avalanche. In some devices, such as the STI-bounded silicon SPAD, the absorption and multiplication regions coincide in the depletion region of a pn junction. The photon absorption probability is [85]:

$$P_{abs}(\lambda) = \exp\left(-\alpha_{Si}(\lambda)\cdot\left(x_j + w_d\right)\right) - \exp\left(-\alpha_{Si}(\lambda)\cdot x_j\right)$$

Equation 3.5

where $\alpha_{Si}(\lambda)$ is the absorption coefficient in silicon for wavelength λ, w_d is the depletion region's width, and x_j is the junction depth.

The depletion region's width depends on the doping profile of the lighter-doped region and on the applied bias, as given by Equation 2.10. As smaller geometries are used, the grading coefficient increases, thereby reducing w_d and thereby the absorption probability. The only way to offset this trend is by increasing the applied voltage across the junction. Increasing this voltage also has the benefit of increasing the probability that once an electron-hole pair has been generated, the opposite charges will be swept and multiplied by impact ionizations, resulting in an avalanche [134].

Unfortunately, the applied voltage cannot be increased indefinitely, because SPAD noise increases with applied field. In these devices, noise stems from thermally generated carriers, direct band-to-band tunneling and from after-pulsing [63]. Thermally-generated carrier noise, exhibits a Poissonian distribution and increases with applied field, due to a higher avalanche-initiation probability and because of increased carrier emission via the Poole-Frenkel effect [63]. Direct band-to-band tunneling, also with a Poissonian distribution, becomes a dominant noise source when high fields are applied across narrow junctions, as is the case in deep-submicron

devices. Finally, for fast SPADs, the total number of avalanches resulting from after-pulsing also increases with higher fields for the same reason outlined above.

After-pulsing follows a multi-exponential time distribution, depending on the lifetimes of traps involved [121]. Because the released carriers may trigger a false avalanche that is not the result of an impinging photon, it is imperative to reduce the probability that such events occur. This can be done by ensuring that the device is biased below its breakdown for a sufficiently long time following an avalanche, such that only a negligible residual trap population remains filled for the next detection window. This results in a dead time during which photons cannot be detected, and is the minimal time between consecutive avalanches. Because trap lifetime is inversely proportional to temperature, cooling the device only exacerbates this noise. Dead time limits the fastest phenomena which can be measured in fluorescence correlation spectroscopy [136], determines the total acquisition time for 3D imaging applications using SPADs [73] and sets the saturation level in photon counting applications [70].

In order to maximize timing precision in SPADs, their junction area must be minimized. This reduces the uncertainty in the position of the avalanche generation within the junction, which in turns reduces the uncertainty in delays between the photon arrivals and avalanche detections [70]. An added benefit of smaller junctions is their reduced noise: trap-assisted and direct tunneling currents are linearly proportional to the junction area, while after-pulsing depends on junction capacitance [73].

The small capacitance, down to 30 fF in our SPADs also results in a reduced dead time. A 7 μm-diameter SPAD using the STI guard-ring exhibits a 3 ns dead time, the shortest reported to date [103]. However, this short dead time does not leave sufficient time for some traps to be released, resulting in an unacceptable after-pulsing rate at room temperature. When heated, the dark rate drops significantly, indicating that after-pulsing is responsible. However, at higher temperatures trap-assisted tunneling becomes dominant.

We set about to design an active-recharge circuit to alleviate this problem, making it possible to operate at higher voltages, thus achieving higher detection efficiencies, while reducing the dark current. Unlike conventional active-recharge circuits, which are designed either to reduce the dead time of the device [108] or to improve its timing precision [106], our circuit aims to reduce after-pulses and increase the attainable detection efficiency. It should preferably not significantly increase the device dead time, so that the benefits of the small geometry are not relinquished, so it must perform its processing and feedback within approximately 3 ns – the dead time of the passive device. This can only be achieved by using fast transistors in close proximity to the diode. Moreover, such a circuit should not impinge on, and should preferably improve upon the quenching behavior of the SPAD. Lastly, a compact and noise-free design is desirable so that these structures can be incorporated into multi-pixel arrays.

3.4.2.2. Design and Simulation

Cova et al's groundbreaking work on SPAD active quenching and recharge was mainly motivated by the need to counter the effects of the large parasitic capacitance of an externally passively-quenched SPAD [137]. In passive quenching, a voltage is applied across the diode through a resistor. As an avalanche forms, the current through the resistor rises and a voltage builds across it, resulting in a lower bias across the diode, thus quickly quenching the avalanche. Immediately after completion of the quenching, the diode capacitance, in addition to any parasitics, are recharged through the same resistor, with a time constant $R_q(C_d+C_p)$, where R_q is the quenching resistance, C_d is the depletion region's capacitance and C_p is any parasitic capacitance on the diode node. If the quenching cannot be achieved on the same die as the SPAD, a large parasitic capacitance, on the order of a few picofarads results. Consequently, the dead time increases, and the photon-counting rate is reduced [106]. Moreover, because of the gradual charging of the diode, the overbias varies within the exposure time, resulting in varying detection efficiencies.

An active recharging scheme can improve these issues. Its effect is shown in Figure 3.9. With passive quenching, the SPAD can after-pulse almost immediately after it has been discharged, yet it only achieves its full overbias (and corresponding optimal detection efficiency) after a long delay. A desired active recharge scheme will silence the detector for the duration of the dead time, and will instantaneously recharge it. This will allow trapped charges to be released without inducing after-

pulses, and will ensure a binary ON or OFF operation. The effect should reduce the primary after-pulses, i.e., those directly resulting from the initial pulse, as well as prevent the formation of secondary after-pulses generated by the primary ones.

A number of active quenching and recharge schemes have been implemented [15, 16]. They use analog circuitry to sense the onset of an avalanche. Using a monostable or a multivibrator they then temporarily open a low-impedance recharging path, resetting the diode, subsequently disconnecting this low impedance to allow for complete avalanche quenching. The use of resistors and other sensitive analog circuitry may introduce undesired effects, such as temperature dependence of the dead time and noise. Furthermore, it is desired to make the active quenching unit compact and low-power, to allow for multi-pixel arrays, and to construct it with digital blocks compatible with generic semiconductor processes.

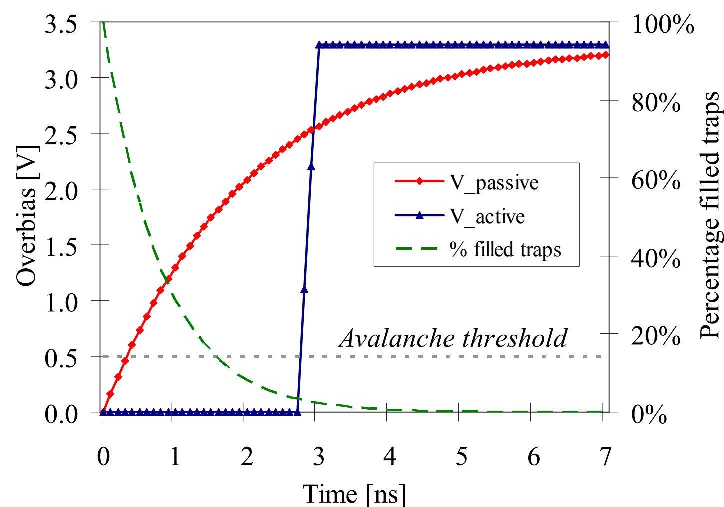


Figure 3.9: Passive and active recharge waveforms. With active quenching, breakdown can occur only after most of the traps have been released.

We designed a compact self-timed actively-recharged circuit which can be incorporated into an array (Figure 3.10). In the active-quenched pixel, the quenching resistor found in the passive pixel, which also serves for recharging the diode, is replaced by two active resistors. A PMOS quenching transistor, *M_q*, with a high "on" resistance, R_{quench} = 1.2 MΩ, ensures fast avalanche quenching with minimal leakage current during non-quenching times. A second, smaller transistor, *M_r*, is used for an ultra-fast recharge, with $R_{recharge}$ = 24 kΩ.

In order to ensure a uniform detection efficiency throughout the detection cycle and in order to minimize after-pulsing, the diode must be kept below its breakdown voltage from the time the avalanche has been quenched until most of the trapped charges are released. It should subsequently be quickly recharged, e.g., through a low resistance. However, this low impedance must be disconnected immediately upon completion of the recharge because at that instant the device is ready to fire-off upon absorption of a photon, but has insufficient resistance to fully quench an avalanche.

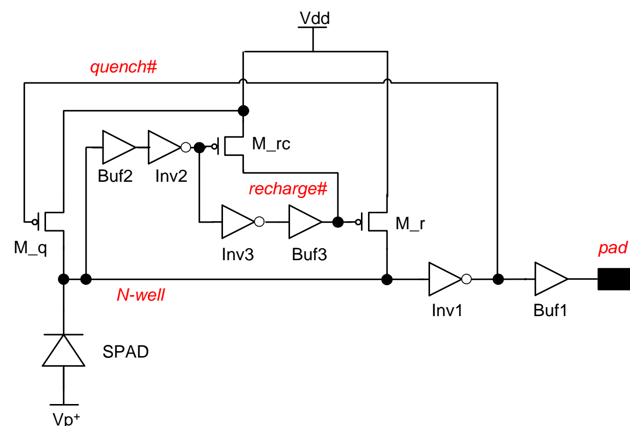


Figure 3.10: Schematic diagram of the active-recharge circuit.

At the beginning of a sensing phase, M_q is in the linear region (*quench#* is low) and M_r is cut-off (*recharge#* is high), with $R_{M_r,OFF} > R_{M_q,ON}$. When a photon arrives and an avalanche builds up, the junction capacitance quickly discharges the *N-well* node and the avalanche is quenched due to an IR drop across M_q. As the N-well voltage drops, M_q moves to saturation and is quickly cut-off by the sensing inverter, Inv1 (*quench#* goes low). This reduces the leakage current through it and in essence freezes the voltage across the junction, so that traps can be emptied without inducing an avalanche. After a longer delay, which is set by Buf3 (based on the expected trap lifetime), M_r is switched to its saturation region through M_rc (*recharge#* goes low), and the diode quickly recharges through the small recharging resistance. When recharging is almost complete, M_r moves to the linear region and the quenching transistor M_q is turned on (*quench#* low). As discussed above, the recharging time must be kept short, and as soon as the excess voltage is attained, the small resistance must be quickly disconnected. This is achieved by Buf2 and Inv2, which turn M_rc on, resulting in the cut-off of M_r.

The circuit simulation shows two discrete SPAD bias levels, corresponding to the desired binary sensitivity states. A dead time of 3 ns was targeted for a significant reduction in dead counts compared with a passively-quenched device having similar detection efficiencies.

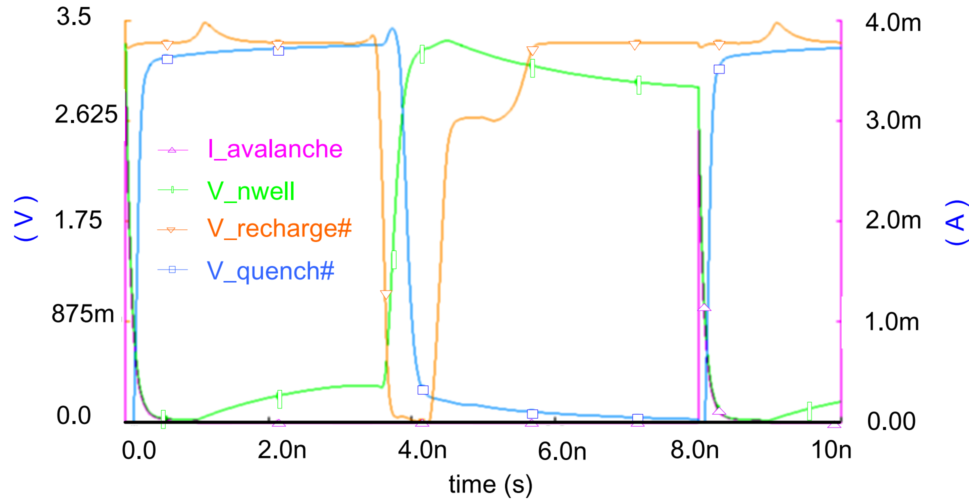


Figure 3.11: Cadence simulation results of an active-recharged SPAD cycle.

3.4.2.3. Delay-afterpulsing optimization

In order to optimize the operation of the STI-bounded SPAD, it is essential to improve upon the noise performance of the SPAD device. On the one hand, recharging can be delayed in order to release trapped charges. However, in order to reap the full benefits resulting from the low-capacitance guard ring, it is desirable to minimize this recharge time in as much as possible. In order to achieve this, we designed an active-recharge circuit similar to the one described in section 3.4.2, but with an externally-controllable delay. The simulation schematic of the new circuit is shown in Figure 3.13 and its results in Figure 3.14.

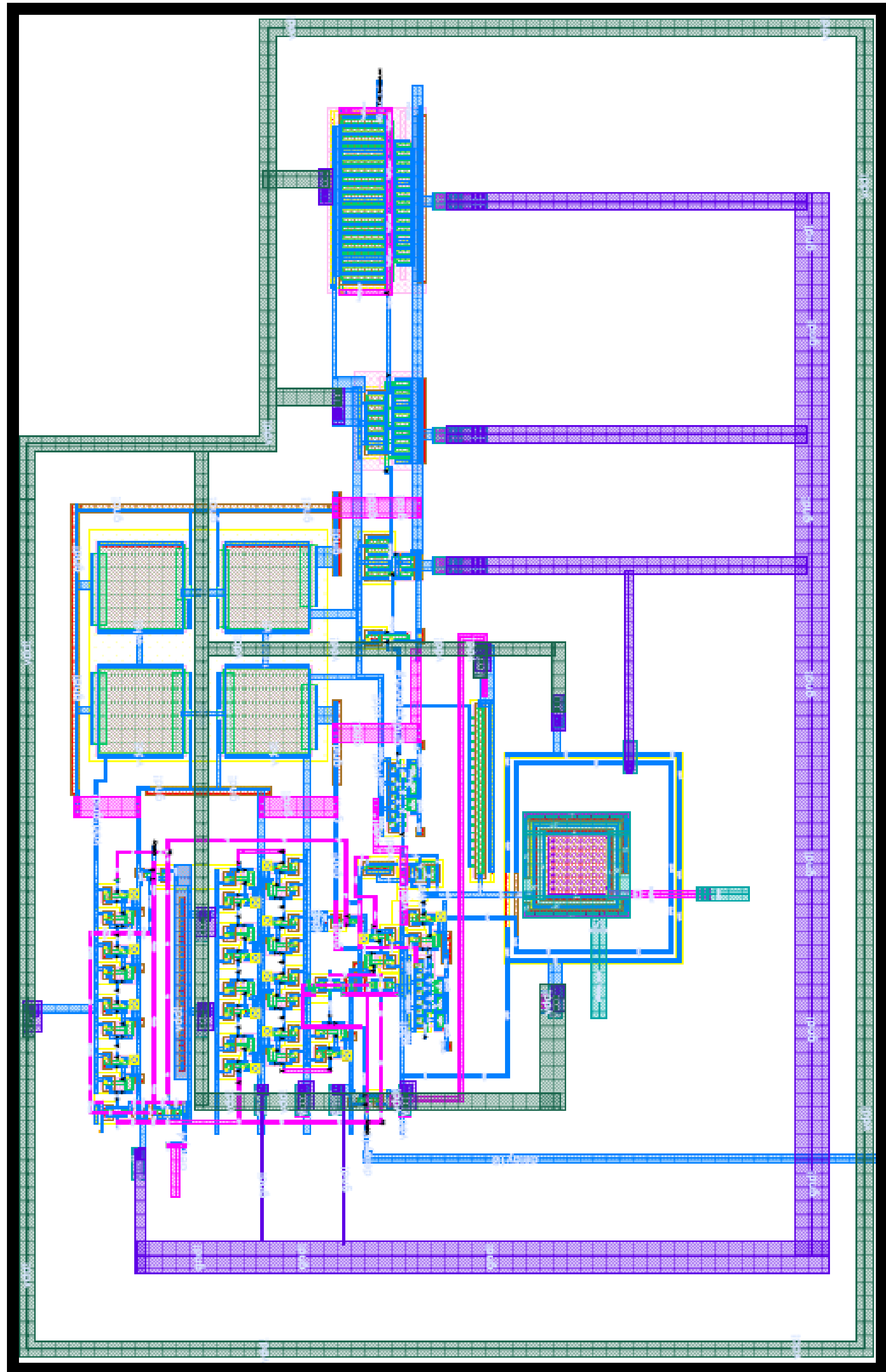


Figure 3.12: Layout of actively-recharged SPAD

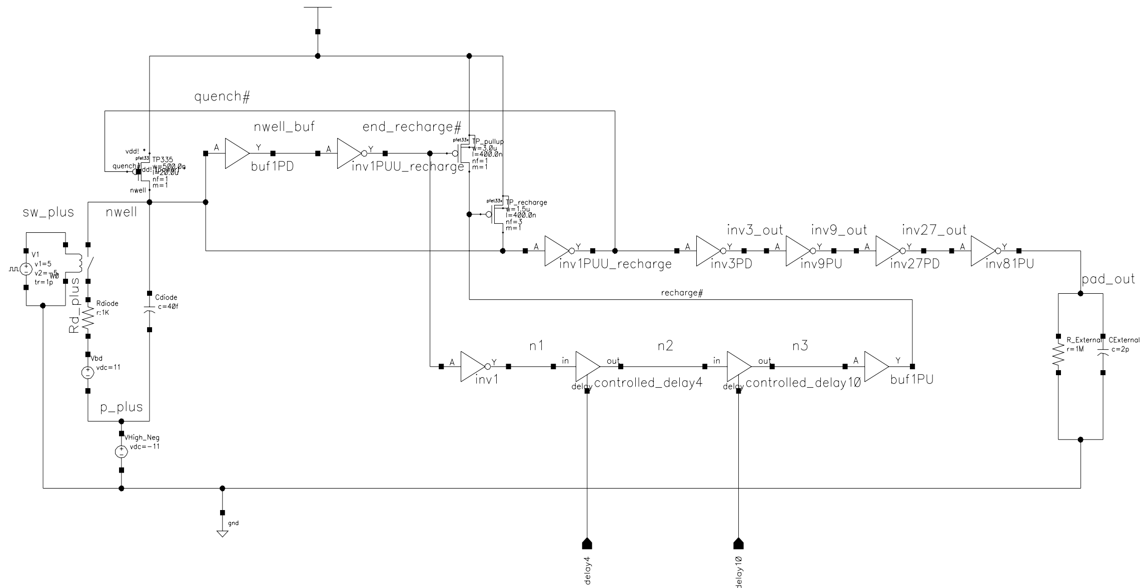


Figure 3.13: Simulation schematic of active-recharge circuit with externally-controlled delay for afterpulse optimization.

Externally-controlled buffers were added to the scheme shown in Figure 3.10 such that each may either be bypassed (via a pass-gate) or add a delay to the *recharge#* signal. Optimization of device operation can then be performed in the lab, based on the noise performance in each of the delay setting.

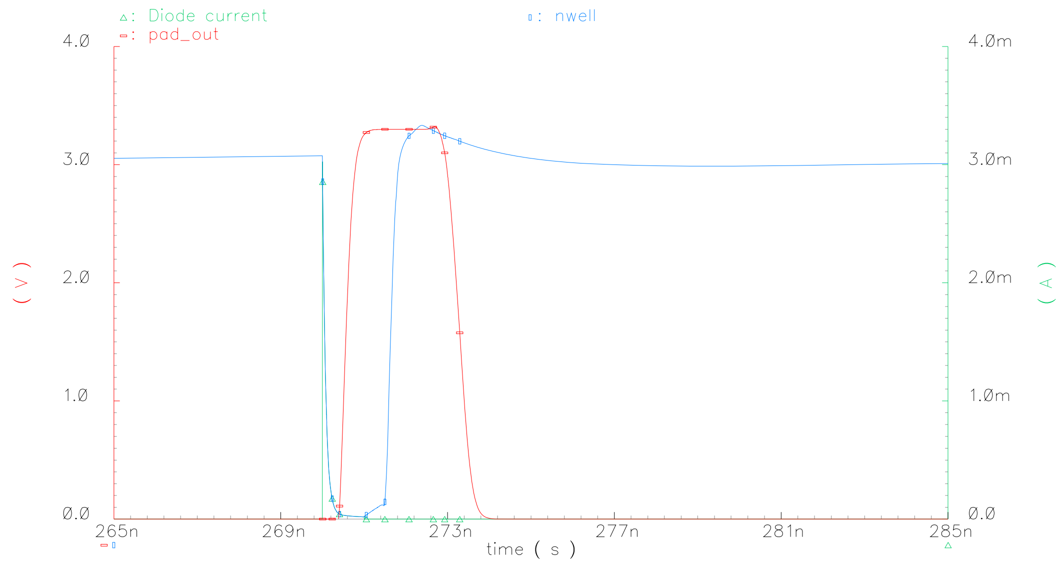
3.5. Dual-Color Single-Photon Detection

3.5.1. Device Concept

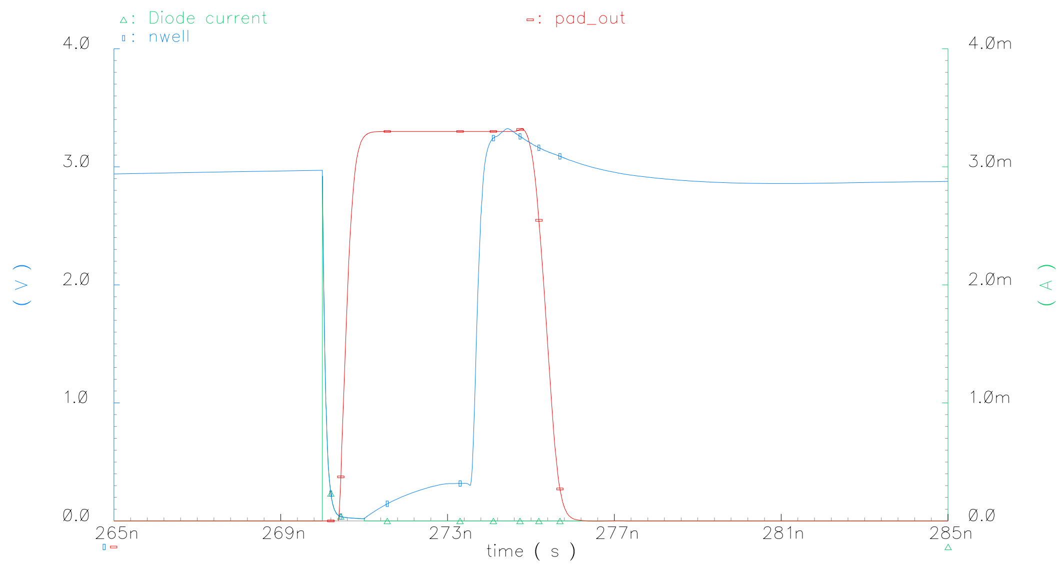
An interesting and useful extension of the STI-bounded SPAD described above is the dual-junction SPAD. Such a device can greatly simplify dual-color fluorescence setups [138], provide greater inter-channel timing synchronization, and can find uses in many other applications. We note from Figure 3.1 that each SPAD pixel actually

consists of two junctions – the shallow p^+/N-well junction, which was described above, and a non-planar N-well/p-substrate junction. Proper biasing of the three regions can result in one or both of the junctions being biased above breakdown. Since both junctions share the N-well cathode. Due to the lower doping of the N-well, compared with that of the p^+ implant, the breakdown voltage of the deep junction is very high, on the order of 100V. However, this junction is non-planar, and exhibits early breakdown, starting at 10.5V reverse bias, close to the 11V breakdown voltage of the planar junction.

While junction planarity has traditionally been viewed as essential for SPAD functionality, we propose to utilize specifically the curved regions of the junction for detection. From Figure 2.20, we see that blue and ultraviolet photons impinging upon the pixel, will be absorbed in the shallow junction with a significantly higher probability than in the deep junction. Red and near-infrared photons, on the other hand, are much more likely to be absorbed in the deep junction. Appropriate optics may split the wavelengths to the different absorption regions.

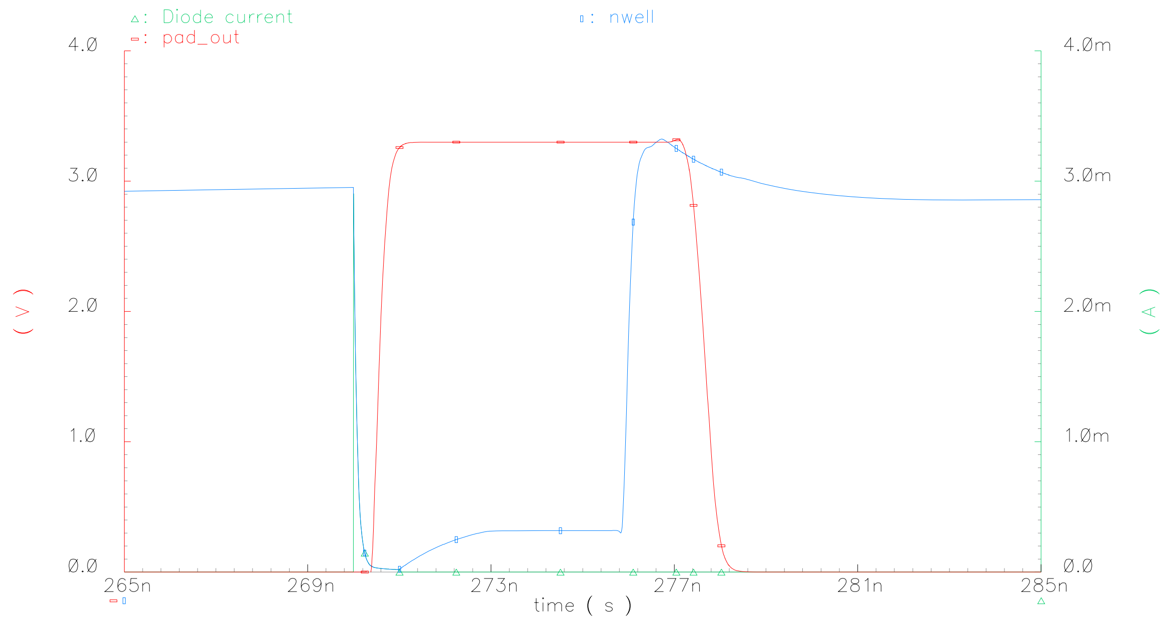


(a)

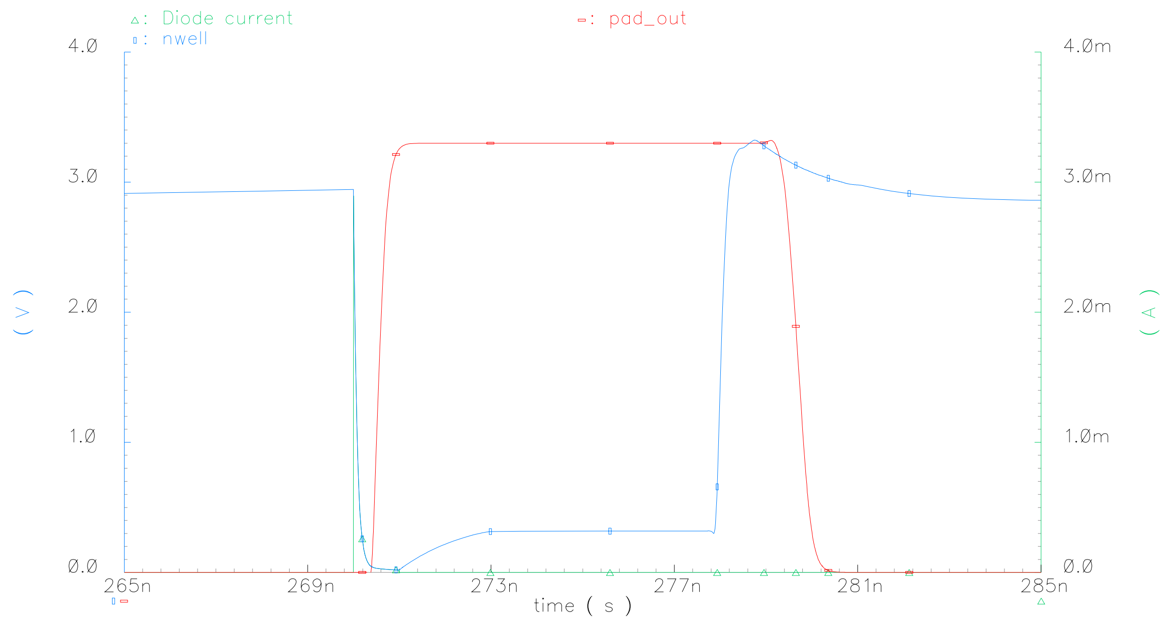


(b)

Figure 3.14: Simulation waveforms of active-recharge with externally-controlled delay. Delays are (a) 3 ns and (b) 7 ns.



(c)



(d)

Figure 3.14 (cont.): Simulation waveforms of active-recharge with externally-controlled delay. Delays are (c) 10 ns and (d) 12 ns.

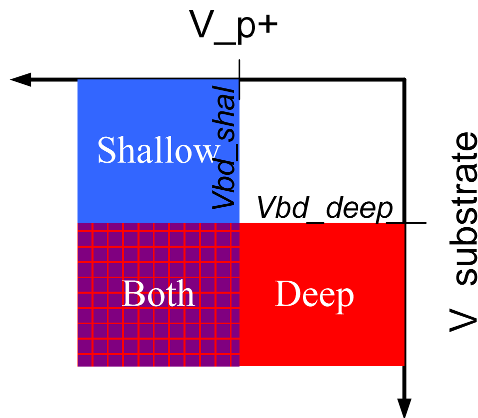


Figure 3.15: Operating principle of dual-junction SPAD: When the substrate is biased beyond *Vbd_deep* – the breakdown voltage of the deep junction – red photons can be selectively detected. The same holds for when the shallow junction is biased beyond *Vbd_shal*. When both junctions are biased beyond breakdown, both junctions detect simultaneously.

A number of challenges and questions must be addressed before such a device can be manufactured:

- What will be the effects of using a curved junction on device reliability?
- Can both junctions be operated simultaneously? In other words, can the large number of charge carriers flowing in an avalanche be swept away from the non-avalanching junction? What will be the effect of photons emitted during an avalanche in one junction (via hot-carrier recombination) on the other junction?
- Since the deep junction utilizes the substrate as the anode, how will adjacent pixels be isolated? Moreover, since the substrate must be biased for proper diode operation, how will the peripheral CMOS circuitry operate?
- How will the avalanche signals from both junctions be read out uniquely?

- Since considerable cross-talk is expected due to inter-channel absorptions, can meaningful information still be extracted from the device?

Preliminary answers to some of these questions are presented in sections 3.5.2 and 4.11.

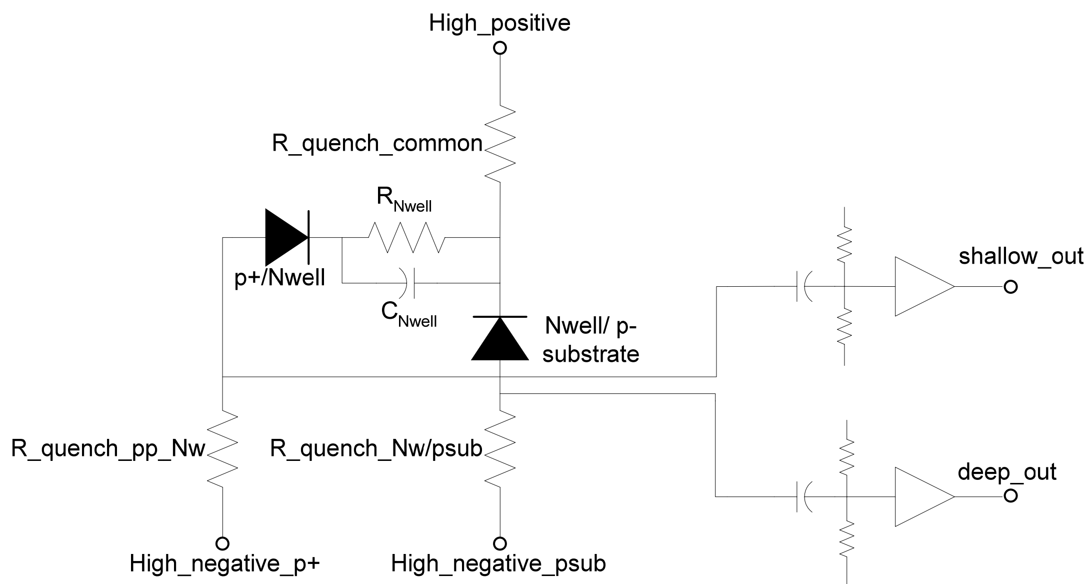
3.5.2. Peripheral Circuitry

Quenching, recharge and readout in the dual-color SPAD present unique challenges. Since the N-well node is common to both diodes, readout must be carried out from the other nodes – p^+ and substrate. These are biased at high negative voltages, so quenching cannot be done with a PMOS transistor, as is done in a traditional SPAD. The high voltage spike must nevertheless be eventually read out and processed through standard CMOS so it must be DC-level-shifted with minimal effect on jitter. Lastly, because dual-color single photon detection looks at cross-correlations between the channels, often in response to a single excitation, it is necessary to match the jitter and delays of both channels.

A conceptual schematic of a circuit we designed for this purpose is shown in Figure 3.16 (a). The N-well node is in fact separated to two nodes coupled by the resistance and capacitance of the N-well. Quenching and recharge is distributed between a common N-well resistor, *R_quench_common*, and diode-specific resistors, *R_quenc_pp_NW* and *R_quench NW/Psub*. A capacitor acts as a high-pass filter, passing the rising edge of the avalanche pulse with high timing precision. A resistor

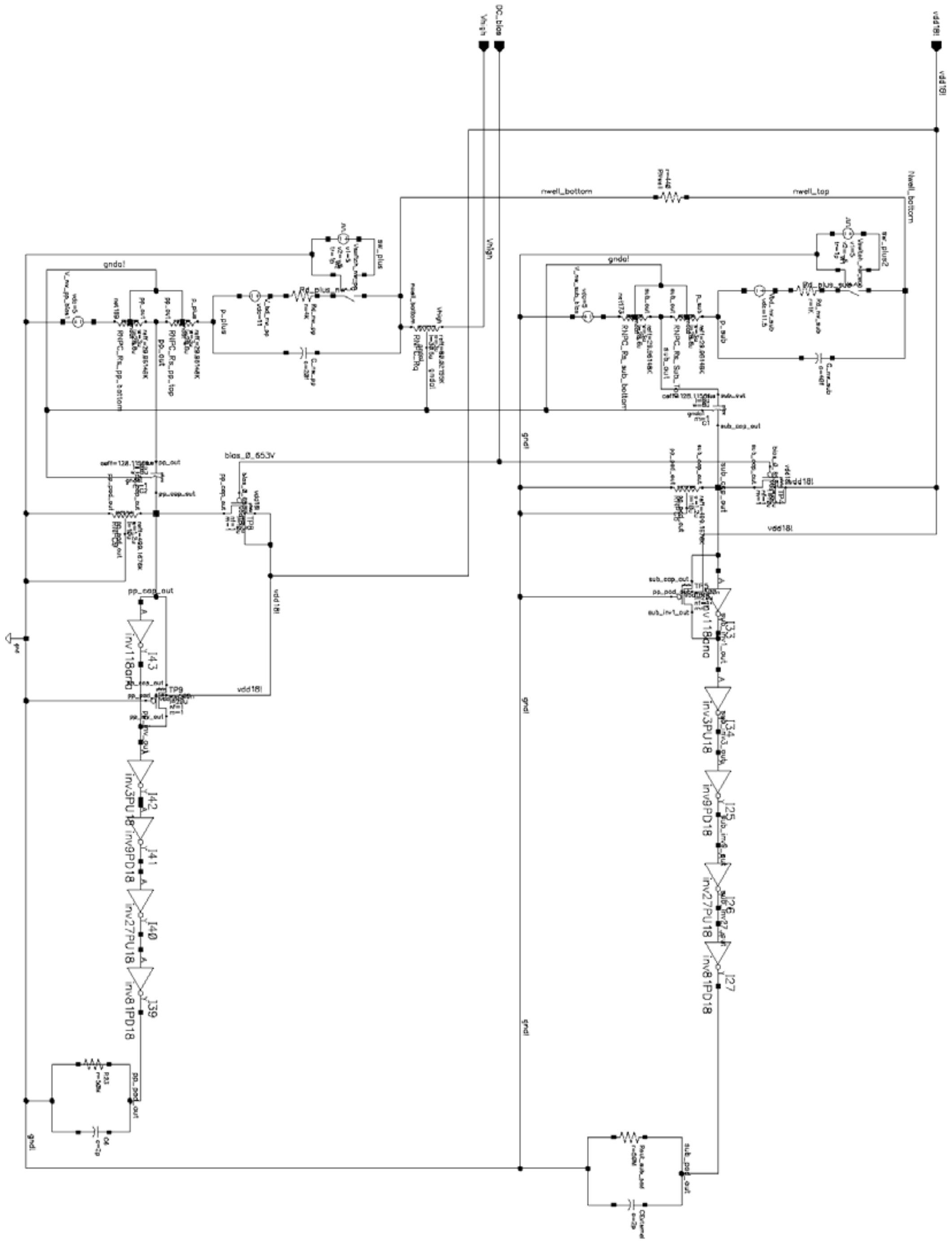
network fixes the DC level of the derived signal, so it can be sensed by a CMOS sense amplifier and processed outside the chip.

One of the main problems with extracting unique signals from each junction is inter-junction cross-talk, modeled by the N-well's resistance and capacitance. Careful tuning of the junction biases and other circuit parameters resulted in robust outputs, shown in Figure 3.17, and referring to the simulation schematic of Figure 3.16 (b). The cross-coupling in the case of three scenarios is shown, but in all cases, the filtering and sampling circuitry removes these artifacts, outputting correct binary signals.



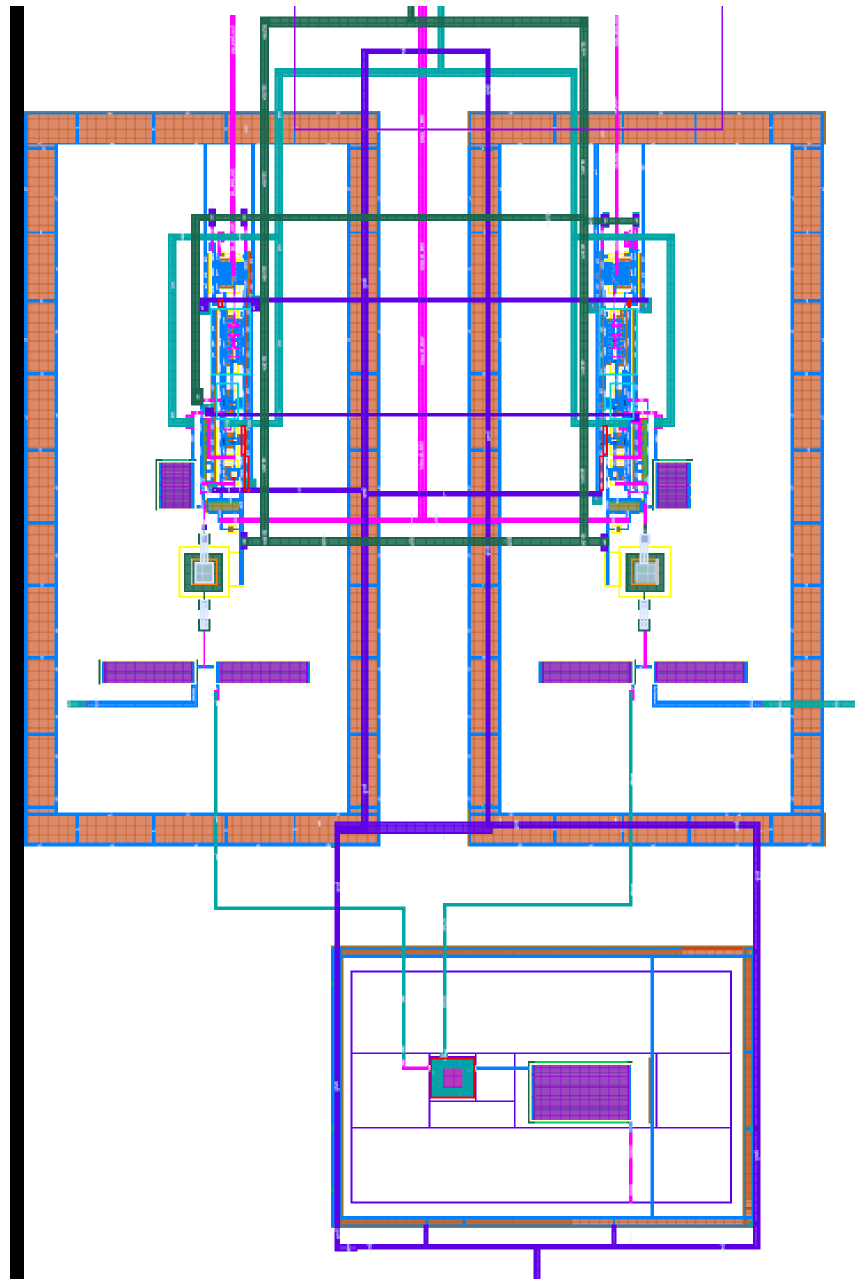
(a)

Figure 3.16: (a) High-level circuit schematic of a dual-junction SPAD and readout circuitry



(b)

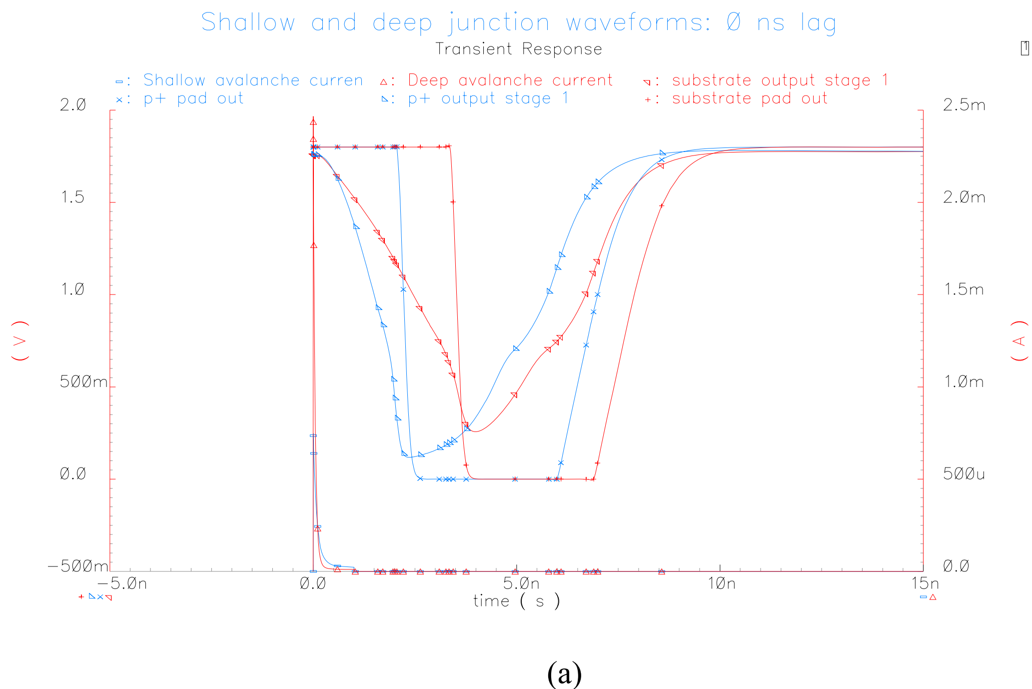
Figure 3.16 (cont.): (b) Simulation model a of a dual-junction SPAD and readout circuitry.



(c)

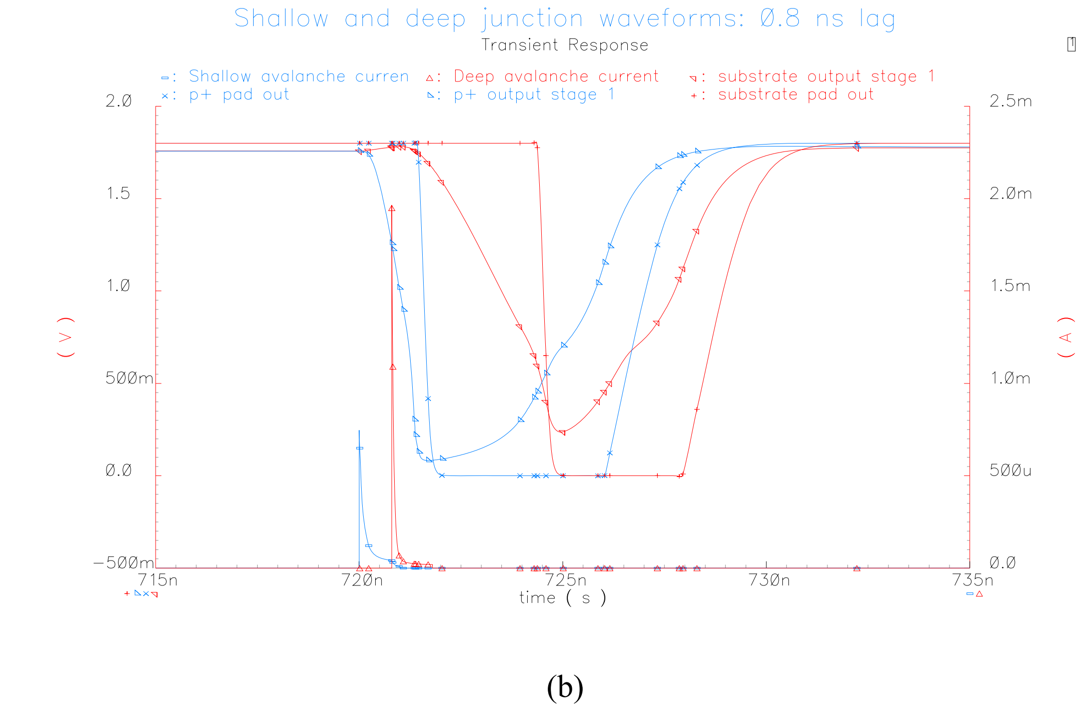
Figure 3.16 (cont.): (c) Layout of a dual-junction SPAD and readout circuitry.

Layout of the circuit also requires special care (Figure 3.16 (c)). Quenching is achieved using analog resistors available in the RFCMOS process, which can sustain the high voltages see in device operation, and which exhibit a relatively small temperature coefficient. Similarly, the high-pass capacitors are implemented using a capacitor structure de-coupled from the substrate, in the form of a Metal-insulator-Metal (MiM) capacitor. The substrate node of the diode, which is biased at a high negative voltage, is isolated from the circuit ground using an undoped substrate ring, as well as substrate tie-downs, which isolate the circuitry (two red rings in the figure). Finally, signal paths of the two channels are matched to reduce inter-channel jitter and match delays.

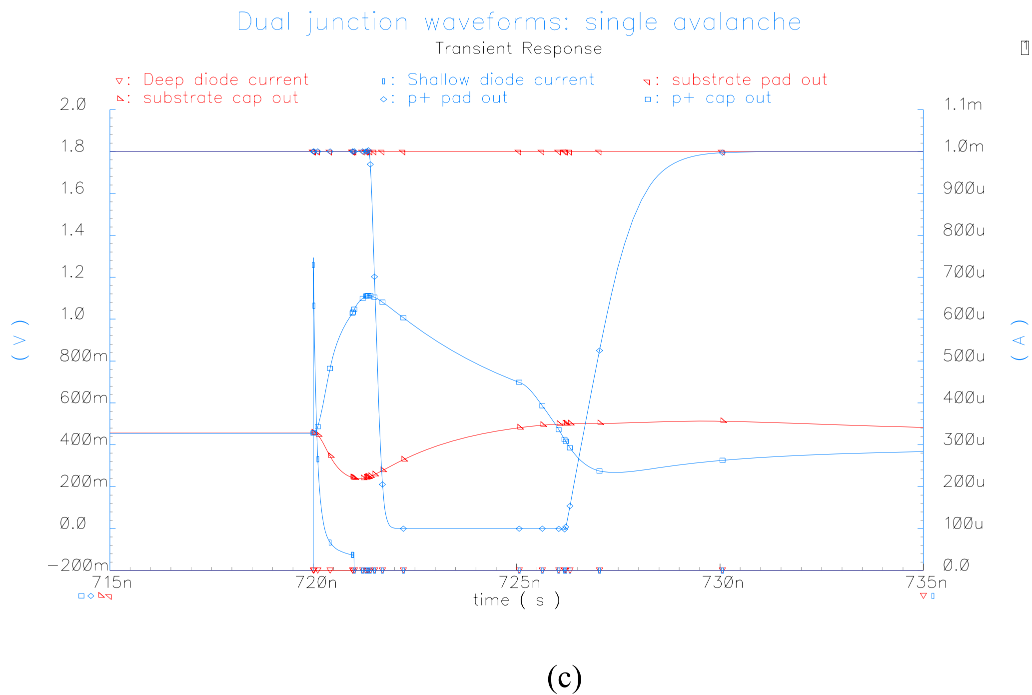


(a)

Figure 3.17: Electrical simulation results for dual-junction SPAD. (a) Photons are absorbed 0.8 ns apart.



(b)



(c)

Figure 3.17 (cont.): Electrical simulation results for dual-junction SPAD. (b) photons are absorbed simultaneously and (c) a photon is absorbed only by the shallow diode.

Figure 3.17 demonstrates the operation of the circuit under various critical conditions. The cross-talk between the two channels causes the internal read-out node of the adjoining junction to ripple, but, regardless of the delay between events, the sampling circuitry eliminates this ripple.

3.6. Conclusion

A new GM-SPAD device has been introduced. The novelty of the new device is in the use of the shallow-trench isolation structure, commonly used in CMOS processing, as the guard ring. We showed that the benefits of the new guard-ring are two-fold:

- It is more efficient in area due to the high dielectric strength of silicon dioxide, compared with silicon.
- It localizes the electric field in a more efficient manner than the diffused ring.

Using process data derived from fab documentation, we built a physical and electrical model of the device. These models estimated high localization of the electric field, and a 5 ns recharge time in passive recharge mode. We also described two novel circuits – one for optimizing the performance of the device through active-recharge, and the other for uniquely detecting two wavelengths simultaneously.

Because deep-submicron technologies utilize high doping concentrations to form shallow junction, the expected spectral response is skewed towards shorter wavelengths. Thanks to the small achievable pixel dimensions and the efficient guard

ring, junction capacitance is smaller than that in SPADs manufactured using traditional guard rings. This translates to lower power consumption. The higher achievable fill factors make large-scale SPAD arrays a real possibility.

Acknowledgement:

This chapter, in part, is a reprint of the material as it appears in the following publications:

H. Finkelstein, M. J. Hsu, S. Esener, "STI-bounded single-photon avalanche diode in a deep-submicron CMOS technology", *IEEE Electron Device Letters*, vol. 27, no. 11, 2006.

H. Finkelstein, M. J. Hsu, S. Esener, "An ultra-fast Geiger-mode single photon avalanche diode in 0.18 µm CMOS technology", *Proceedings of Advanced Photon Counting Techniques, SPIE Vol. 6372*, Boston, MA, 2006.

H. Finkelstein, M. J. Hsu, S. Esener, "A compact single-photon avalanche diode in a deep-submicron CMOS technology", *Proceedings of International Conference of Solid-State Devices and Materials*, Yokohama, Japan, 2006.

The dissertation author was the primary investigator and first author of this paper.

4. DEVICE CHARACTERIZATION

4.1. Introduction

Single-photon detector characterization carries with it many challenges, and is, in many respects, the most difficult phase of detector development. In order to determine the figures-of-merit described in Section 2.8, one must address the following challenges:

- Create and measure incident photon stimuli within a resolution of a single photon, for a wide range of wavelengths.
- Measure arrival times of analog signals with precision of tens of picoseconds without introducing jitter.
- Design test hardware with multi-GHz bandwidth, avoiding even slight reflections, which can manifest themselves in statistical measurements as experimental artifacts.
- Capture large amounts of data (up to 1 billion time stamps per second) in real time, and statistically process this data.

Due to the high speed and low jitter of the novel SPAD, its performance often exceeds that of the best test instrumentation. We resorted to new experimental techniques, as described in this chapter. Test hardware was designed as part of this work and data was initially captured using a Labview program which interfaced with a real time digital oscilloscope. However, due to the slow acquisition times and large amounts of data, we migrated to commercial test equipment specifically designed for

single-photon measurements. These were provided by two specialty companies – Becker & Hickl and Picoquant.

This chapter describes the experimental methodologies and setups, as well characterization results corresponding to the metrics discussed in section 2.8. We conclude by discussing the implications of the measured parameters on use of the new detector for various applications.

4.2. Test Devices

IBM's 0.18 μm 1P6M RFCMOS process was chosen for fabricating the new devices. A 0.18 μm process allows for large-scale integration of supporting circuitry for future arrays of the devices, yet the process is mature enough for defect density to be acceptably low. An RFCMOS process was chosen because such processes are usually well-characterized for high-frequency operation, and usually contain many analog structures, such as metal-insulator-metal (MiM) capacitors, which may become useful in some of the more advanced SPAD applications.

We fabricated two test chips using the MOSIS service. The first chip, shown in

Figure 4.1(a), contained various structures aimed to prove the planarity of the STI-bounded device. It was packaged in a small form-factor, 44-pin plastic lead chip carrier (PLCC) package. The second chip, shown in

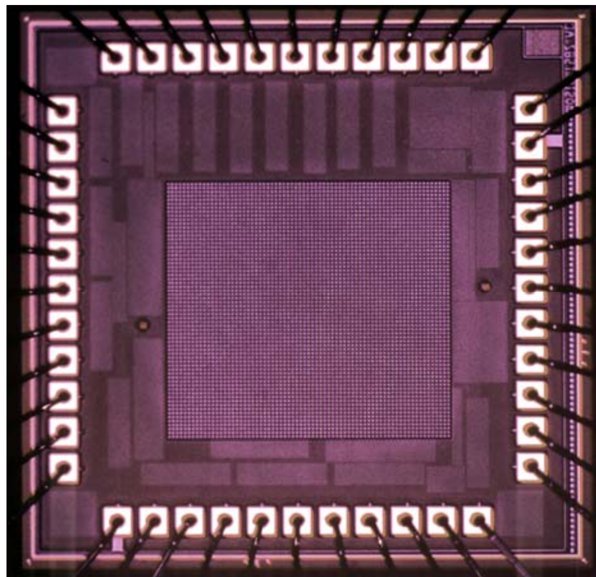
Figure 4.1 (b), contained shallow and deep junction SPADs, including various quenching, recharge and support circuits. It was packaged in a 121-pin plastic grid array (PGA) package.

The second chip has two important improvements over the first one. The first is the removal of auto-fill patterns from above the active area. These are isolated distributed metal polygons added by the foundry for reducing mechanical stress across the wafer (Figure 4.2). While having no electrical effect, they block parts of the active area, thus reducing the effective fill factor. The second improvement is the removal of the cobalt silicide layer from the active area. This layer is used to improve the electrical contact to the source/drain regions in CMOS transistors. However, despite its thinness (35nm – 50nm) it has significant absorbance, mainly in longer wavelengths [139]. Its removal is expected to improve the quantum efficiency.

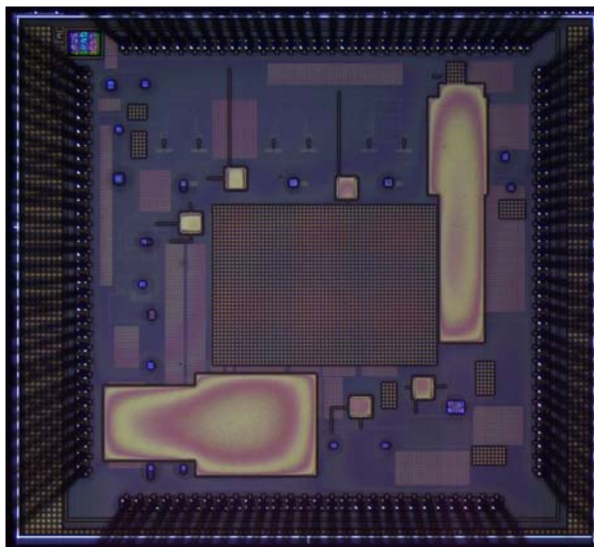
4.3. Validation of Junction Planarity

In order to validate the planarity of the STI-bounded junction, we examined three different junction geometries similar to those theoretically analyzed in Section 2.4. Non-planar junctions were formed by placing a polysilicon gate over the curved junction regions (Figure 4.3). Because the shallow-trench isolation is automatically formed by the fab only in non-active areas, it is not created at the gate-covered edges of the diode, which are considered to be a transistor channel. Consequently, the edges of the junction create a semi-spherical junction and its corners correspond to a semi-

spherical junction (Figure 4.4). For the former, a 45 degree polysilicon corner replaced the 90 degree one. Measurements were made on a socketed device through a ZIFF adapter (Figure 4.5), using an HP4156A Semiconductor Parameter Analyzer.

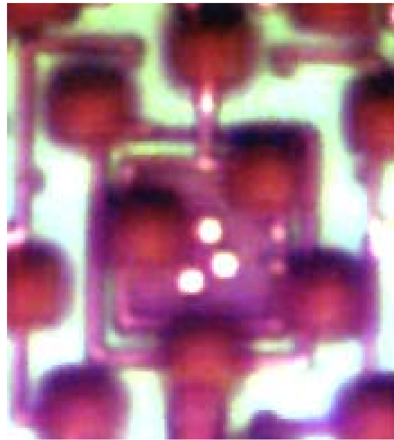


(a)

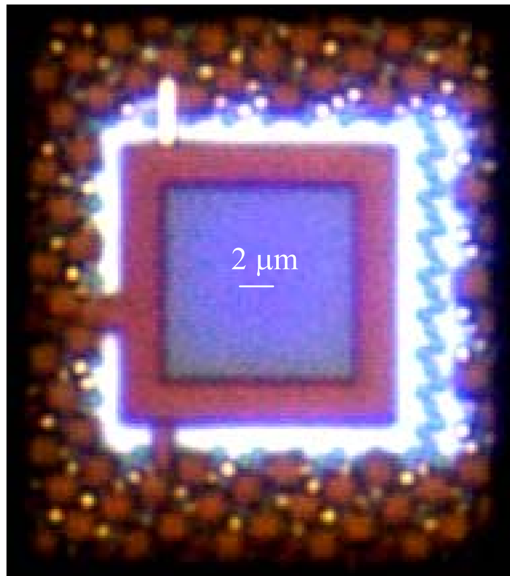


(b)

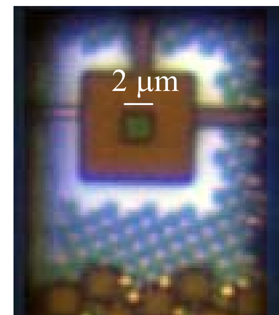
Figure 4.1: Microscope image of (a) first test chip, aimed at proving the new SPAD device and investigating its avalanche behavior, and (b) second test chip containing integrated quenching and recharge devices.



(a)



(b)



(c)

Figure 4.2: Microscope image of (a) a high fill-factor SPAD with a 14 μm side active area (internal square) and 25% fill factor; and (b) a high-speed small-pitch SPAD with 2 μm side active area and 3% fill factor. The surrounding spheres are fab-generated structures for stress reduction.

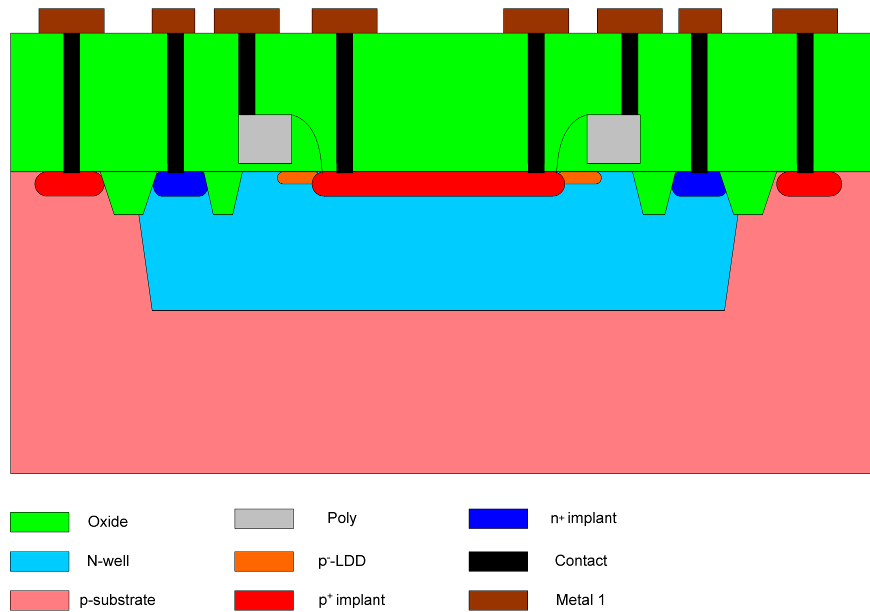


Figure 4.3: Cross-section of non-planar p+/N-well junction

These junctions do not precisely correspond to spherical and cylindrical junctions because of the formation of a lightly-doped drain (LDD) implant at the edge of the implant. An LDD results from reduced doping of the drain region in order to control the drain-substrate breakdown [91]. The reduced doping gradient between the drain and channel lowers the electric field in the channel in the vicinity of the drain. Therefore, we expect that junctions bounded by LDD will exhibit a somewhat higher breakdown voltage than those without such an implant.

The I-V plots of the three structures is shown in Figure 4.6. Two effects are visible. The non-STI-bounded junctions exhibit a leakage current, and there is a clear difference in the breakdown voltages between the different junctions. The former is due to tunneling through the high field across the oxide at the polysilicon edge, as can be seen by the localized high field in an ISE simulation of the structure (Figure 4.7).

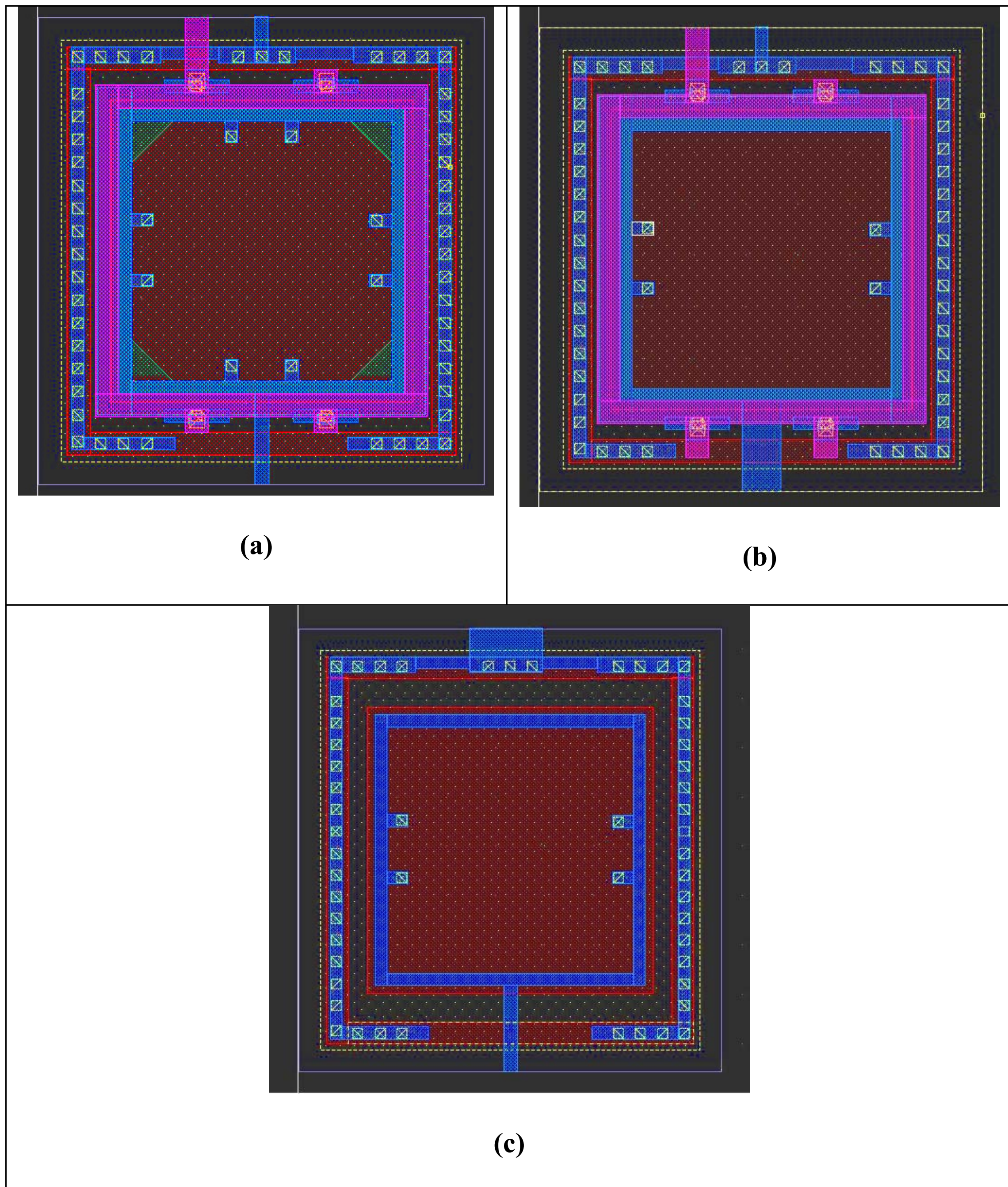


Figure 4.4: Layout of (a) a semi-cylindrical, (b) a semi-spherical and an STI-bounded diode.

Figure 4.5: Socketed test device

The measured breakdown voltages for the semi-spherical and semi-cylindrical junctions were 9.2V and 10V, respectively. The STI-bounded junction exhibited breakdown at 11V. From Equation 2.25 we calculate:

$$\frac{V_{cyl}}{V_B} = \left[\frac{1}{2}\left(\eta^2 + 2\eta^{6/7} \right) \ln\left(1 + 2\eta^{-8/7} \right) - \eta^{6/7} \right]$$

Equation 4.1

and

$$\frac{V_{sp}}{V_B} = \left[\left(\eta^2 + 2.14\eta^{6/7} \right) - \left(\eta^3 + 3\eta^{13/7} \right)^{2/3} \right]$$

Equation 4.2

These ratios are 0.84 and 0.90 for the semi-spherical and semi-cylindrical junctions, compared with the theoretical values from Equation 4.1 and Equation 4.2, which maximize at 0.62 and 0.64, in agreement with the above discussion on the effects of the LDD.

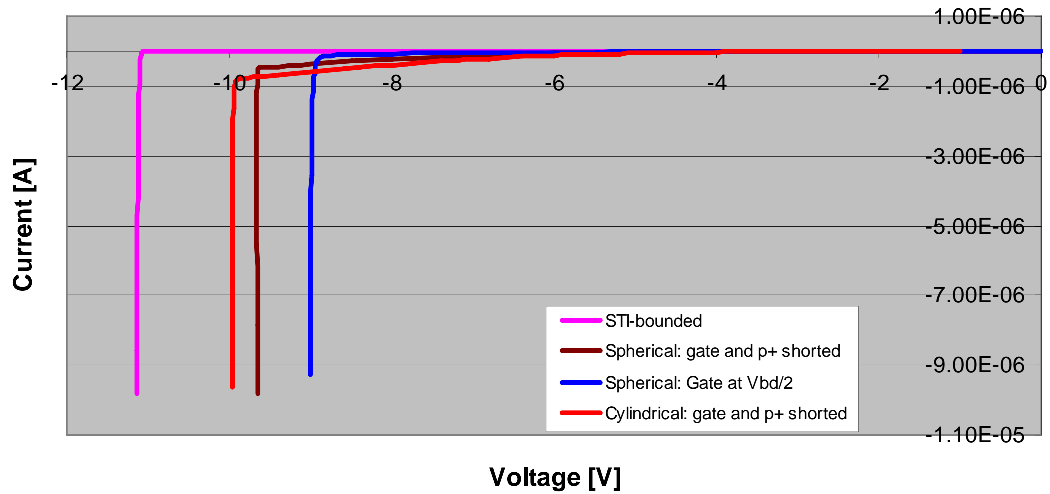


Figure 4.6: Reverse-bias I-V curves for various junction structures.

In order to determine whether the measured breakdown voltage of the STI-bounded junction is in agreement with the assumption of planarity, we start by calculating the reverse-bias junction capacitance from the expression provided by the fab [114]:

$$C = \frac{C_0}{\left(1 - \frac{V_A}{V_{bi}}\right)^m}$$

Equation 4.3

where V_A is the breakdown voltage; and C_0, the zero-bias capacitance, the built-in voltage, V_{bi}, and the fitting parameter, m, are given. From Equation 2.11 and Equation 4.3 we calculate the depletion region's width at breakdown, W = 0.2659 μm. From Equation 2.6 we can now calculate $N_B = 1.32 \times 10^{17}$ cm^{-3}. Equation 2.19 results in a calculated value of V_B = 8.91V. This value is lower than the measured value. The

higher measured breakdown voltage certainly does not imply a premature breakdown. It is probably due to the series resistance of the diode, which lowers the effective voltage seen by the diode, especially in the breakdown regime where the current is large.

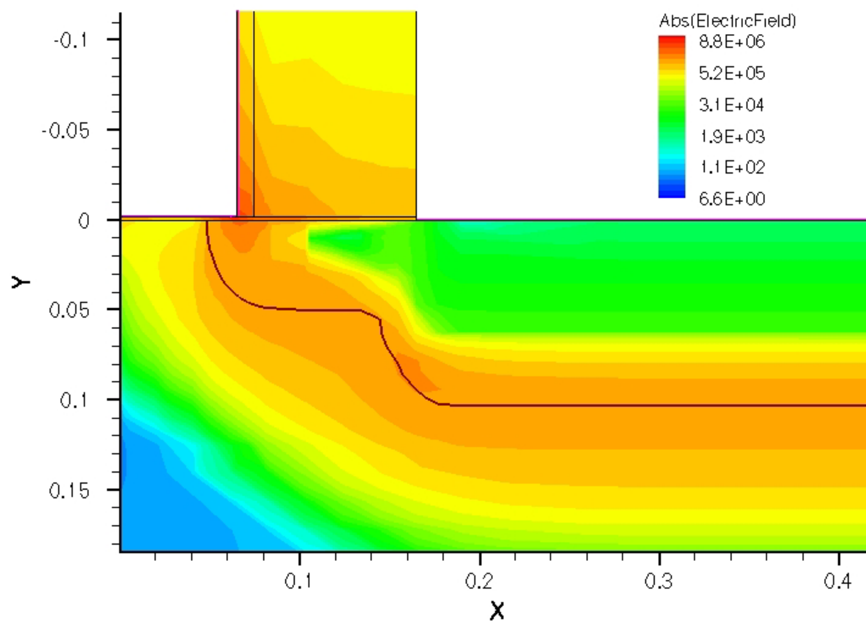


Figure 4.7: Electric field distribution at the edge of a gate (not shown) defining a non-planar LDD-bounded pn junction.

In order to confirm that the junction does not break down at its edges, we focused a pulsed laser beam to a 5 μm spot size and scanned its position across a 7 μm SPAD, counting the output avalanche pulses using a Becker-Hickl MSA-1000 counter. A detailed description of the setup appears in Section 4.7. If edge breakdown occurred, we would expect the photocounts to peak at the edge positions and exhibit a valley between these peaks. A planar junction will exhibit a detection peak at its center. The results, shown in Figure 4.10, indicate a planar junction.

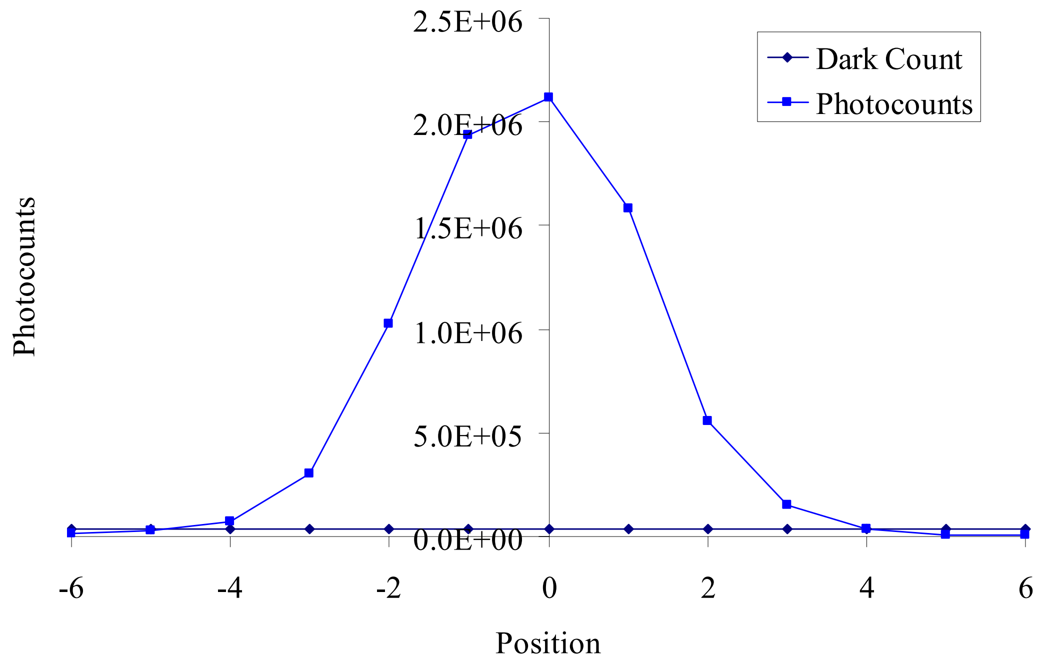


Figure 4.8: Photocounts as a function of beam position on the device.

4.4. Avalanche Pulse

The avalanche pulse was first measured in the first test chip using a device with external quenching. In order to minimize the parasitic capacitance, which is expected to increase afterpulsing and the recharge time, we design a printed-circuit board (PCB), shown in Figure 4.11 (a), with a surface-mount resistor. Measurements were performed using a real-time Tektronix TDS3032 Oscilloscope with 300 Msample/sec sampling rate (Figure 4.11 (b)).

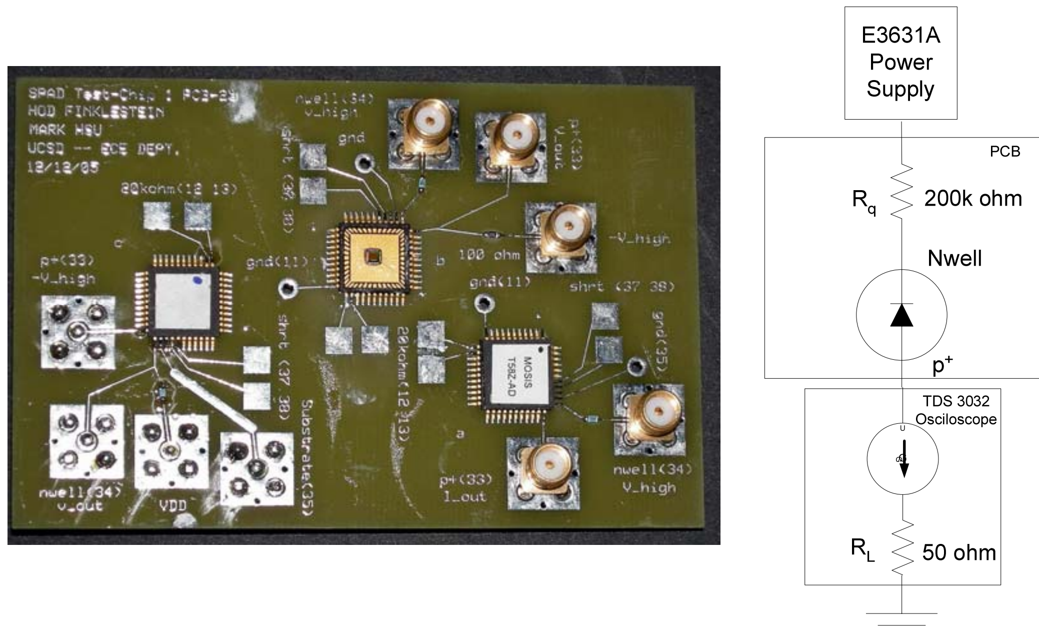


Figure 4.9: (a) Photo and (b) schematic of PCB for measuring avalanche behavior of externally-quenched SPAD

A single avalanche pulse is shown in Figure 4.13. The fast rise tine, simulated to be in the tens of picosecond is distorted by the limited bandwidth of the scope as well as by RC delays of the package and cables. The overall pulse duration is approximately 3 ns. By integrating the area under a single pulse, we determine that at an overbias of 0.4V, 5.5 x 10^6 electrons flow during an avalanche. This is on the same time scale as other measured results for hybrid-quenched SPADs [70]. The corresponding junction capacitance is the ratio of the measured charge and the overbias, 2.2 pF. This is dominated by the parasitic capacitance since the junction node is connected to the external resistor via SMA connectors and board traces.

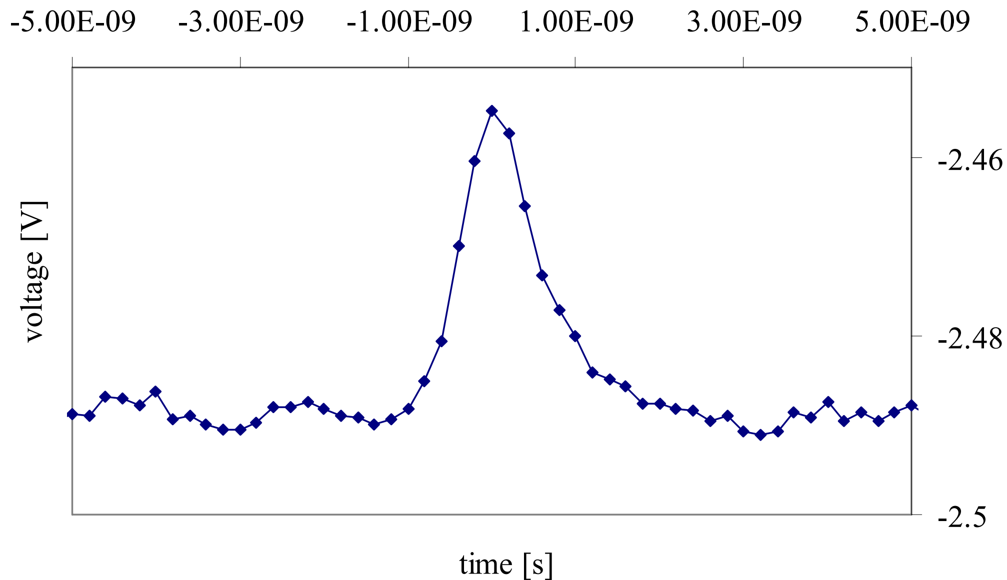


Figure 4.10: Single avalanche current pulse at 0.4V over-bias, corresponding to approximately 5.5x10^6 electrons.

The Q-V relationship of the junction under avalanche is shown in Figure 4.11, with a slope corresponding to a junction capacitance of 1 pF. Avalanche charge is expected to be greatly reduced when the quenching resistor is integrated on the same die (Figure 4.12) as the SPAD and an output stage buffers the junction from the load capacitance. However, once the junction has been isolated with a buffer and integrated resistance, the avalanche charge cannot be measured directly (it is possible to measure it indirectly by recording the average current drawn from the p+ supply, but statistical variations in the pulse amplitude renders this option useless). Integrated SPAD devices were fabricated on a second test chip and are described in the following sections.

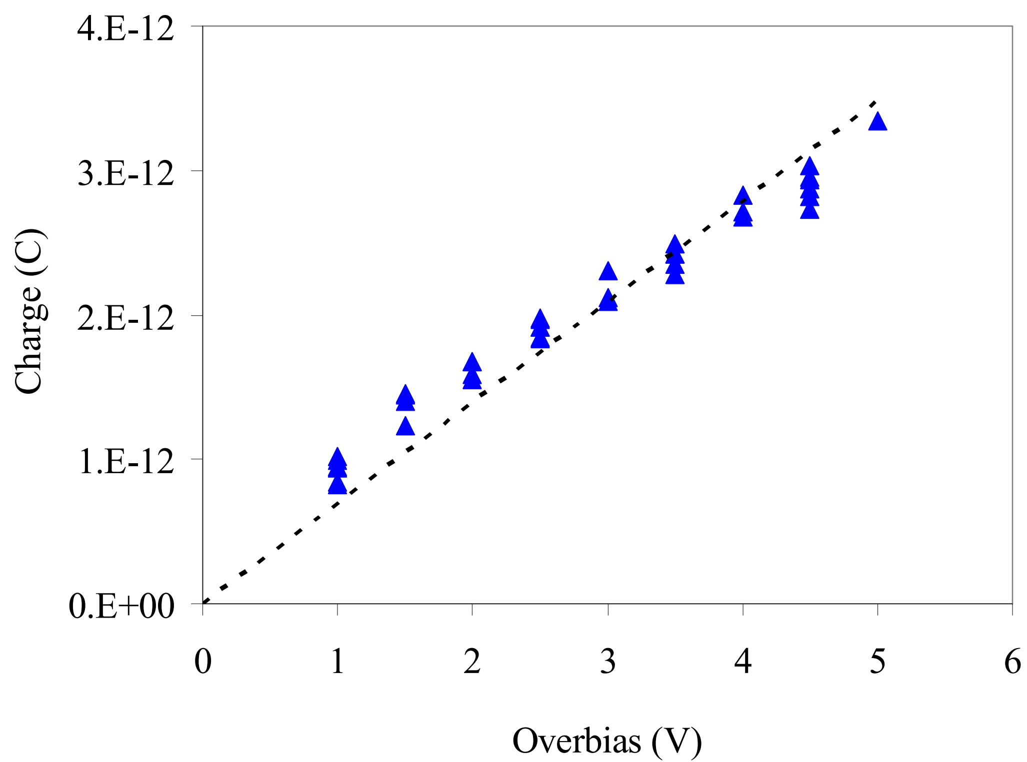


Figure 4.11: Avalanche charge as a function of overbias for hybrid-quenched SPAD.

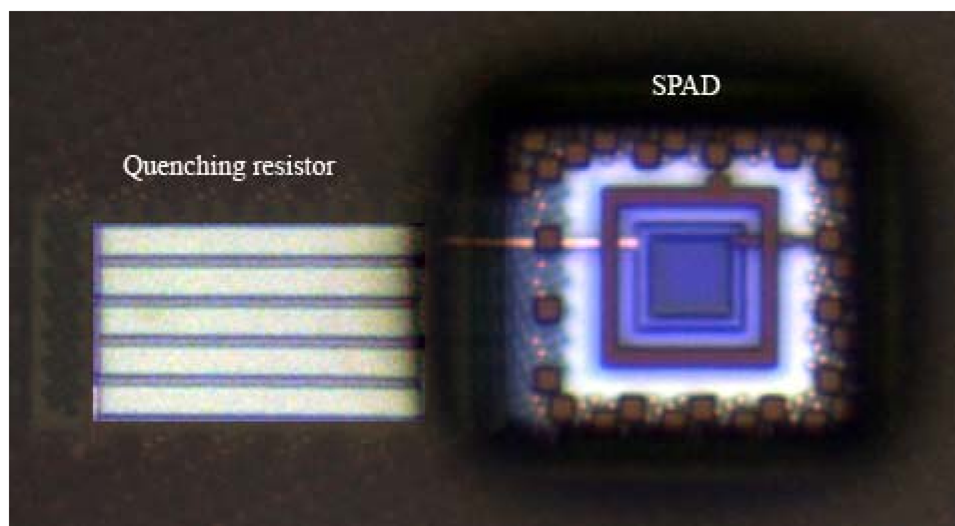


Figure 4.12: Microscope image of SPAD with integrated quenching resistor.

4.5. Dead Time

As described in Section 2.8.2, the dead time of the device is crucial for observing fast phenomena in real time and for increasing the dynamic range of the detector. We developed several novel methods to measure this dead time.

Using a real-time Tektronix TDS 3032 oscilloscope, we set the threshold of a covered SPAD to the highest level where a signal can be observed, and set the scope to “Infinite Accumulate” mode (Figure 4.13). In this manner, an oscilloscope sweep is recorded with the triggering dark pulse at the center of the sweep. We can therefore view an accumulation of all pulses 1 μs before and 1 μs after the arrival of this trigger.

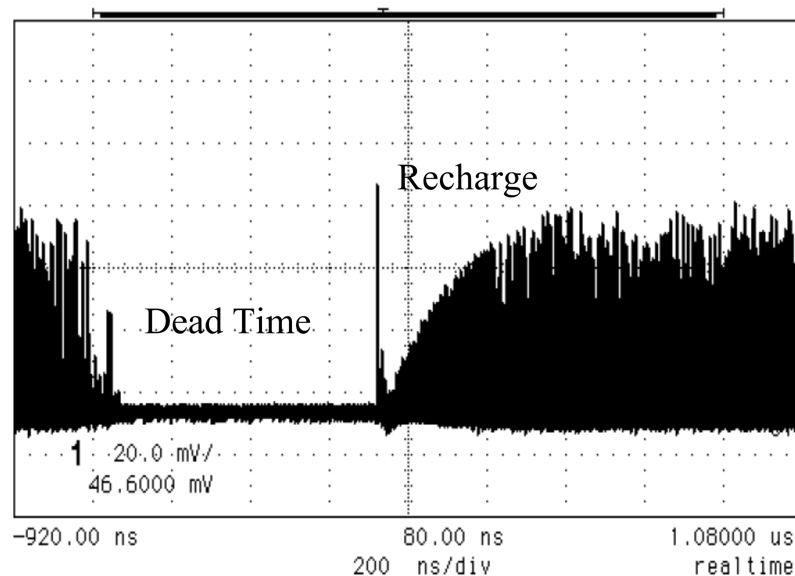


Figure 4.13: Oscilloscope image showing hybrid-quenched device dead time and recharge.

The results show that for a certain time before the triggering pulses arrival, no pulses appear. This is due to the dead time of the device. In other words, a certain time is required for the recharging of the junction before a high pulse can be generated. Following the arrival of the pulse, pulse heights, due to afterpulsing, follow an exponential curve, due to the recharging of the junction. Shorter pulses are far more likely to occur than the higher pulses, due to the exponential behavior of trap lifetimes, discussed in section 2.8.3.3. The implications of this behavior are discussed in more detail in Section 4.6. When the trigger threshold is reduced, so do the dead time and recharge time.

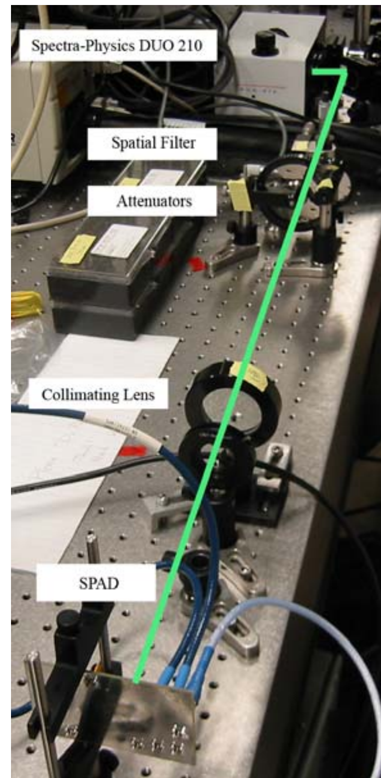
Another new method for measuring the device dead time uses a zero-dead-time counter, such as the Becker-Hickl MSA-1000, in combination with a pulsed light source (Figure 4.14). In our experiment, we used a Spectra-Physics VSL-227ND-S pulsed UV nitrogen laser with a DUO-201 dye laser extension, with a low-repetition-rate (10 Hz) 3-9 ns pulse width. The counter's trigger was the laser's electrical output and the counter's data input was connected to the SPAD output. The trigger cable length was made to be 1 m shorter than the SPAD output cable, artificially creating a delay between the two signals, thus making it possible to observe the SPAD output before the laser pulse arrival.

The resulting time histogram, shown in Figure 4.15, shows a peak corresponding to the correlated detection of the laser pulses, a zero-count region corresponding to the dead time of the device (at the counter data threshold) and the remaining tail of the afterpulsing. Unlike the previous setup, where a sweep only

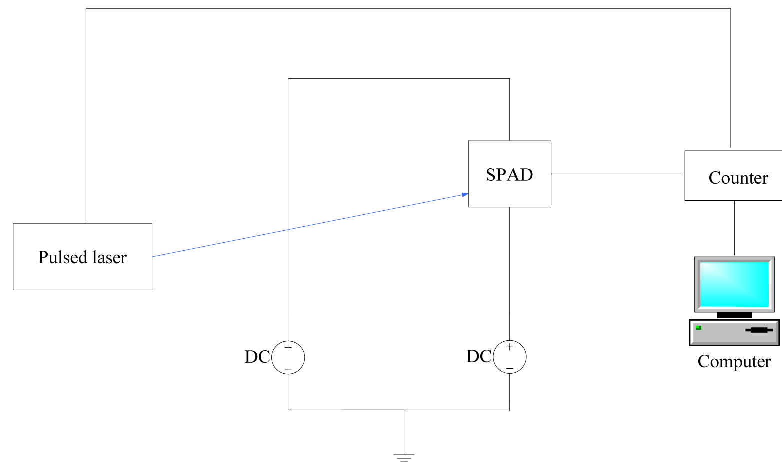
occurred upon a SPAD pulse being observed, the counter is set to sweep without a priori knowledge of whether the photon will be detected. Consequently, some noise is observed before the laser pulse arrival, due to uncorrelated detector dark counts. A dead time of approximately 30 ns is observed using this method and the given threshold conditions.

In the second test chip, the quenching was performed by an integrated PMOS transistor, and resulted in a significantly lower junction capacitance. In order to test this chip, several impedance-matched high-speed test boards were designed. Due to the reduced capacitance resulting from the integrated quenching, the dead time decreased by a factor of 170, to approximately 3 ns (Figure 4.16). This is the fastest SPAD dead time reported to date.

As can be seen from the scope image, the short dead time is accompanied by a high afterpulsing probability. The reasons for this and remedies are discussed in Section 4.8.



(a)



(b)

Figure 4.14: (a) Photograph and (b) schematics of dead-time measurement setup using time-histograms

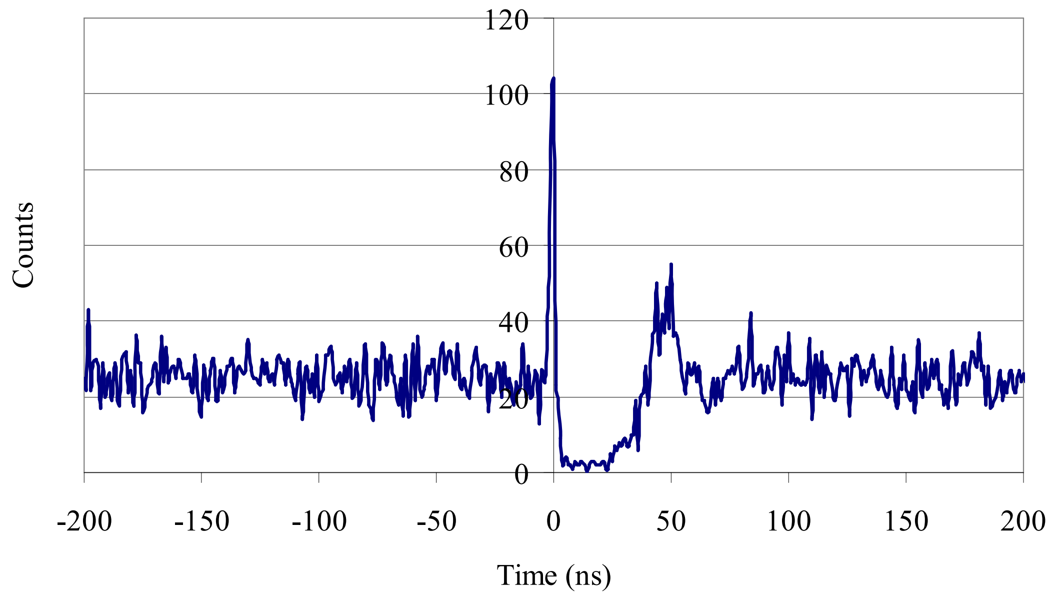


Figure 4.15: Time-histogram of dark pulse arrivals for the first test chip pixel.

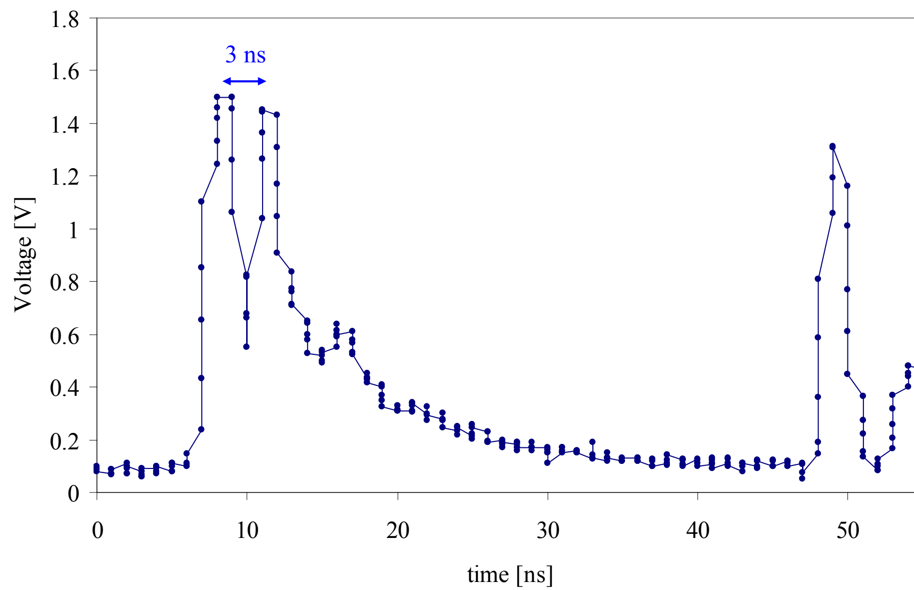


Figure 4.16: Oscilloscope snapshot of integrated 7 μm SPAD with source follower output.

4.6. Dark Current

In the first prototype chip, avalanche behavior was studied using an external surface-mounted 200 kΩ quenching resistor on a custom-designed test board (Figure 4.9). The measured output capacitance of the SPAD is approximately 500 times higher than the capacitance simulated in the integrated quenching scheme. Consequently, more charge carriers take part in each avalanche, a larger proportion of traps become filled, and after-pulsing increases (Figure 4.17), in agreement with previous SPAD results [70].

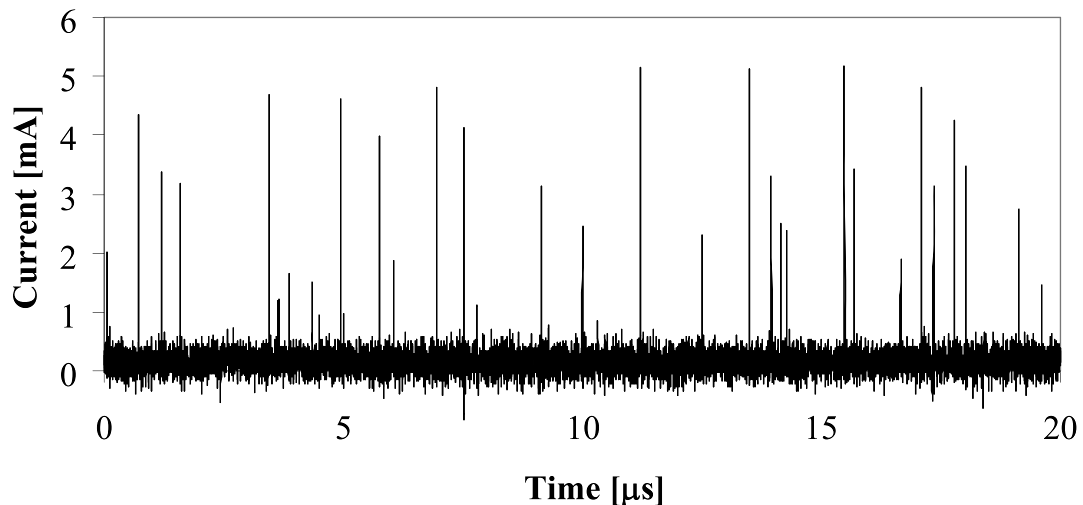


Figure 4.17: Avalanche pulses at 2.5V over-bias.

In order to verify that these dark counts indeed arise out of released traps, we calculated the auto-correlation of the times-of-arrival of the dark counts. This was

done using Microsoft Excel by calculating differences of the time-shifted series. The results, shown in Figure 4.18, confirm the exponential time-probability characteristic of afterpulsing.

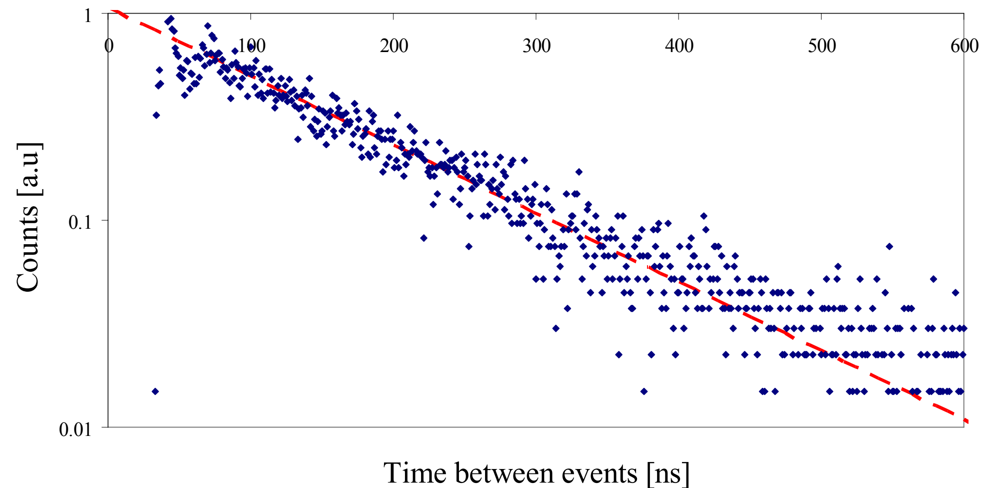
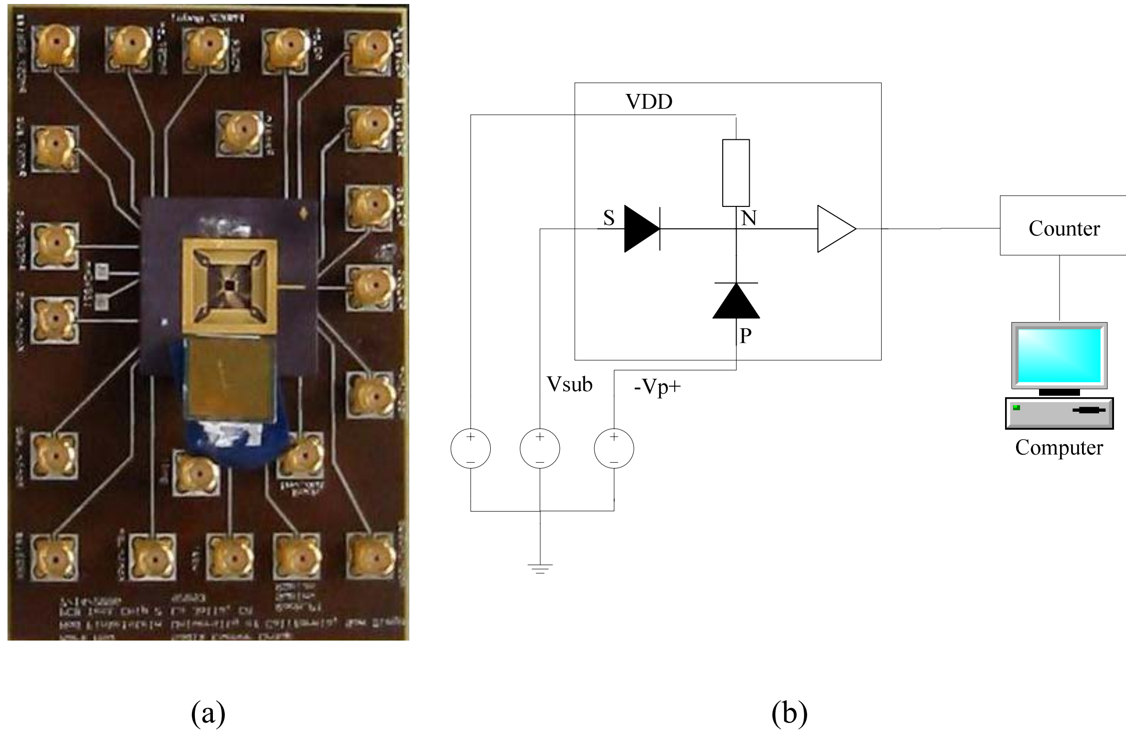


Figure 4.18: Time-histogram of dark counts processed by time-shifting method.

The dark current statistics of the pixels with integrated quenching (second test chip) has been studied using two setups. In the first, the output of a covered, biased, SPAD, was connected to a Tektronix TDS 3032 oscilloscope and read out to a Labview program. The program detected the pulse peaks and recorded their time of arrival to within the scope's precision (0.2 ns). The results were processed offline in Matlab and Excel. The benefit of this method lies in the high time precision of the measurements. Its deficiencies are: (i) Single oscilloscope frames are captured at a time, limiting correlation studies to 2 μsec and biasing results towards short correlation times (there are 9,999 pairs of points per frame separated by 0.2 ns but only one pair separated by 1.998 μs). (ii) Because oscilloscope images are captured,

data file sizes explode to tens of megabytes. (iii) Since all data processing is done in software, processing times are very long (6 hours using a fast workstation for a single typical experiment).



(a) (b)

Figure 4.19: (a) Printed circuit board designed for second test chip characterization. (b) Photo of experimental setup for measuring dark pulses.

The second setup for measuring dark pulse statistics is illustrated in Figure 4.20. The SPAD output is split using a high-speed power-splitter (for signal integrity) and fed to the *trigger* and *data* inputs of an MSA 1000 counter. The counter, which has no dead time between counts, is used to construct a time histogram of the dark pulses. The setup, first introduced in this work, relies on the time-invariance of the dark counts, when randomly chosen and measured over a long enough duration. In

other words, although afterpulsing is a correlated process, originating from a single "seed" pulse, by randomly choosing the triggering pulse, we ensure an un-correlated triggering signal. Moreover, at high count rates, afterpulsing becomes a self-perpetuating process, and, as such, each pulse in fact contributes as a "seed".

The benefit of this method is in its simplicity and in the instantaneous derivation of the histogram, because most of the processing is performed in hardware. The main drawback is the skewing of the data towards shorter correlation times. Unlike the first setup, where each pairs of arrival times is considered, here we only consider the correlations between the (randomly selected) first pulse of a frame, and the following ones. This results in a loss of "coherence" for large value of τ where correlation with the original pulse diminishes.

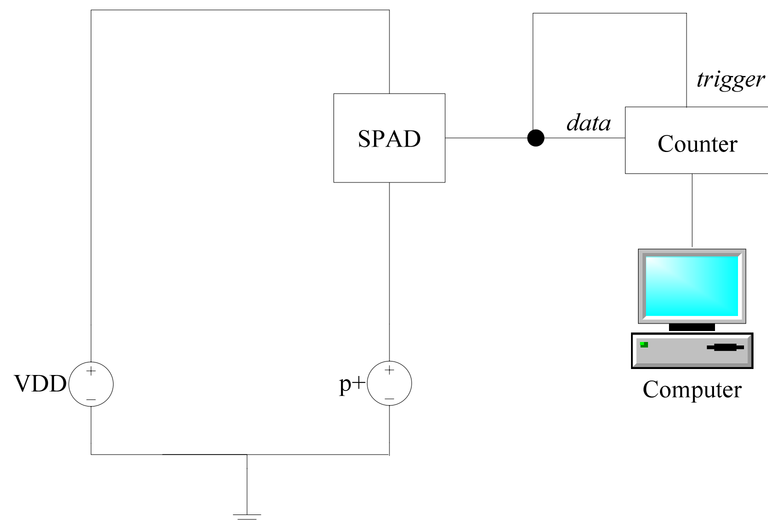


Figure 4.20: Setup for measurement of dark current statistics

The results using the first setup, shown in Figure 4.21 show four distinct regions:

a) The initial dead time, of approximately 3 ns.

b) The beginning of an exponential drop-off, characteristic of afterpulsing.

c) A slow decline in counts, lasting approximately 750 ns, down to a rate of 1000 counts per second.

d) A fast decline in counts, until counts are negligibly correlated.

One hypothesis for explaining this observation assumes that at the time the counter is triggered an avalanche occurs, filling a large number of traps. Due to the short dead time of the SPAD, many of these traps are released, with an exponentially-decaying probability, thus accounting for the behavior in the second region. However, soon afterward (starting one dead-time period after the first afterpulse), these afterpulses cause secondary afterpulses. These result in a pseudo-steady-state population of filled traps, and a self-sustaining process, which lasts for approximately 1 μsec, as, for example in [140]. When the device is biased at different voltages, the absolute dark counts change, but correlation is always preserved for the same period of time. A similar pattern is observed using the second setup, shown in Figure 4.22.

The counter-based setup was used to measure the effects of overbias on dark counts. Results, shown in Figure 4.23, agree with theory. As the overbias increases, so does the electric field, and consequently the avalanche initiation probability. This increase is not monotonic. At low overbias, the avalanche charge, C_jV_{ob} barely suffices to trigger the first inverter in the output buffer chain. As the overbias increases, so does the avalanche initiation probability. Finally, the dark current rate exceeds the dead time of the SPAD, resulting in saturation.

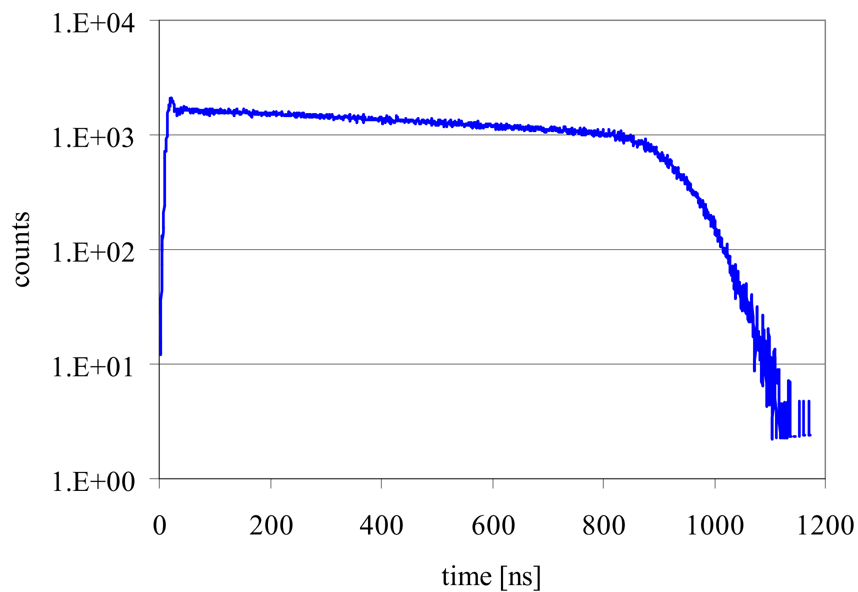


Figure 4.21: Auto-correlation of SPAD dark pulses with integrated quenching, using Labview for post-processing.

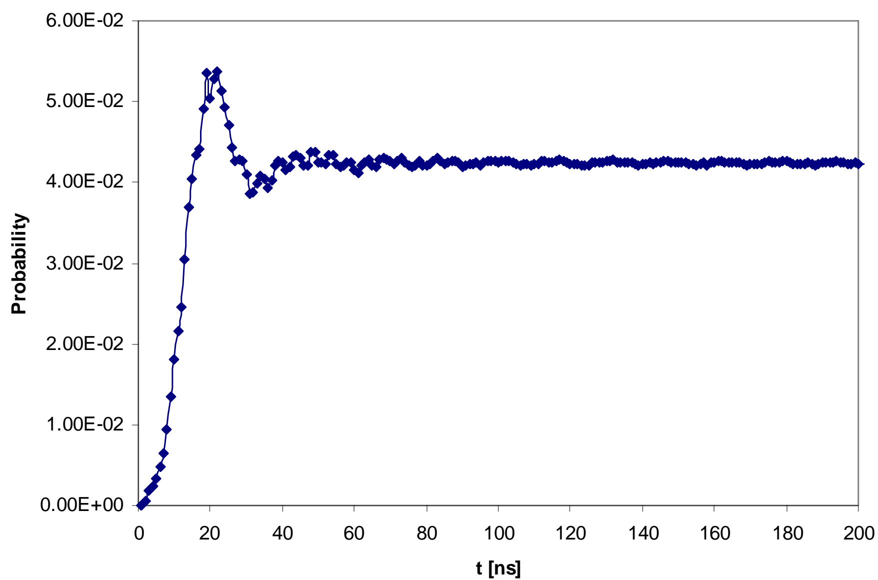


Figure 4.22: Histogram of afterpulses with counter triggered by a dark pulse

The temperature dependence of the dark counts is crucial in determining the mechanism responsible for their formation. Unlike other noise sources which increase with temperature, afterpulsing decreases with temperature. In order to precisely measure the junction temperature, a high-resistance polysilicon resistor with a high temperature coefficient was designed on the test chip, in proximity to the junction. Its resistance was monitored as a hot-air gun heated the junction. Measurements were conducted when the temperature reached a steady-state and its stability was monitored throughout the experiment.

The temperature dependence of the dark counts, shown in Figure 4.23, is typical for afterpulsing and demonstrates that it is the dominant dark noise mechanism at room temperature. The observed shift is a result of accelerated release of trapped charges during the device dead time, as well as an increase in the breakdown voltage with increasing temperature [141].

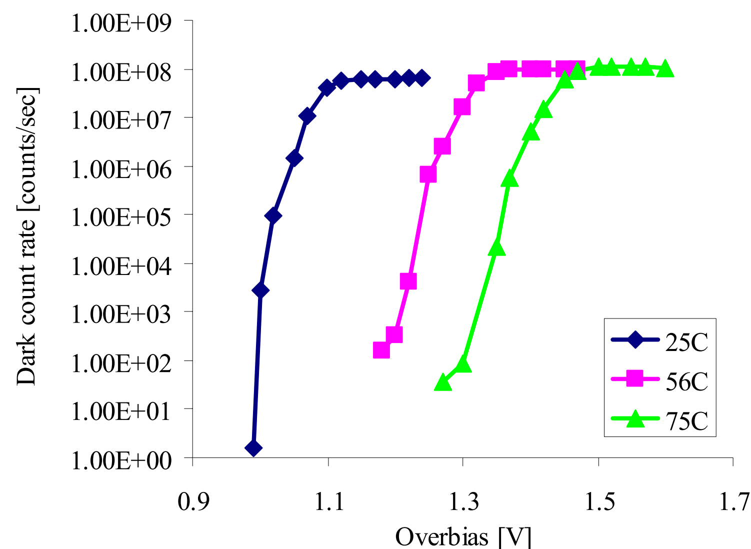


Figure 4.23: Dark count rate dependence on overbias and temperature for a 7 μm SPAD with integrated quenching and inverter-stage output.

The dark count rate we measured is unacceptably high for proper operation. Other silicon SPADs report dark count rates down to 100 counts per second. The causes of this effect are twofold: the short device dead time and the narrow depletion region at breakdown, which facilitates for a relatively high tunneling rate. As we showed in this section, the autocorrelation of the majority of the dark counts follow an exponential relationship, indicating that the former mechanism is the dominant one. It is possible that the tunneling current, which is small but higher than in older-technology devices, facilitates for the higher rate but supplying seed avalanches. This hypothesis is examined in section 4.8, where we describe the characterization of the custom peripheral circuit, which was designed to alleviate this behavior, and which was described in Section 3.4.2.

4.7. Detection Efficiency and Spectral Response

Measurements of detection efficiencies for the single photon detectors require careful calibration. Two approaches were used for these measurements. The first approach is based on a time-correlated method. In order to understand the benefits of time-correlated measurements, consider the scope output in Figure 4.24, where a pulsed laser illuminates the SPAD. It is impossible to distinguish between "signal" pulses, which are correlated to the laser and noise. If we were able to create a histogram of the times-of-arrival of pulses relative to the laser pulses, we could

determine how many correlated pulses were produced by the SPAD. That is the basis of our measurement technique.

A schematic of the setup is shown in Figure 4.25. A 635 nm pulsed laser source, BHL-630, first illuminated a calibrated power meter (Newport 2832C with an 818SL power head calibrated from 400-1100 nm) placed at the plane of the SPAD. Because of its slow response, it, in effect, integrated the photon flux over time. An iris was used to clip the beam such that a sufficiently low photon flux was produced. The electrical output of the laser triggered an MSA-1000 counter. The laser spot was then focused on the active area of the relevant pixel such that the spot under-filled the pixel, and a time histogram was collected by the counter. Under-filling of the pixel is verified by scanning the pixel position perpendicularly to the beam and verifying that a plateau of maximal counts is observed (Figure 4.26). After subtracting the dark counts, the SPAD counts were summed at times correlated to the laser pulses. The number of counts was divided by the number of laser pulses.

Given a beam power W, photon energy $h\nu$, with pulse repetition rate f, and given that the beam under-fills the active area of the detector, the mean number of photons impinging on the detector per pulse is

$$N = \frac{W}{h\nu f}$$

Equation 4.4

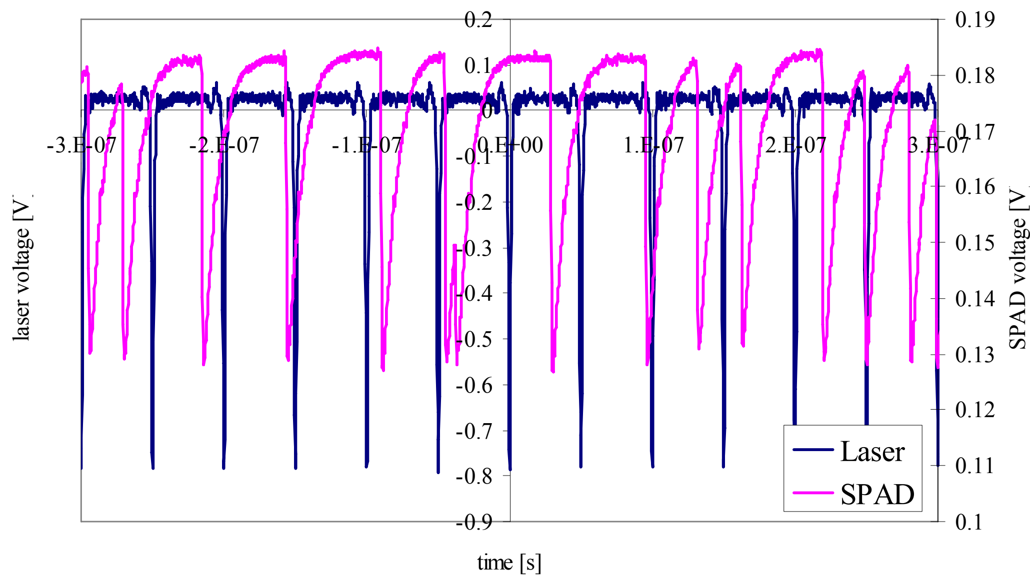


Figure 4.24: Scope image of a SPAD illuminated by a pulsed laser source.

The probability for at least one detected photon is:

$$P_{\geq 1} = 1 - (1 - P)^N$$

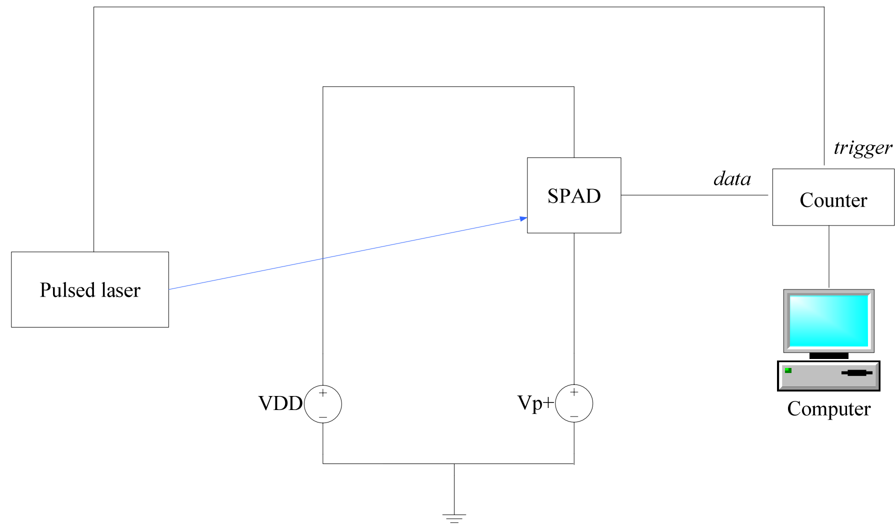
Equation 4.5

with P the single photon detection probability. Using this equation, we calculate the single-photon detection efficiency. A typical time histogram is shown in Figure 4.27, before subtraction of the dark pulses.

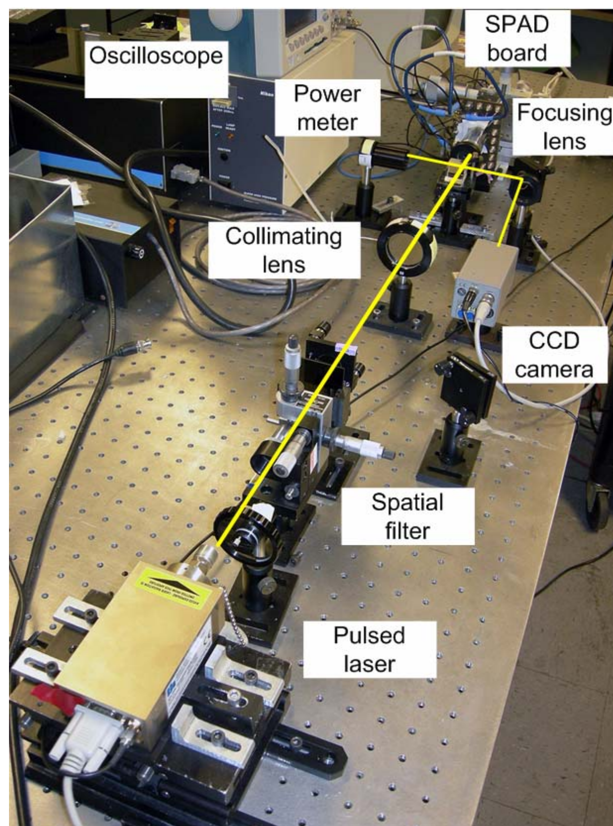
The clear benefit of this method is its sensitivity. By using time-gating, one is able to clearly differentiate between uncorrelated noise pulses, and the correlated SPAD output. Moreover, because the excitation comes from a pulsed source, SPAD statistics can be readily measured. The disadvantage of the method is that it requires a low-jitter pulsed laser source for each relevant wavelength, if a spectral response measurement is desired.

The second method we used for spectral response measurements used a CW source with a broad emission spectrum (100 W Hg arc lamp). The lamp's output was passed through a Digikrom 240 monochromator with a 600 grooves/mm diffraction grating (Figure 4.28). The slit size of the monochromator was externally controllable and the monochromator output could be tuned from 350 nm – 700 nm. An iris and additional optics at the output of the monochromator selected a single mode and controlled the beam size. A beam splitter split the beam to a power meter and to the illumination plane.

Initially, the beam splitter was calibrated. An identical power meter was placed at the imaging plane and the total beam power on the two power meters were measured, thus calculating the wavelength-dependent response of the beam splitter. Specifically, we recorded the slit widths for each wavelength, which corresponded to an identical number of photons per unit area impinging the detector area. Next, the second power meter was replaced by the SPAD, the spot was focused so it under-filled the SPAD active area, and the spectral response was measured. The photon flux was designed such that the average inter-photon arrival times were more than twice the dead time of the device and approximately 10 times slower than the dark count rate. Results are shown in Figure 4.27.



(a)



(b)

Figure 4.25: (a) Schematic and (b) photograph if experimental setup for measuring SPAD detection efficiency and output statistics.

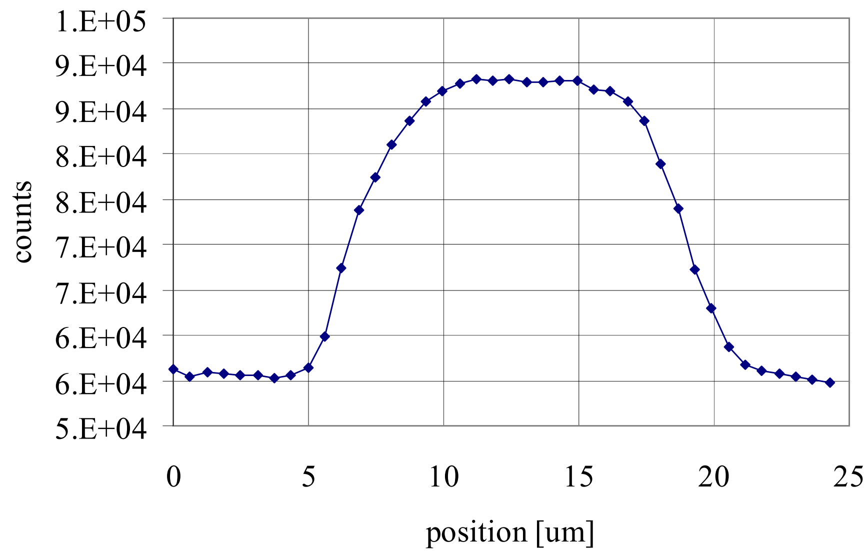


Figure 4.26: SPAD counts as a function of position perpendicular to a pulsed laser source, for optical alignment

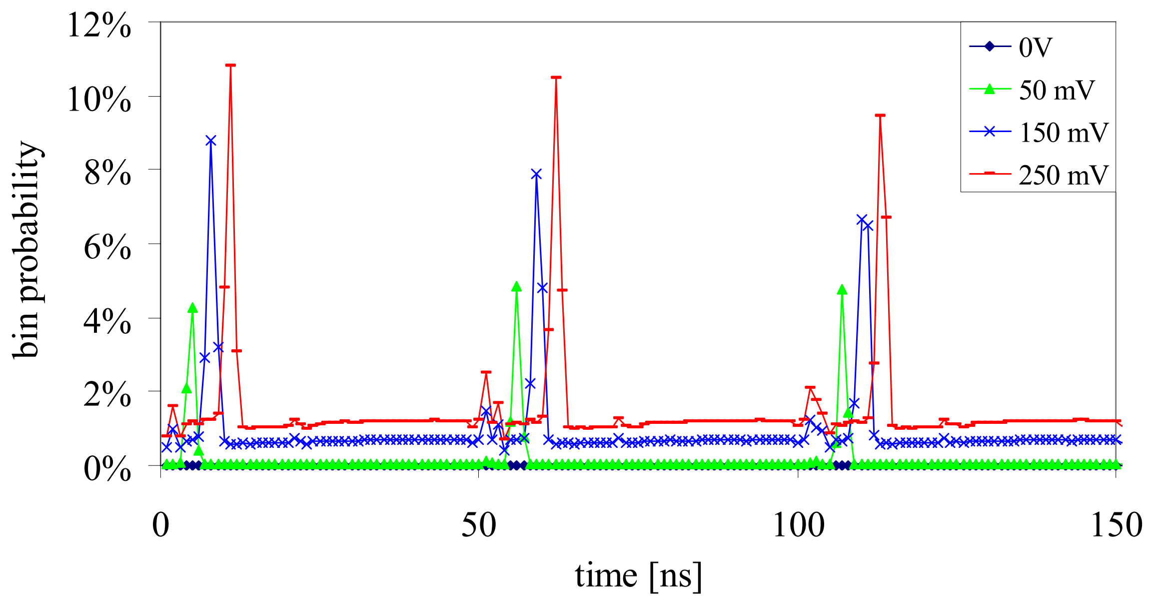


Figure 4.27: Time histogram with pulsed-laser excitation at 635nm for various overbias voltages.

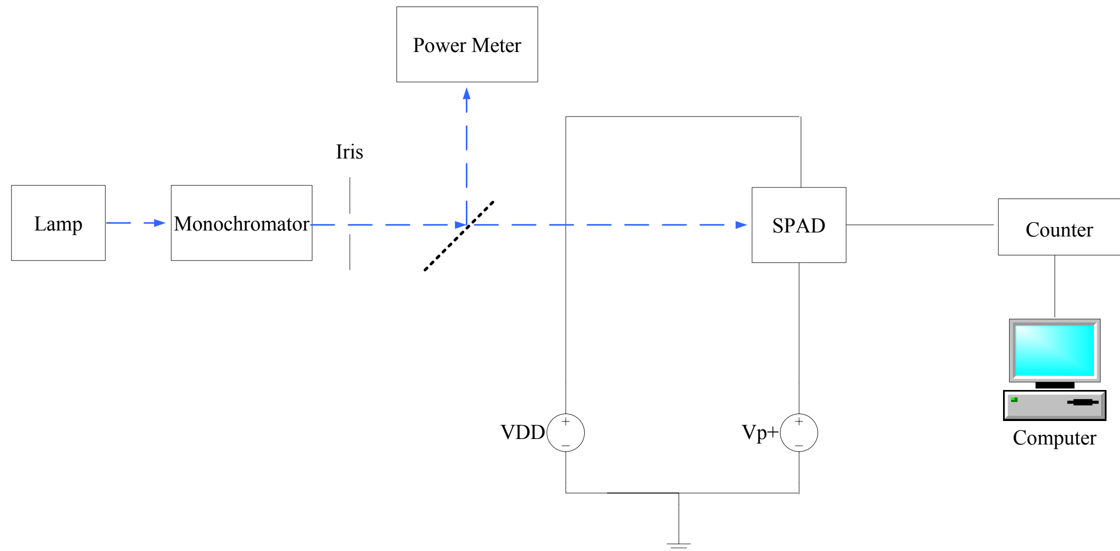


Figure 4.28: Setup for measuring spectral response using a CW source.

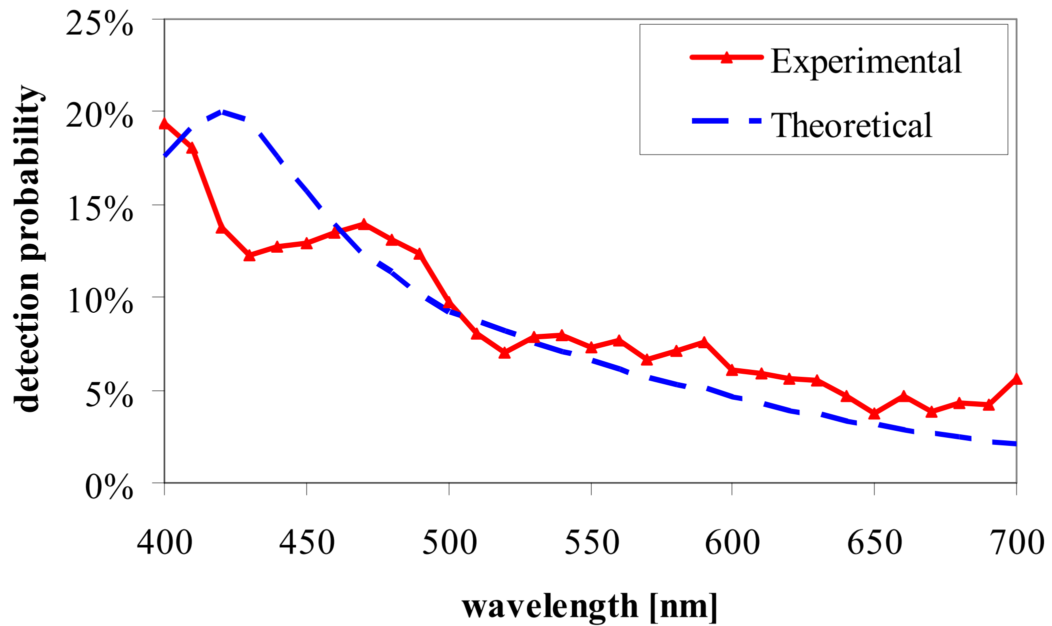


Figure 4.29: Spectral response for STI-SPAD with source-follower output stage at 1.03V overbias using monochromator technique.

4.8. Optimization Using Active Recharge

A microscope image of the circuit described in section 3.4.2 is shown in Figure 4.30. It measures 44 μm per side, 13 times smaller than the smallest active-recharge circuit to date [108].

As discussed in Section 3.4.2, unlike previous active recharge circuits, which were designed in order to *accelerate* the recharging of the device [98, 108, 142-144], the aim of the circuit described here is to release a maximum number of filled traps without initiating an avalanche and without significantly increasing the dead time of the device.

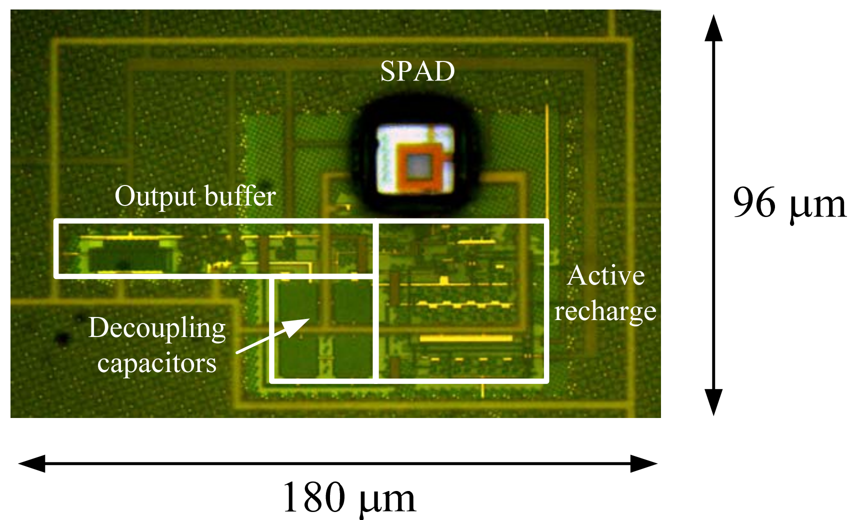


Figure 4.30: Micrograph of the new active-recharge circuit, together with the SPAD, supply-noise decoupling capacitors and the output buffer.

We first evaluated the dead-time performance of the actively-quenched SPAD by operating it at room temperature with a high over-bias, forcing a high dark count. The device output was measured using a Tektronix TDS 3032 oscilloscope and is shown in Figure 4.31, with discernible peaks separated by 3 ns, in agreement with our simulations. This corresponds to a 3 ns dead time – three times faster than the shortest actively-recharged dead time reported to date [108].

In order to gauge the effectiveness of the active-recharge circuit, we compared the dark count rates of two identical SPADs, one having a passive recharge circuit and the other actively recharged. The results, shown in Figure 4.32, demonstrate a significant improvement in dark counts with the new circuit. A 1V overbias results in 10,000 dark counts per second in the passively-recharged device, yet the actively-recharged device can operate with a 120 mV higher overbias with the same dark count. Moreover, the passively-recharged SPAD saturates at 1.1V while the actively-recharged one can operate up to 1.28V. The higher operating voltage resulting in a higher electric field should produce increased detection efficiency.

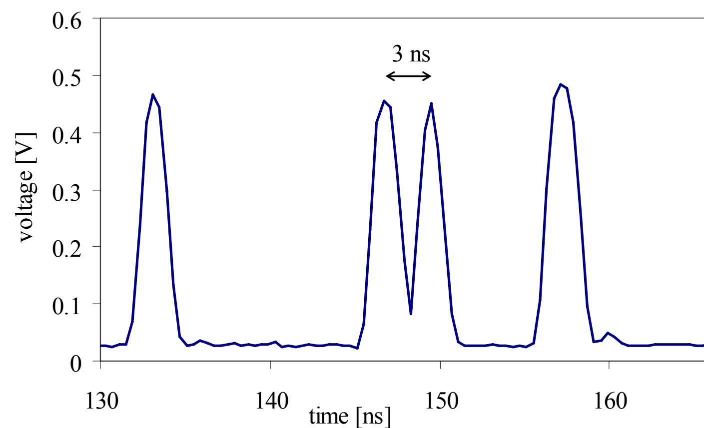


Figure 4.31: Oscilloscope image of SPAD output with 3 ns dead time.

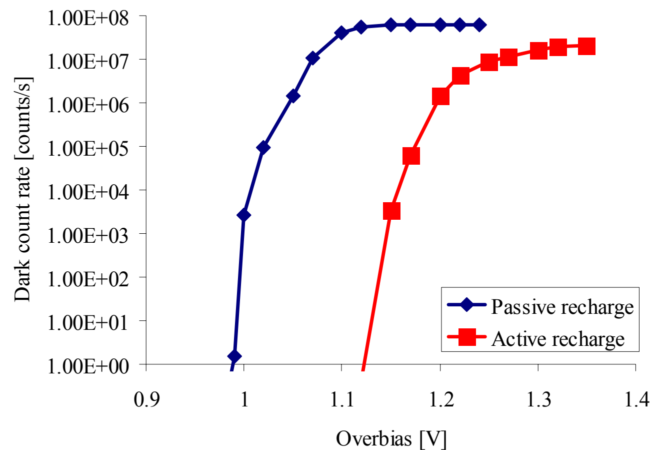


Figure 4.32: Dark count rates for identical diodes with passive and active recharging at 25C.

An auto-correlation investigation of the dark counts (Figure 4.33) reveals a vastly different correlation relationship than the passive scheme. The active-recharge circuit reduces the afterpulsing rate close to the non-correlated noise level without exhibiting the afterpulsing tail. This uniform noise density allows for device operation at higher exposure rates.

Finally, we measured the detection efficiencies of the diode with the two recharging schemes. Because different quenching transistors are used in the two schemes, it is conceivable that the difference in behavior may be due to a difference in the junction electric field for a given overbias, due to different IR drops across the quenching transistors. Therefore, it is useful to examine the detection efficiency as a function of dark counts as a measure of the effectiveness of the active-recharge circuit.

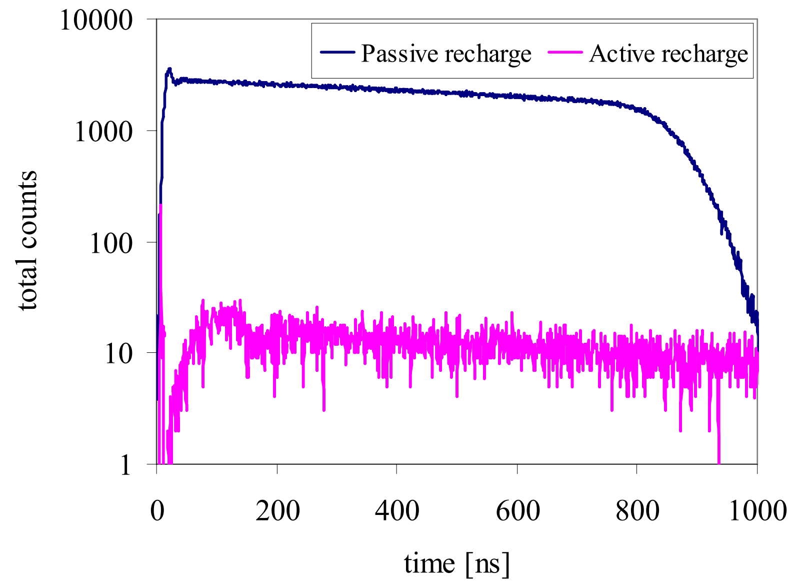


Figure 4.33: Autocorrelation curves of passively- and actively-recharged SPAD.

For this purpose, we used a Becker-Hickl BHL-600 laser to illuminate the devices at 635 nm with 50 ps pulses at a 20 MHz repetition rate. Using a calibrated New Focus 2031 photodiode as a reference, we attenuated the photon flux to 0.35 photons per pulse. The laser's electrical trigger served as the "Trigger" channel input of an MSA-1000 and the SPAD output was fed to the "Signal" channel of the counter. Histograms were collected for dark and illuminated devices with passive and active quenching under the same temperature and bias conditions.

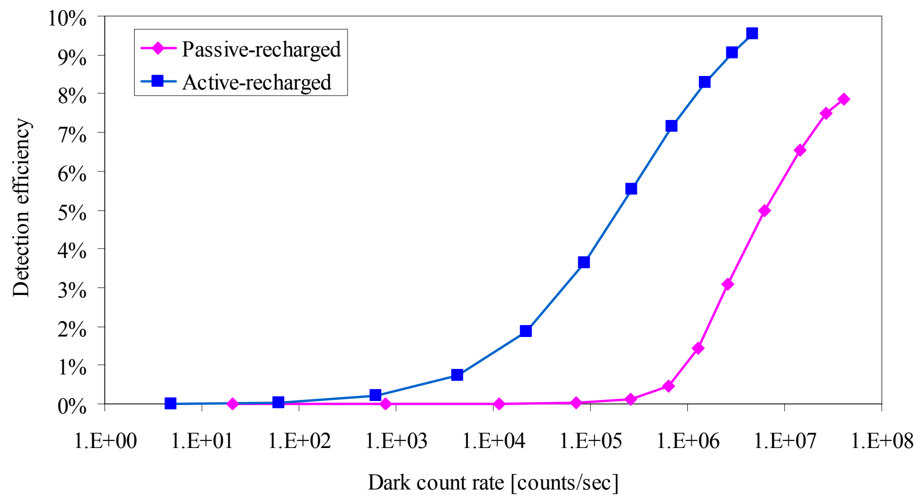


Figure 4.34: Detection efficiency versus dark count rate for a passively- and an actively-recharged SPAD with 635 nm illumination.

A plot of the detection efficiency versus the dark count rate, shown in Figure 4.34, demonstrates the efficacy of the new recharge scheme. It makes it possible to achieve identical detection efficiencies as in a traditional passive recharge scheme, with only 1-3% of the dark counts. Similarly, if we would like to operate with a maximal dark count of $1x10^6$ counts per second, after-pulse suppression using the new circuit makes it possible to improve detection efficiencies from 1% to 7.8%, with a similar improvement factor in shorter wavelengths where the device is more sensitive. At this regime, time-gated operation should be used in order to only collect signal pulses. Based on our theoretical calculations [102] and the performance of a SPAD with a similar junction depth [108], we expect our device's detection efficiency to peak at 450 nm, with a three-fold higher efficiency than at 635 nm.

4.9. Cross-Talk

In order to assess the extent of inter-pixel cross-talk, an array of six detectors was designed in proximity to each other. A central SPAD was isolated from surrounding identical SPADs using various schemes, as shown in Table 4.1.

Table 4.1. Summary of cross-talk experiments.

Structure name	Isolation Scheme	Optical cross-talk results
Top	9 μm spacing with substrate tie-down ring	None
Bottom	24 μm spacing with undoped substrate isolation	None
Right	9.6 μm spacing	None
Left	9 μm spacing with thin substrate tie-down strip	None

Cross-talk measurements were performed by simultaneously biasing the central SPAD and one of the surrounding devices. The output of the central SPAD served as the trigger of a MSA 1000 counter and the output of the surrounding SPAD was fed to the signal input. If optical cross-talk is present, we would expect to see a

correlation peak between the two devices some time after the arrival of the triggering avalanche pulse.

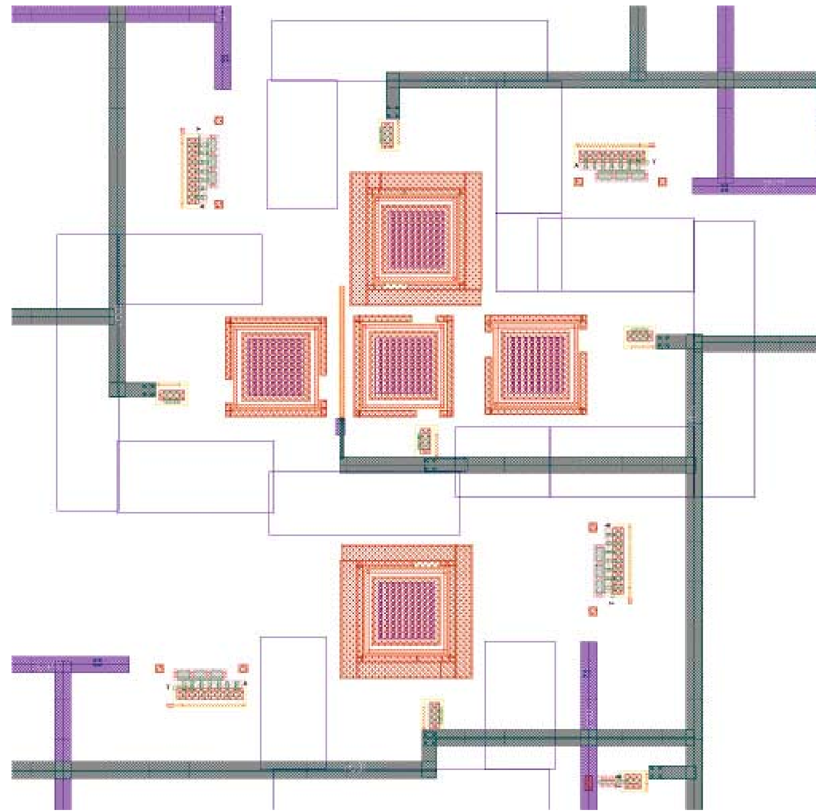


Figure 4.35: Layout of SPAD array used for cross-talk experiments.

In order to ensure that the trigger from the center SPAD arrives before the signal from the neighboring SPAD, a shorter cable was used for the triggering signal. We located the signal compared to the trigger by using a power splitter and routing the output from one of the SPADs both to the trigger and signal inputs of the counter, thus emulating a “100% crosstalk” case.

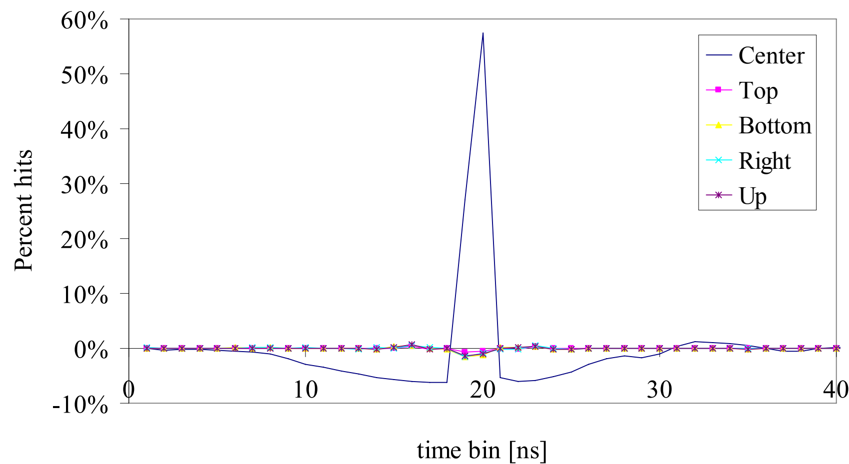


Figure 4.36: Time histogram (background subtracted) of cross-talk experiment triggered by the Center SPAD.

The results, shown in Figure 4.36 indicate a 20 ns delay due to the different cable lengths. As can be seen there are two dips, before and after the correlation peak. The former can be explained by the dead time before the arrival of the pulse. Since a trigger only occurs once an output of a certain amplitude arrives, there must be a dead time before its arrival. The dip is gradual because various avalanche amplitudes can result in a triggering pulse, corresponding to various dead times. The dip and rise following the peak are indicative of the recharge of the junction.

For the cross-talk experiments, the same cables were used, but without the power splitter. The results, shown in Figure 4.36, do not show any noticeable optical cross-talk events. A minor (<0.4%) electrical correlation can be observed, whereby there is a dip in counts correlated to the arrival of the trigger, and an increase in counts corresponding to the dead and recharge of the primary junction. This effect cannot be due to optical cross-talk because it starts before the triggering avalanche. It can be

explained by power supply issues. Specifically, when an avalanche occurs, it draws large instantaneous current from the supply, thus reducing its voltage. With a lower overbias, the probability for dark avalanches reduces, explaining the observed dip. A similar explanation relates to the smaller dips (< 0.3%) on either side of the peak.

In conclusion, we found no indication of optical cross-talk in any of the isolation schemes studies. A slight reduction in breakdown probability, which we expect to correspond to a slight decrease in detection efficiency is observed. In reality, the probability of exactly instantaneous avalanches is low, so this scenario is not troubling. Moreover, the issue can be overcome altogether by increasing the capacitance of the power supplies, thus making more charge available for the avalanches.

4.10. Performance Comparison with Diffused-Ring SPAD

In order to compare the effectiveness of the STI guard-ring, a SPAD with the traditional diffused-ring structure was included in the test chip (Figure 4.37). This device is slightly different than the traditional diffused-ring structure due to the automatic formation of a guard-ring by the fab anywhere across the wafer which is not designated as either a source-drain implant or a channel. Unlike the STI-bounded SPAD, the electric field across the STI is now much lower, because the junction edge (shown in white) is surrounded by the lightly-doped p-well. Consequently, fill factors are much lower.

In order to compare the two SPADs, we illuminated a passively-recharged SPAD and a SPAD with a diffused-ring. Both diodes had the same active area – 7 µm – and were buffered via a chain of inverters. The maximum detection efficiency was found to be at a bias of 12.05 V for the STI-bounded bias, and 12.2 V for the diffused-ring device. The setup of Figure 4.28 was used for spectral response measurements.

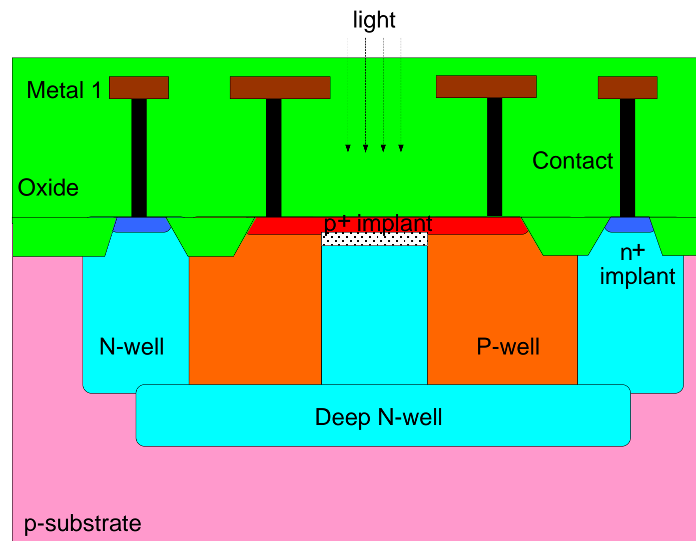


Figure 4.37: Cross-section of triple-well SAPD with diffused guard-ring.

The results, shown in

Figure **4.38**, are surprising. Since the avalanching junction is identical, the spectral response was expected to be identical. However, while qualitatively, both junctions peak at short wavelengths, the detection efficiency of the diffused SPAD is approximately twice that of the STI-bounded SPAD.

Part of this behavior may be explained by the higher bias of the diffused-ring device. But the reason for the difference in optimal operating voltages still remains to be explained. We therefore measured the dark count rate as a function of voltage, as

shown in Figure 4.39. Results show a stark difference in performance. The voltage at which dark counts start to appear is 500 mV lower in the diffused-ring SPAD. Moreover, the operating voltage range of the diffused-ring SPAD is 1100 mV, whereas the STI-bounded SPAD saturates within 100 mV.

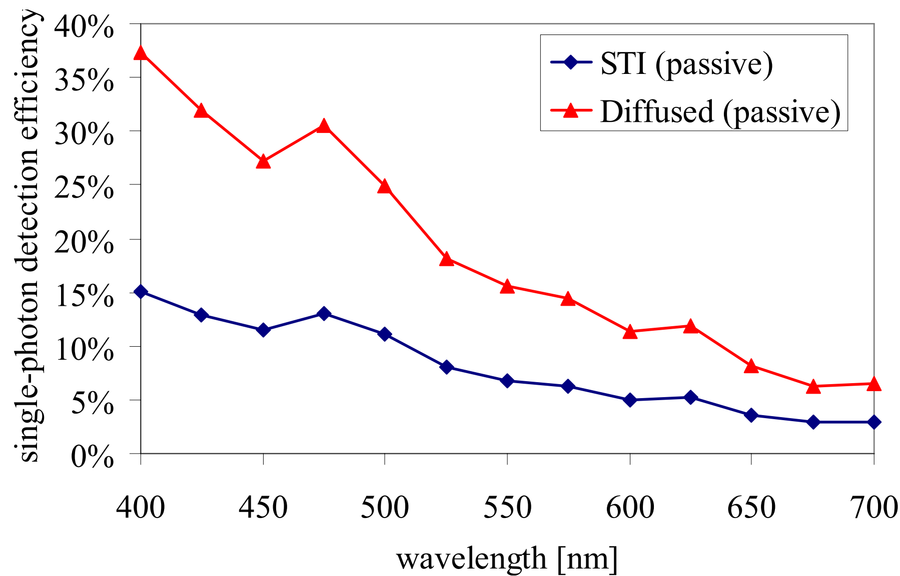


Figure 4.38: Spectral responses of passively-quenched STI-bounded and diffused-ring SPADs biased at 12.05V and 12.2V, respectively.

There are a number of possible explanations for our observations. The lower breakdown voltage may be indicative of localized premature breakdown, for example at the contact points between the edges of the N-well and the p^+ implant. This hypothesis is supported by the initial high slope of the DCR curve, which may result from an increasing high-field region, in addition to the increase in the electric field.

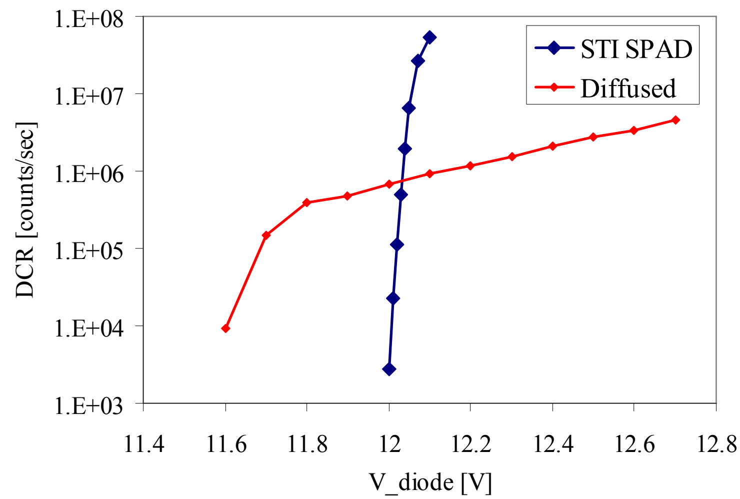


Figure 4.39: Dark count rate vs. voltage for SPADs with diffused and STI guard rings.

The difference in the voltage operating range is more perplexing. One explanation may be that a small but significant electric potential develops across the long N-well/deep-N-well/N-well path in the diffused SPAD, so that, for a given voltage, the junction field is lower than in the STI SPAD. The STI-bounded SPAD on the other hand, suffers from a high dark count rate, and as a result seldom fully recharges (Figure 4.40). This translates to lower detection efficiencies compared to an identical overbias of the diffused-ring device (Figure 4.41).

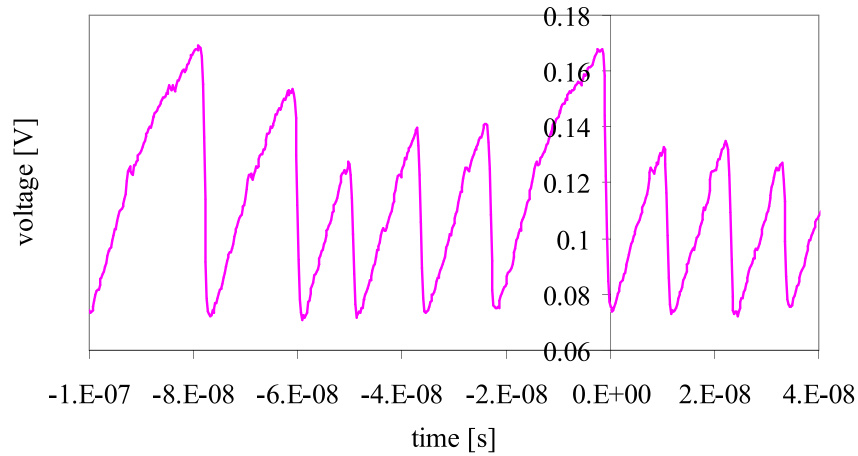


Figure 4.40: Scope waveform of saturated STI-bounded SPAD with source-follower output

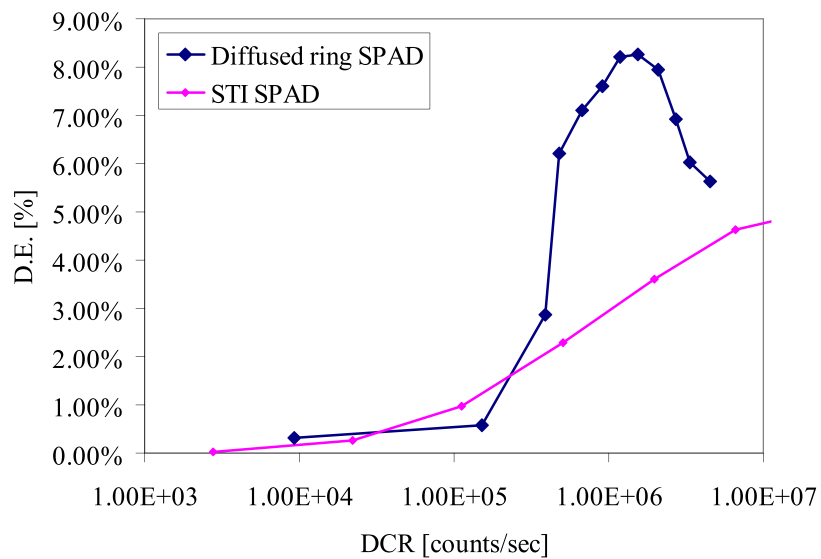


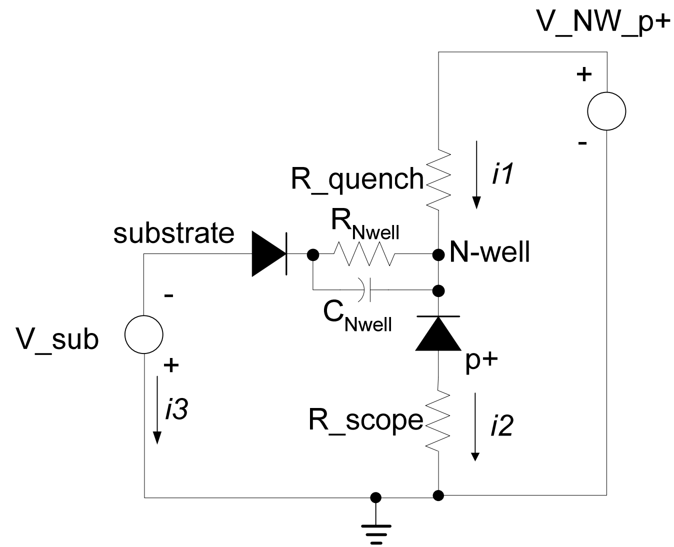
Figure 4.41: Detection efficiency vs. DCR for STI-bounded and diffused-ring SPADs.

4.11. Dual-Color Single-Photon Detection

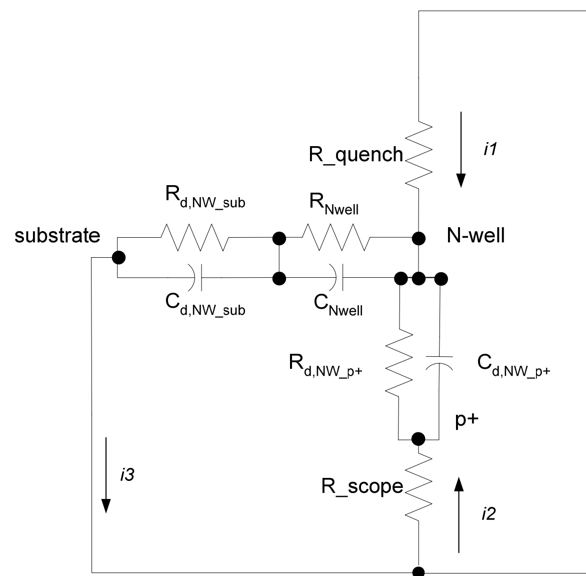
In order to prove the feasibility of the dual-color detection concept, we developed the setup shown in Figure 4.42 (a). The dual-junction SPAD is represented by two back-to-back diodes, with a distributed N-well node, separated by the resistance and capacitance of the N-well layer, R_{Nwell} and C_{Nwell}, calculated to be 400 Ω. Quenching and recharge for both diodes is performed by a single quenching resistor, $R_{_quench}$, tied to the *N-well* node. The *N-well* node is biased with a positive voltage, *V_pos* and the substrate is biased relative to p^+ by *V_sub*. The output signal is observed through an ammeter. In our case, the digital scope is operated in current mode with *R_scope* = 50 Ω input impedance.

Each of the diodes can be modeled using its junction resistance and capacitance (Figure 4.42 (b)). This resistance is very high when the device is in reverse-biased in sub-Geiger and is very low when avalanching.

When the device is biased such that only the shallow, p^+/N-well junction is in Geiger node, and the latter breaks down, the scope measures a positive current. When the device is biased such that only the deep, substrate/N-well is in Geiger mode, we have, due to conservation of current, $i3 = i1 + i2$, i.e., $i0 = 0$, and current flows in the direction shown in Figure 4.42 (b), i.e., in opposite polarity to the previous case. Because some of the avalanche current *i3*, flows to *i1*, we expect the amplitude of the observed spikes when the deep junction breaks down to be smaller than when the



(a)



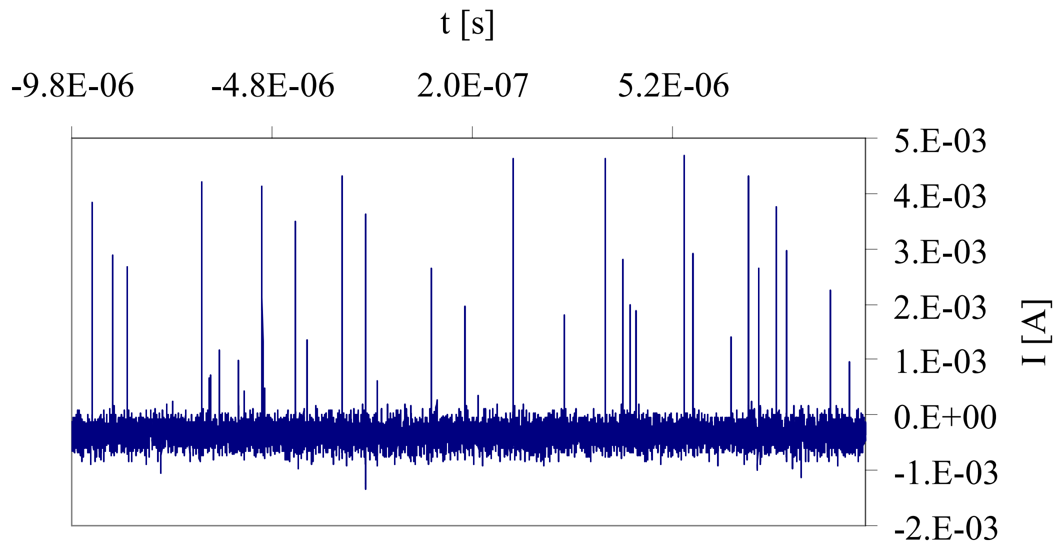
(b)

Figure 4.42: (a) Experimental setup for characterizing externally-quenched dual-junction SPAD and (b) small-signal equivalent circuit when only the deep junction breaks down.

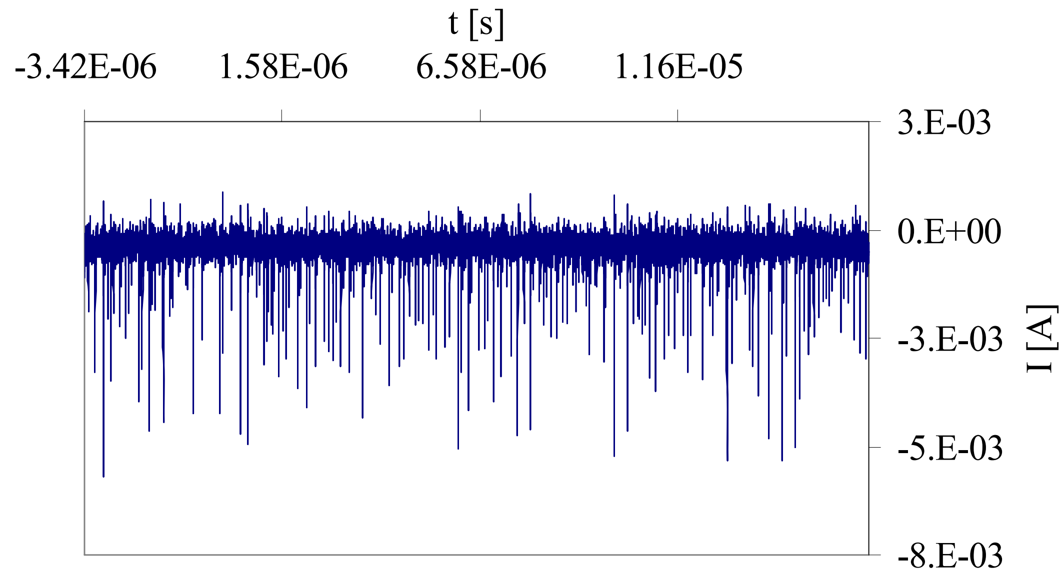
shallow junction breaks down. When both junctions are biased above breakdown, unique signals can be obtained from each breakdown.

The device was biased when not illuminated in order to test its operation with dark pulses. Figure 4.44 illustrates the various modes of operation. When only the shallow junction is biased above breakdown, positive pulses are measured by the ammeter-scope. When only the deep junction is biased above breakdown, negative pulses result. When both junctions are biased above breakdown, both positive and negative pulses are observed. Reliable operation of the non-planar junction over millions of cycles proved the feasibility of operation using these non-planar structures,

A calculation of the cross-correlation between the positive and negative dark pulses was performed using a Labview program, which captured scope waveforms, and an Excel spreadsheet, which calculated the cross-correlation function. The results (Figure 4.45) provide important insights into the behavior of the dual-junction device. The first observation is that the probability for cross-talk is close to unity. This can be observed from the peak at delta-time of 0. In addition, a dead time and an afterpulsing tail of the shallow junction can be seen at the right half of the cross-correlation curve. Interestingly, the left side of the graph corresponds to charges generated by the shallow junction causing an avalanche in the deep junction – this explains the apparent a-priori knowledge exhibited by negative-time correlation.

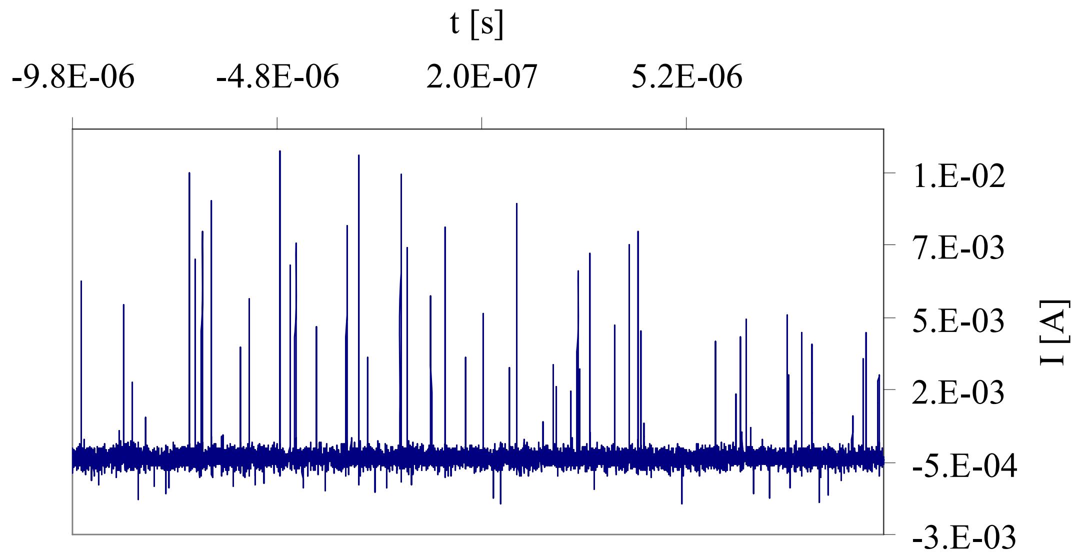


(a)



(b)

Figure 4.43: Scope waveforms for externally-quenched dual-junction SPAD. Biases are (a) V_{deep} = -9.6V, $V_{shallow}$ = -13.6V; (b) V_{deep} = -13.6V, $V_{shallow}$=0V



(c)

Figure 4.44: Scope waveforms for externally-quenched dual-junction SPAD. (c) V_{deep}=-11.6V, $V_{shallow}$=-13.6V.

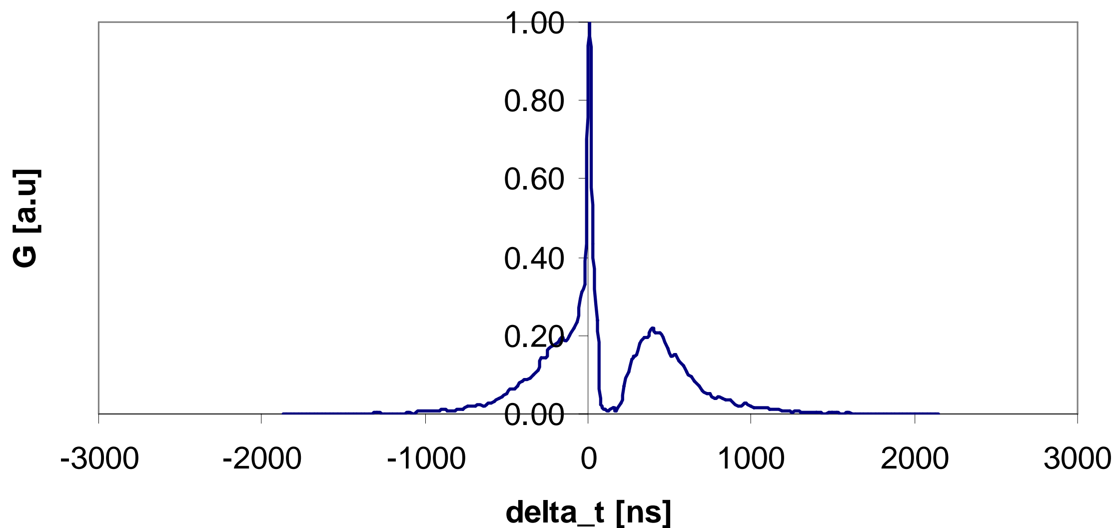


Figure 4.45: Cross-correlation between shallow- and deep-junction dark pulses.

The causes of the observed cross-talk may either be the diffusion of injected charges from one junction's avalanche to the other junction (electrical cross-talk), or

the absorption of luminescent photons emitted by one avalanching junction and absorbed by the other (optical cross-talk). The implication of this cross-talk is that instantaneous operation of both junctions is not feasible. On the other hand, because deep avalanches result virtually always in shallow avalanches, this can serve as a readout mechanism for the deep-junction avalanches, although, some timing information will be lost.

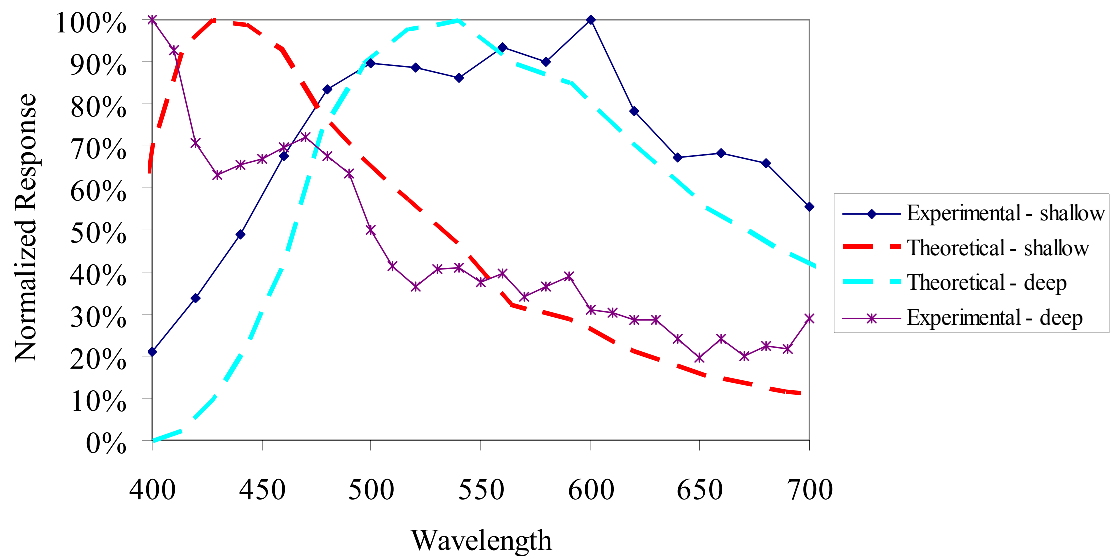


Figure 4.46: Spectral response of shallow and deep junctions.

We measured the spectral response of the two junctions using the setup of Figure 4.28 on the first test chip. Because of the fill patterns, we could only obtain relative spectral responses. These are plotted in Figure 4.46, and agree well with the expected response. As can be seen, spectral cross-talk can be minimized by using probes that emit at very short and very long wavelengths.

Future work is warranted on this device, both in terms of optimization of the junction depth to reduce optical cross-talk, and in terms of the required processing circuitry.

4.12. Conclusions

In this chapter, we summarized the characterization results of the new STI-bounded SPAD. We demonstrated its reliable operation over a long period of time, without any DC-level shifts, indicating that oxide charge-up is not an issue with the device. By measuring the breakdown voltage of the device, and comparing it both to devices manufactured with curved junctions, and to the computed breakdown value, we showed that the bounded junction is indeed planar.

Avalanche behavior was studied on two generations of devices – an externally-quenched and an integrated-quenched SPAD. Avalanche charge followed well the voltage and capacitance dependences developed in chapter 2. We demonstrated several new techniques for measuring the dead time of the device, and showed that by utilizing the efficient STI guard ring, dead times were reduced to 3 ns – shorter than any SPAD to date.

Detection efficiencies were measured using several techniques, the most efficient of which used time-gating and a fast counter. Detection efficiencies peaked, as expected by theory, at the ultraviolet, though their value was lower than expected. This was explained by the high dark current, dominated by self-sustaining

afterpulsing, which does not allow the junction to fully recharge. The source of the noise was deduced by looking at the autocorrelation of the noise signal. We hypothesized that the excessive afterpulsing resulted from the very short recharging of the junction. In order to reduce this dark current without slowing the detector, we characterized the new compact active-recharge circuit, and showed that, as expected, we were able to reduce the noise for a given detection efficiency, compared with the passive-recharging scheme. An investigation of several inter-pixel isolation schemes concluded that, probably due to the low avalanche current, cross-talk is not observed.

A comparison of the STI-SPAD with devices based on the traditional diffused-ring structure showed some of the trade-offs in the choice of the STI ring. The diffused-ring device exhibited a considerably lower noise and a wider operating-voltage range. Due to its lower noise, higher detection efficiencies were achieved, up to 37%. A possible explanation for these observations is the larger parasitic capacitance which results in longer recharge times. This capacitance also results in a 10x reduction in fill factor in the diffused-ring device.

Finally, we summarized some preliminary characterization results relating to dual-color operation of the SPAD. We showed that it is possible to obtain unique outputs from the deep and shallow junctions, and that the spectral response of the junctions is in accordance with theory. Cross-correlation measurements demonstrated a high cross-talk between the junction, which, we argue, can be used for a simple readout of the deep junction, when the two junctions are biased alternately.

The results presented here indicate that the STI-bounded SPADs are advantageous especially in applications requiring large arrays, or when high data rates are beneficial. Their weakness is in their lower detection efficiencies, shorter-wavelength sensitivity and higher noise. Another expected advantage of the devices in superior jitter performance – because of the STI isolation, the jitter tail characteristic of SPADs with a diffused ring is expected to disappear, resulting in much improved timing precision. This behavior was not characterized in this work because only packages with long wire bonds were available, resulting in large inductance. The applications that stand to benefit are FLIM, where high-repetition-rate images are advantageous, and quantum key distribution, where time gating can remove the effects of the excessive noise, and where the excellent jitter performance can further reduce gate widths.

Acknowledgement:

This chapter, in part, is a reprint of the material as it appears in the following publications:

H. Finkelstein, M. J. Hsu, S. Esener, “STI-bounded single-photon avalanche diode in a deep-submicron CMOS technology”, *IEEE Electron Device Letters*, vol. 27, no. 11, 2006.

H. Finkelstein, M. J. Hsu, S. Esener, “An ultra-fast Geiger-mode single photon avalanche diode in 0.18 μm CMOS technology”, *Proceedings of Advanced Photon Counting Techniques, SPIE Vol. 6372*, Boston, MA, 2006.

H. Finkelstein, M. J. Hsu, S. Esener, “A compact single-photon avalanche diode in a deep-submicron CMOS technology”, *Proceedings of International Conference of Solid-State Devices and Materials*, Yokohama, Japan, 2006.

The dissertation author was the primary investigator and first author of this paper.

5. SINGLE-PHOTON FREQUENCY UPCONVERSION VIA HOT-CARRIER ELECTROLUMINESCENCE

5.1. Introduction

Single-photon detection at infrared wavelengths has gained relevance in recent years due to its central role in quantum communications [145], eye-safe laser detection and ranging (LIDAR) [146], optical time-domain reflectometry (OTDR) [147] and in semiconductor failure analysis [148]. IR single-photon detectors should ideally operate at high frequencies (tens to hundreds of MHz), consume minimal power (< 1 nW/bit), operate reliably at non-cryogenic temperatures over many cycles and be manufacturable at a low cost. When operated in arrays, such devices should also have a small pitch and low pixel-to-pixel cross-talk.

Several figures of merit are used to evaluate single-photon detectors. The *single-photon detection probability* is the product of the probabilities of a photon being absorbed in the material and of it initiating a detectable avalanche. The device's *spectral response* describes the wavelength-dependence of this detection probability. These metrics depend on the percentage of pixel area which collects photons (fill ratio); on the absorbing layer's composition, depth and thickness; and on the electric field distribution in the multiplication region, as will be detailed in Section III.

During the recharge process following an avalanche, the SPAD is temporarily biased below its breakdown voltage, and cannot generate an avalanche pulse in response to a photon. This time is called the *device dead time* and depends on the recharge mechanism, on the overbias above breakdown and, most significantly, on the junction's capacitance.

Dark counts result from avalanches which are not induced by absorbed photons. They can originate from thermally-generated carriers; from band-to-band tunneling; via trap-assisted tunneling; and by afterpulsing – the release of carriers trapped in prior avalanches. The latter mechanism is an important factor in determining the device dead time.

The time-delay spread between the photon absorption event and the clocking of the resulting electrical signal depends on the diameter of the SPAD as well as on the timing circuitry. It determines the *timing resolution* of the single-photon detector. Other factors such as active area, pixel pitch and manufacturing cost are also important in the evaluation of single-photon detectors.

Visible wavelength single-photon detection has been extensively investigated. Silicon SPADs have recently been demonstrated on commercial deep submicron technologies, allowing for compact pixels with 25% fill factor, with less than 5 ns of dead time and with digital outputs [103]. Such devices can be integrated into arrays with the quenching, recharge and processing circuitry on the same die.

Various techniques have been employed for single-photon detection in the IR. Superconducting transition-edge sensors detect the subtle temperature change in

small-volume tungsten microcalorimeters in response to the absorption of a photon. [3]. These devices offer excellent detection efficiencies, a very low dark current, and no afterpulsing. However, their dead time is excessive (a few microseconds) and they require cooling to 4.2 K in addition to very-low-noise amplifiers which must also be cooled. Superconducting niobium [149] and niobium-nitride [43] nanowires have also been used for single-photon detection, achieving gigahertz operation, but still requiring cryogenic cooling. Furthermore, production of such devices is expensive and has not been demonstrated in large arrays or in mass scale.

Solid-state IR SPADs have been demonstrated using a planar geometry and a standard semiconductor processing flow [150]. These detectors traditionally have separate absorption and multiplication regions, whereby, for example, photons are absorbed in a thick (several microns) lightly-doped InGaAs layer. The photo-generated carriers are swept towards an InP high-field multiplication region where impact ionizations provide gain. When the extraction rate of carriers from this multiplication region falls below the creation rate, an avalanche breakdown is said to occur and the gain becomes “infinite”. In this mode, known as Geiger Mode (GM), single-photon detection becomes possible. The avalanche must be quenched to avoid damage to the junction.

Avalanche quenching can be achieved either passively, by using a voltage-limiting resistor in series with the device, or actively, using a circuit which senses the onset of the avalanche, and subsequently quenches it. Once the avalanche has been

quenched, the diode capacitance must be re-charged, either passively, through the quenching resistor, or actively, using a recharging circuit.

IR Geiger-mode single photon avalanche diodes (GM-SPADs) have been shown to have detection probabilities on the order of 33% when operated 5% above their breakdown voltage [151]. They are amenable to integration in large arrays and to mass production because they can be manufactured using standard lithographically-defined processing techniques. However, the support circuitry, including the quenching, recharging and processing circuitry must be implemented externally, usually in silicon [152]. Recently, one of the authors demonstrated an InP/InGaAs Metal-Oxide-Semiconductor SPAD, which integrates the quenching function into the device [153]. This is done by depositing a resistive layer on top of the pn junction. When an avalanche occurs, charges accumulate on the interface between this resistive layer and the multiplication layer, thereby creating a negative feedback loop which quickly quenches the avalanche. Because no external quenching is required, this new technique is expected to significantly reduce the junction capacitance, and consequently, the recharging time and the power dissipated by the device.

The main deficiency of GM-SPADs lies with their excessive noise, which originates from four sources [12]. Hole-electron pairs, thermally generated at the edge of the high field region through Shockley-Read-Hall generation and separated by the strong electric field, can cause a “false” avalanche [85]. Trap-assisted tunneling depends on the defect density as well as on the doping concentration and may be exacerbated at high electric fields by barrier lowering via the Poole-Frenkel effect [12,

14]. Direct band-to-band tunneling requires strong electric fields above $7x10^5$ V/cm and occurs in devices with a breakdown voltage lower than $4E_G/q$, where E_G is the bandgap energy and q is the electron charge [13, 14]. Finally, afterpulsing which results from the release of charges trapped during previous SPAD cycles, increases with high defect densities and is linearly dependent on the total charge flowing during an avalanche [12]. The rate of emission from these deep traps follows an exponentially decaying distribution which depends on the activation energy of each deep trap mechanism.

Whereas the first three mechanisms can be reduced by cooling the device, deep trap lifetimes increase exponentially as temperature is decreased. Because thermal generation and tunneling increase with narrower bandgap, it is necessary to cool IR SPAD, usually to about 200˚K. At these temperatures, afterpulsing becomes the dominant noise source with rates on the order of tens of kHz and it becomes the main bottleneck for device bandwidth [7, 8]. Time-gating can reduce the effects of afterpulsing [9, 15]. However, due to the exponential time distribution of afterpulses, the separation between gates must be made long compared with the afterpulse lifetime, so that the probability of experiencing an afterpulse during an exposure time gate is acceptably low. Furthermore, time gating is only possible when the arrival time of the photon is known to within the duration of the gate. At low temperatures, SPADs can still be operated in free-running mode with minimal afterpulsing effects, but only if a sufficiently long hold-off time is ensured following an avalanche, thereby severely limiting the detection rate [98].

Afterpulsing can be reduced by limiting the charge flowing during an avalanche. This may be done by active quenching but is achieved more efficiently by reducing the junction capacitance. The capacitance in IR SPADs is dominated by the capacitances of the readout, recharge and quenching circuitry, because these functions are implemented off-chip, either on a board or on a silicon die. Connections are made either by wire bonding [150] or by using indium bumps [16]. This limits the pixel pitch and may result in capacitances on the order of pF, with a resulting deterioration in the noise performance of the device due to afterpulsing.

Recently, a scheme was described for integrating the various SPAD components using 3-wafer direct bonding with cross-visa [17]. While this technique holds much promise for reducing the junction node's capacitance, it requires complex and expensive processing and has yet to be demonstrated for large arrays.

Another recent report demonstrated a novel method for detecting and reading-out the signal from a single IR photon, while simultaneously reducing jitter and afterpulsing [18]. A continuous wave laser is used to seed a periodically-poled lithium-neonate waveguide. Impinging IR photons are converted to visible wavelength and are subsequently detected by a low-jitter silicon SPAD. Optimal performance has been achieved for a single channel with a pump power of approximately 300 mW and an overall conversion efficiency of 5-7%. For large detector arrays and for power-sensitive applications improved performance is desirable. Furthermore, the requirement for the impinging photons to have a specific polarization with respect to the optical axis of the detection setup is often undesirable.

In this chapter we propose and model a new interconnection and readout scheme which does not require any electrical interconnection between the IR SPAD and the CMOS readout circuitry, offers superior upconversion efficiencies with low power, does not impose any requirements on the polarization of impinging photons, and is scalable to large arrays. By bypassing the requirement for electrical bonding between the SPAD pixels and the readout circuitry, the capacitance seen by the junction is significantly reduced. As discussed above, this results in a reduction in afterpulsing, which is the dominant noise source at low temperatures, while simulataneously decreasing the device dead time. The proposed method also promises to greatly simplify the manufacturing of integrated IR-SPAD devices and to improve their fill factor.

The proposed scheme is based on wavelength upconversion using a byproduct of the avalanche process, namely hot-carrier luminescence from the multiplication layer of the device. This luminescence is shown to have a significant component at higher energies than the bandgap of the absorbing material, and can therefore be extended to many IR detection materials. Readout of the luminescent photons is achieved by a coupled silicon single-photon avalanche diode.

One of the main advantages of the proposed device lies in its simple manufacturing flow. Heterogeneous integration severely limits the combinations of materials which can be directly interfaced. Moreover, electrical connections from III-V devices require additional masking layers and reduce the pixel pitch to tens of microns [154] – unacceptably high for large arrays. Here, we propose to utilize a

mature wafer-level glass-to-glass fusing technology [155] in order to connect between a III-V SPAD which detects the primary IR photon, and a silicon CMOS SPAD, which detects the up-converted photons, and which processes the information on the same die.

Section 5.2 of this paper reviews the physics of hot-carrier luminescence in avalanche photodiodes. Section 5.3 describes the model of the optical readout and derives an expression for the upconversion efficiency as well as the power consumption per avalanche. Section 5.3.6 provides a numerical calculation of upconversion efficiency and power consumption, and compares the proposed scheme's performance with traditional readout setups.

5.2. Hot-Carrier Luminescence in Avalanche Photodiodes

Hot-carrier luminescence has been extensively studied since the 1950's [132, 156]. It is believed to be the primary mechanism responsible for avalanche spreading [126] and has traditionally been viewed as a detrimental side-product of the avalanche, resulting in optical cross-talk [152] and a potential source for eavesdropping in quantum communication channels [63]. The same effect is being widely used to investigate defects in switching integrated circuits [148].

Hot-carrier luminescence most likely results from recombination of hot electrons with holes (Figure 5.1) [131, 157]. In a *direct* recombination process, a photon is emitted whose energy equals the difference between the hot electron's initial

energy and the bandgap. During a phonon-assisted *indirect* recombination process, both energy and momentum must be exchanged, and the probability of such events is considerably lower, depending on the availability of suitable phonons. This is manifested by the different electron temperatures in these processes. In GaAs, the direct recombination process is characterized by an electron temperature of 800˚K while the indirect process in the same material exhibits a temperature of 3000˚K [157]. The higher temperature stems from a longer mean lifetime, consistent with a less probable recombination event.

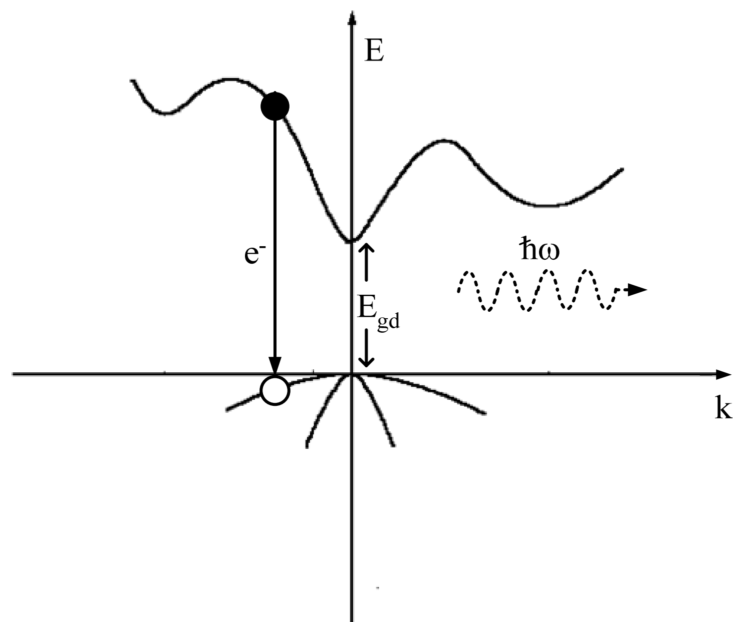


Figure 5.1: InP energy band diagram illustrating hot-carrier luminescence in a direct recombination processes. A hot electron accelerated by the strong electric field recombines with a hole at the valence band. The excess energy is released in the form of a photon with energy ħω>Egd. A low-energy, infrared, component due to transitions between the light and heavy hole bands has also been observed.

For a direct bandgap material, such as GaAs or InP, the photon emission rate under parabolic band approximation is given by:

$$R_d(\hbar\omega) \propto \hbar\omega\left(\hbar\omega - E_{gd}\right)^{1/2} f(E)\left[1 - f(E - \hbar\omega)\right]$$

Equation 5.1

where $\hbar\omega$ is the emitted photons' energy, E_{gd} is the direct bandgap; E is the electron energy above the bottom of the conduction band; $f(E)$ and $\left[1 - f(E - \hbar\omega)\right]$ are the hot electron and the hole distributions, respectively, both of which strongly depend on the hot-carrier temperature [157]. The emission spectrum of InP was calculated using Equation 5.1, and is shown in Figure 5.2. Because the photons carry the excess energy after the recombination, $\hbar\omega > E_{gd}$ thereby achieving energy upconversion.

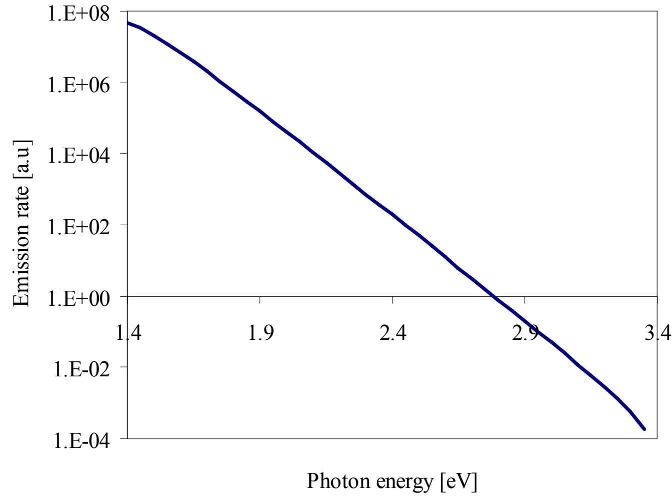


Figure 5.2: Calculated spectrum of electroluminescent photons emitted from an InP pn junction with a doping of 2.5 x 10^{18} cm^{-3}. A bandgap of 1.35 eV [126] and an ionization threshold of 2.05 eV [158] were used for the calculations.

The efficiency of the upconversion process strongly depends on the electroluminescence yield, i.e., the photon emission rate per unit avalanche charge. This figure is quite difficult to measure due to self-absorption by the emitting device, the detector's spectral response, the effect of defects, and collection uncertainties due to reflections. For a silicon pn junction, Kurtsiefer [158] reported a figure of 39 photons per steradian in an avalanche with 4×10^{8} electrons, resulting in a lower limit of 2.5×10^{-6} photons per electron after partially accounting for the detector's sensitivity but not for self-absorption or for total internal reflections at the semiconductor-air interface. A measurement accounting for both the optical system and self-absorption in silicon was presented by Lacaita [159]. She reported an emission efficiency of 2.9×10^{-5} photons with energy higher than 1.14 eV per carrier crossing the junction. Electroluminescence yield for InP has not been reported to date. As outlined above, it is expected to be significantly higher than that of silicon and is conservatively estimated to be 2.9×10^{-4} photons per hot-carrier for the purposes of our calculation.

5.3. Model for Upconversion Efficiency

5.3.1. Upconversion Model

The model proposed in this work is based on the hot-carrier luminescence effect described above. Figure 5.3 illustrates a device based on this concept, comprised of an

InGaAs/InP SPAD fused to a silicon CMOS SPAD. A primary IR photon is absorbed in the narrow-bandgap InGaAs layer, and the photogenerated charges are swept to the high-field multiplication region. During avalanche multiplication, secondary photons are emitted from the multiplication layers and are detected by the silicon SPAD, and processed on the same die. This upconverting hybrid pixel can be scaled to large arrays for parallel operation, for example in single-photon NIR imaging applications. A metal masking layer is used to minimize inter-pixel cross-talk.

We wish to investigate whether a sufficient number of secondary photons are emitted per primary avalanche such that they can be detected with a high probability. Furthermore, we would like to assess the bandwidth and power consumption of the hybrid scheme.

In order to calculate the overall detection probability of the up-conversion scheme, we need to multiply the primary NIR detection probability in the InGaAs/InP SPAD by the emission probabilities (as a function of wavelength) of the electroluminescent photons emitted towards the silicon SPAD. We should then account for self-absorption in the InP device. Finally, we need to determine the probability of absorption of these photons in the silicon SPAD's depletion region, and calculate the avalanche initiation probability of the photogenerated charge carriers. Mathematically, the upconversion probability can be expressed as:

$$\eta_{uc} = \int_{E_{gs}}^{E_{\max}} \int_{x_{j2}}^{x_{j2}+x_d} \frac{\Omega}{4\pi} N_{sp}\left(\hbar\omega\right)\left[1 - P_{sa}\left(\hbar\omega, x_{j1}\right)\right] \times P_{abs}\left(\hbar\omega, x_{abs}\right) P_{av}\left(x_{abs}\right) d\left(\hbar\omega\right) dx_{abs}$$

Equation 5.2

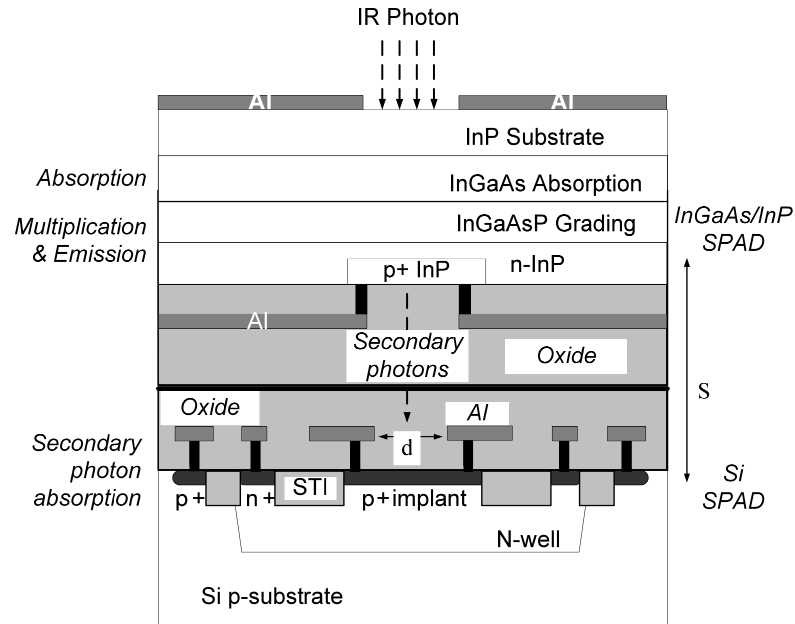


Figure 5.3 Cross-section of the proposed device: An InGaAs SPAD is direct-bonded to a Si SPAD for optical readout. IR photons are incident on the back surface of the InGaAs SPAD (e.g., [150, 151]). As carriers recombine during the avalanche, they release visible photons which are detected by the silicon device (e.g., [103]). The devices are fused through their silicon-dioxide passivation layer, eliminating the need for lattice matching between the two semiconducting materials.

where Ω is the solid angle subtended by the silicon junction when observed from the InP junction; N_{sp} is the number of electroluminescent photons at energies, ħω, emitted in a primary avalanche; $P_{sa}\left(\hbar\omega, x_{j1}\right)$ is the probability of self-absorption of the secondary photons, which is a function of their energy and generation depth x_{j1}; P_{abs} is the absorption, probability in the silicon SPAD's depletion region, which extends from x_{j2} to $x_{j2}+w_d$; and $P_{av}(x_{abs})$ is the probability for an electron-hole pair photo-

generated at x_{abs} to induce a detectable avalanche. Photons in the silicon absorption band - from its bandgap E_{gs} to E_{max}, where the electroluminescent photon yield diminishes to zero – must be considered.

In the following subsections, we will develop expressions for these parameters that will allow us to compute the overall up-conversion efficiency of the device.

5.3.2. Secondary Photon Emission towards the Silicon Junction

An avalanche event in a SPAD can be viewed as a discharge of the junction capacitance, C_j, from an initial voltage, in excess of the diode's breakdown voltage to approximately the breakdown voltage. The total number of electrons flowing during an avalanche is:

$$N_e = \frac{1}{q}(C_j + C_p)V_{ob}$$

Equation 5.3

where q is the electron charge (in Coulomb) C_j is the junction capacitance, C_p is any additional capacitance seen by the junction, including interconnection and sensing capacitances and V_{ob} is the overbias above breakdown.

The number of secondary photons emitted from the primary junction is:

$$N_{sp}(\hbar\omega) = \eta_e s(\hbar\omega) N_e$$

Equation 5.4

where η_e is the luminescence yield per electron (in relevant energies for Si absorption) and $s(\hbar\omega)$ is the normalized spectral distribution of the secondary photons, shown in Figure 5.2. In reality, additional photons are expected to be emitted from the grading region between the absorption and multiplication regions, where the high electric field is lower than the breakdown field, and thus recombination events are highly likely. In this analysis we conservatively disregard these photons, and assume all secondary photons are emitted from the maximum field region, at the junction plane.

Assuming an isotropic emission from the junction plane, only a fraction of the emitted photons, $\frac{\Omega}{4\pi}$, is actually transmitted towards the silicon junction. The solid angle Ω for the case of a planar InP junction emitting towards a parallel planar silicon junction can be approximated by assuming all photons are emitted from one of the vertices of the InP rectangular junction (Figure 5.4). This will provide a lower bound on the actual flux emitted towards the silicon SPAD. The solid angle subtended by the detector from this vertex is given by [160]:

$$\Omega = \tan^{-1}\frac{d^2}{S\sqrt{2d^2+S^2}}$$

Equation 5.5

where *d* is the side dimension of the Si junction and *S* is the vertical distance between the junctions.

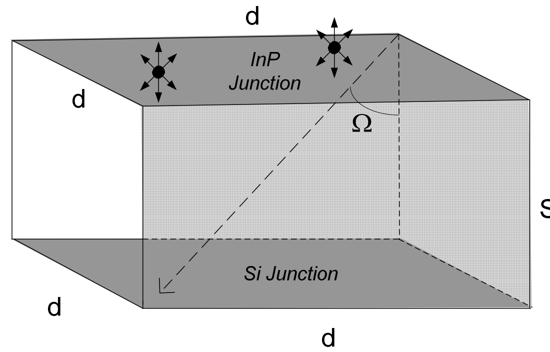


Figure 5.4: Geometrical construction for calculating the percentage of photons emitted from an InP junction plane onto a Si junction plane. A conservative approximation places all avalanching charges at one of the vertices of the emitting plane.

Equation 5.3, Equation 5.4 and Equation 5.5 can be combined to give:

$$\frac{\Omega}{4\pi}\mathrm{N_{sp}}\left(\lambda\right)=\frac{1}{4\pi \mathrm{q}}\tan^{-1}\frac{d^2}{S\sqrt{2d^2+S^2}}\eta_{\mathrm{e}}s\left(\lambda\right)(\mathrm{C_j}+\mathrm{C_p})\mathrm{V_{ob}}$$

Equation 5.6

5.3.3. Secondary Photons Self-Absorption Probability

Self-absorption in InP reduces the number of photons which reach the surface. For simplicity, we assume all luminescence occurs at the junction plane, a distance x_{j1} from the surface. The emitted photon population will be:

$$\mathrm{N_{surf}}\left(\hbar\omega\right)=\frac{\Omega}{4\pi}\mathrm{N_{sp}}\left(\hbar\omega\right)\left[1-P_{sa}\left(\hbar\omega,x_{j1}\right)\right]=\frac{\left[\tan^{-1}\frac{d^2}{S\sqrt{2d^2+S^2}}\eta_{\mathrm{e}}s\left(\hbar\omega\right)(\mathrm{C_j}+\mathrm{C_p})\mathrm{V_{ob}}\right]\exp\left(-\alpha_{InP}\left(\hbar\omega\right)x_{j1}\right)}{4\pi \mathrm{q}}$$

Equation 5.7

where $\alpha_{InP}(\hbar\omega)$ is the absorption coefficient in InP.

5.3.4. Secondary Photons Absorption Probability in Silicon

Having calculated the spectral distribution of the emitted photons, we can estimate the probability for these photons to be absorbed by the Si SPAD, as well as their probability of generating an avalanche. Because the two devices are directly fused at the silicon dioxide layer, we can safely assume that reflections at the interfaces do not substantially affect the number and spectral distribution of secondary photons.

We have shown that in an STI-bounded shallow junction the high field region is highly localized in the depletion region of the junction [103], so we can assume all absorption occurs within the depletion region. The probability for N photons of energy $\hbar\omega$ to generate at least one electron-hole pair within this layer is:

$$\mathrm{P}_{\mathrm{abs}}^{N}(\hbar\omega) = 1 - \left\{1 - \left(\exp\left[-\alpha_{Si}(\hbar\omega)\cdot w_d\right] - \exp\left[-\alpha_{Si}(\hbar\omega)\cdot\left(x_{j2} + w_d\right)\right]\right)\right\}^{\mathrm{N_{surf}}(\hbar\omega)}$$

Equation 5.8

with α_{Si} being the absorption coefficient in silicon, w_d the depletion width and x_{j2} the junction depth in the silicon device. The depletion width of the junction can be determined from the analytical expression for a one-sided linearly-graded junction (e.g., as given by [85]):

$$w_d = \left(\frac{3V_B \varepsilon_s}{2qa} \right)^{1/3}$$

Equation 5.9

where V_B is the sum of applied and built-in voltages, ε_s is the dielectric constant of silicon, q the electron charge and a the grading coefficient of the linearly-graded junction.

The total upconverted photons' absorption probability in silicon can now be calculated using Equation 5.7 - Equation 5.9 over all relevant wavelengths.

5.3.5. Secondary Avalanche Initiation Probability

Next, we need to calculate the probability that an absorbed photon will induce an avalanche. For a one-sided, linearly-graded pn junction, Poisson's equation translates to a field distribution:

$$E(z) = \frac{qa}{2\varepsilon_s}\left(w_d^{\,2} - z^2\right)$$

Equation 5.10

We approximate the avalanche probability as a function of the position of generation of the electron-hole pair by solving the coupled differential equations [134]:

$$\frac{dP_{be}}{dz} = \left(1 - P_{be}\right)\alpha P_{bp}$$

(a)

$$\frac{dP_{bh}}{dz} = \left(1 - P_{bh}\right)\beta P_{bp}$$

(b)

Equation 5.11

where P_{be} and P_{bh} are the avalanche initiation probability by an electron and a hole, respectively, α and β are the ionization rates of electrons and holes, respectively, and P_{bp} is the joint avalanche initiation probability:

$$P_{bp} = 1 - \left(1 - P_{be}\right)\left(1 - P_{bh}\right) = P_{be} + P_{bh} - P_{be}P_{bh} .$$

Equation 5.12

These equations can be solved numerically to provide the avalanching probability (Figure 5.5).

5.3.6. Numerical Calculations and Design Considerations

In order to test the feasibility of the proposed upconversion scheme, we use the model developed in the previous sections, with parameters based on the self-quenched InGaAs/InP SPAD described in [153]. The device has an active area of 10 μm per side and a junction capacitance of 150 fF, dominated by the capacitance of the depleted region. Due to the optical readout, the off-chip routing and the sensing circuit's capacitance, which can be on the order of a picofarad in SPADs with electrical readout, are eliminated. The passivation thickness of a 6-metal layer silicon device (as

was used in [103]) is on the order of 7 μm, so InP-SPAD/Si-SPAD capacitance is negligible.. It operates at an overbias of 5V with a junction located 200 nm below the surface. For the silicon detector, we compare the performance of two commercially-available detectors [109, 161].

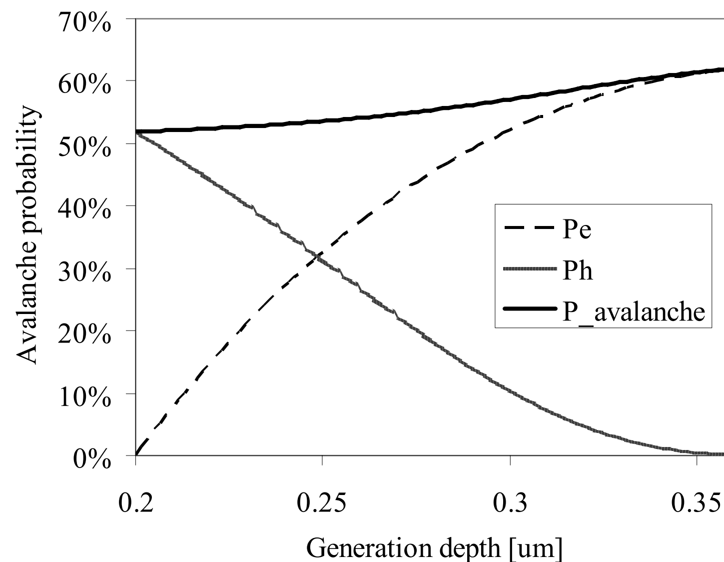


Figure 5.5 Numerical analysis of electron, hole and total avalanche initiation probabilities (P_e, P_h and P_p) as a function of photon absorption depth in a Si SPAD.

From Equation 5.3, we can estimate that 4.7 x 10^6 electrons flow during an avalanche, and from Equation 5.4 we calculate that 131 photons are emitted isotropically in the silicon absorption band from the junction. The spectral density of these photons is shown in Figure 5.6, both at the junction and, using Equation 5.7, at the surface of the InGaAs/InP SPAD.

The upconversion efficiency can now be determined using the emitted spectral

density and the sensitivity of the silicon detector. Results indicate a 97% detection probability for the Cova device (at 5V overbias) [109] and 91% for the Rochas SPAD [161]. These numbers can be further raised if the total charge flowing in the InGaAs/InP SPAD is increased (Figure 5.7), or by increasing the silicon junction's depletion width, thereby increasing its detection efficiency for long wavelengths.

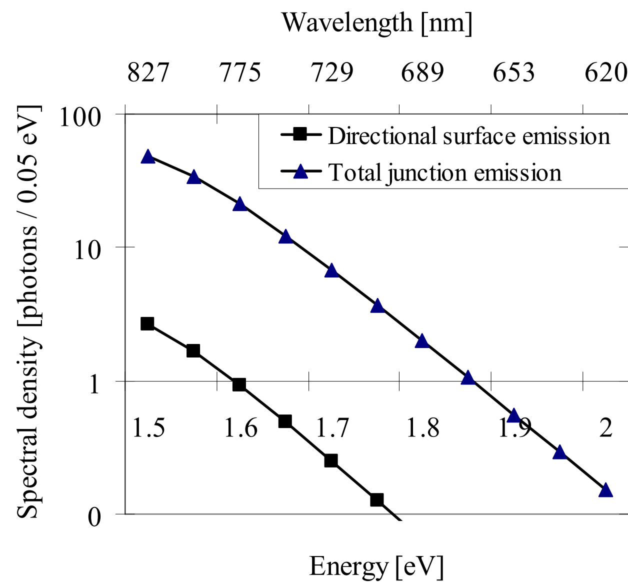


Figure 5.6: Junction and surface electroluminescence spectral densities for the 200 nm deep InP junction described in the text. The lower curve accounts for only those photons emitted towards the silicon junction. Absorption coefficients were taken from [162].

Because the silicon SPAD will also generate electroluminescent photons, it is important to prevent a positive feedback loop between the junctions. This can be achieved by controlling the dead time of the IR SPAD so it overlaps the avalanche time of the silicon device.

The power dissipated during the upconversion process is the sum of the powers

dissipated during the InP and silicon avalanche. These can be estimated as the product of the junction capacitances by their overbias, resulting in approximately 1 pW per detected photon, significantly lower than reported schemes [2].

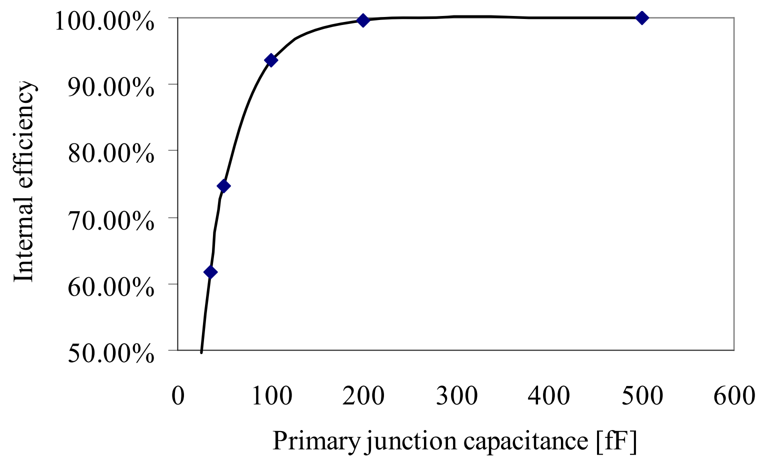


Figure 5.7 Numerical simulation of internal upconversion efficiency as a function of primary SPAD's junction capacitance.

5.4. Experimental Measurements of Hot-Carrier Electroluminescence

5.4.1. Experimental Methodology

The calculations presented in the previous sections demonstrated that electroluminescence yield, i.e., the number and spectrum of emitted photons per hot charge carrier, is the determining factor for upconversion efficiency. As stated above, to date, only the electroluminescence yield of silicon has been measured. In this

section, we will describe the methodology, setup and results of electroluminescence yield measurements for a device manufactured for efficient emission by Mr. Kai Zhao, under Prof. Yu-Hwa Lo's supervision. This device has a separate absorption and multiplication structure (SAM), whereby absorption is in an InGaAs layer and multiplication is performed in an InAlAs layer. Additional layers are used for lattice matching and for additional functionality.

Measuring the electroluminescent yield requires the determination of the absolute photon emission as a function of wavelength and hot-carriers. This is quite challenging due to the unavoidable losses due to limited collection efficiency and optical response of any optical system. In order to overcome these obstacles, we used the following methodology:

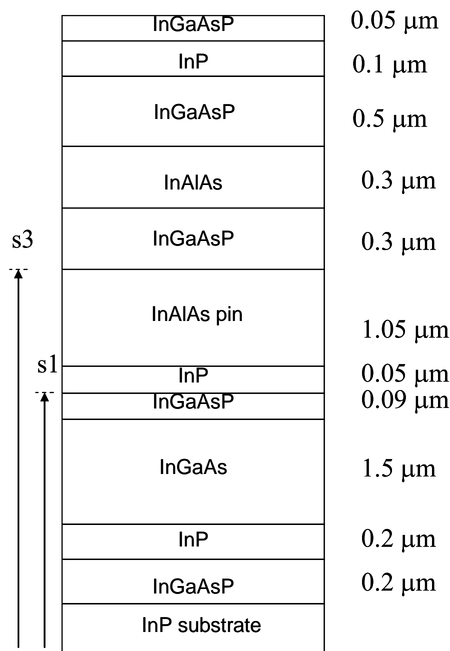


Figure 5.8: Cross-section of InGaAs/InAlAs device used for emission measurements (fabricated by Kai Zhao).

- Calibrate the spectral output of a CCD camera and monochromator using known emission lines of a mercury lamp.
- Measure electroluminescent spectrum and intensity of a silicon SPAD of similar dimensions as the III-V SPAD, as a function of avalanche current.
- Determine junction emission in intensity units by accounting for self-absorption.
- Determine the collection efficiency of the setup by comparing the measured intensity results with published photon emission data for silicon (assuming the emission rates depend mainly on the avalanching material rather than device structure).
- Measure the electroluminescence surface emission and spectrum as a function of avalanche current and convert to photons.

Measurements were made using a cooled PIXIS 400B CCD camera, with 3.5 e^- rms readout noise and a dark current of 0.005 e^- per second [163], and with a Princeton Instruments SpectraPro 2150i (Acton Research SP150) monochromator with 300 grooves per millimeter. Data was captured in Princeton Instruments' Winspec, and was analyzed in Excel.

5.4.2. Optical Response

For the silicon emission measurements, the diffused-ring SPAD was used, rather than the STI SPAD. The latter's emission was found to be stronger at the

periphery of the active area, with shorter wavelength emission. This suggests that emission is through the transparent silicon-dioxide trench which does not absorb the shorter wavelengths. In the diffused-ring SPAD, emission is uniform and assumed to emanate from the junction depth of 0.2 μm (Figure 5.9). Integration of the image intensity results in a total collected charge by the camera of 6.7×10^4 electrons.

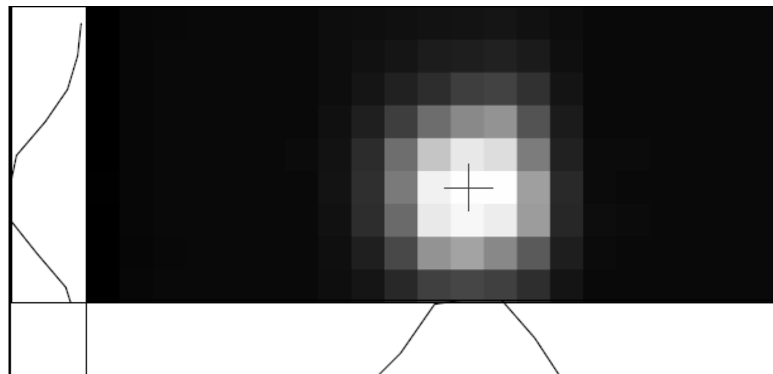


Figure 5.9 Hot-carrier electroluminescence from a diffused-ring silicon SPAD.

Average avalanche current was determined by measuring the current sunk from the p^+ high-voltage supply. Emission intensity was integrated over the whole image for various currents, and was found to be linearly proportional to the current, indicating all the measured current takes part in the avalanche.

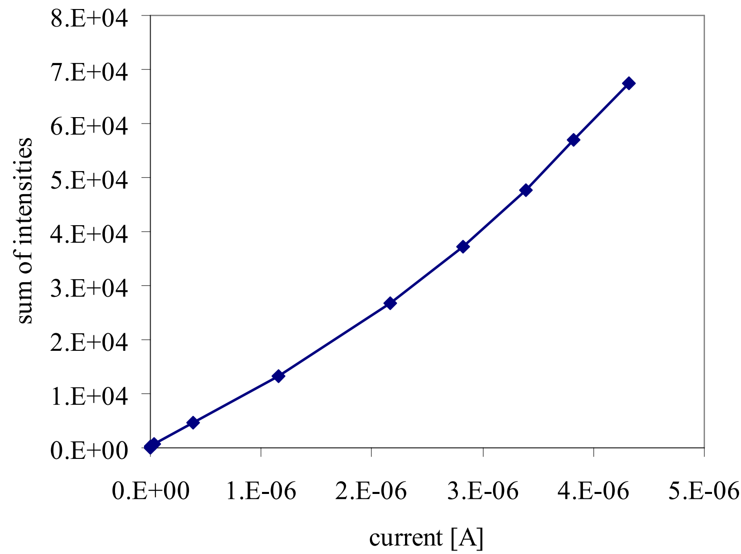


Figure 5.10: Total emission intensity as a function of avalanche current.

The emission spectra were obtained through a thin slit, which scattered different wavelengths to specific pixels on the CCD. Each intensity value in the camera represented the integral of intensities over the wavelength range corresponding to that pixel. The spectral density is more relevant to our analysis. For a pixel intensity $N(\Delta\lambda)$, the corresponding spectral density is:

$$N(E) = \frac{N(\Delta\lambda)}{\Delta E}$$

Equation 5.13

where ΔE is the energy range corresponding to $\Delta\lambda$. Figure 5.11 shows the spectral histogram and spectral densities obtained with a 140 μm-wide slit with the SPAD biased at 12.9 V. The average current was 4.36 μA and integration time was 100 seconds.

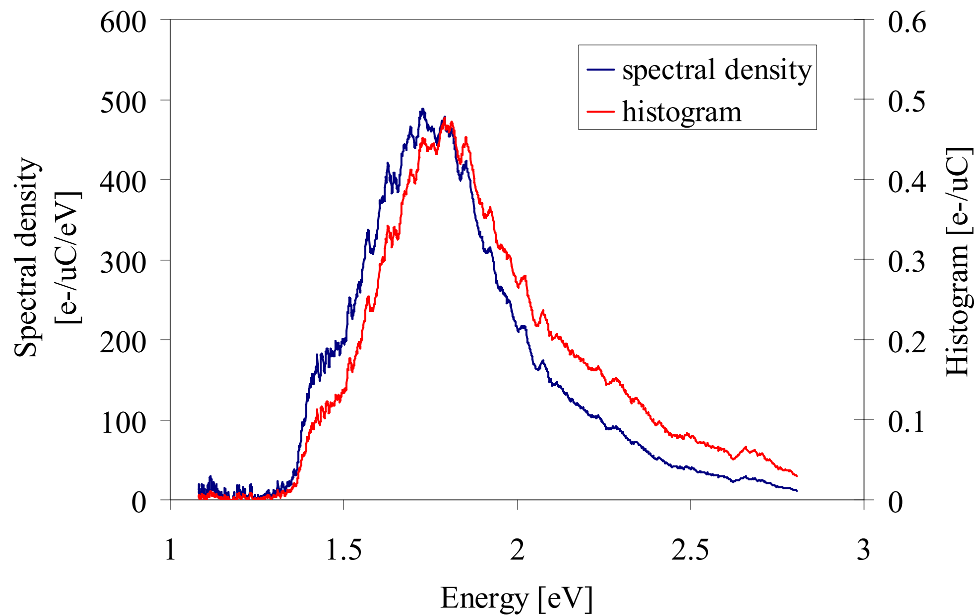


Figure 5.11: Measured silicon electroluminescence spectrum.

The area under the graph above is the total collected charge. We normalize to get the collected spectrum in units of electrons per avalanche charge per energy band.

In order to determine the optical response of the system, we compare our experimental measurements to published absolute electroluminescence data [129], corrected for 0.2 μm self-absorption in silicon [164]. The two agree well, except for the low energy range where the silicon CCD has little sensitivity.

The optical response curve of the system can be readily computed from the ratio of the curves above, and is shown in Figure 5.13.

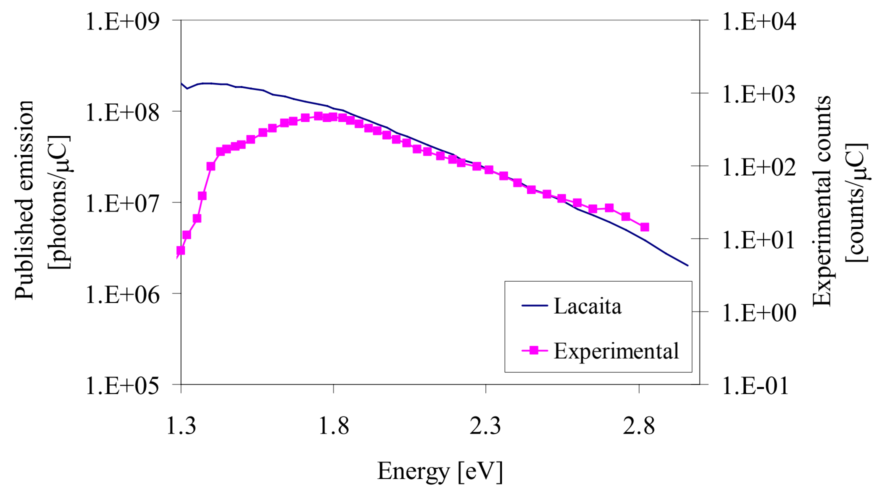


Figure 5.12: Experimental and published electroluminescence curves for 0.2 μm-deep silicon junction.

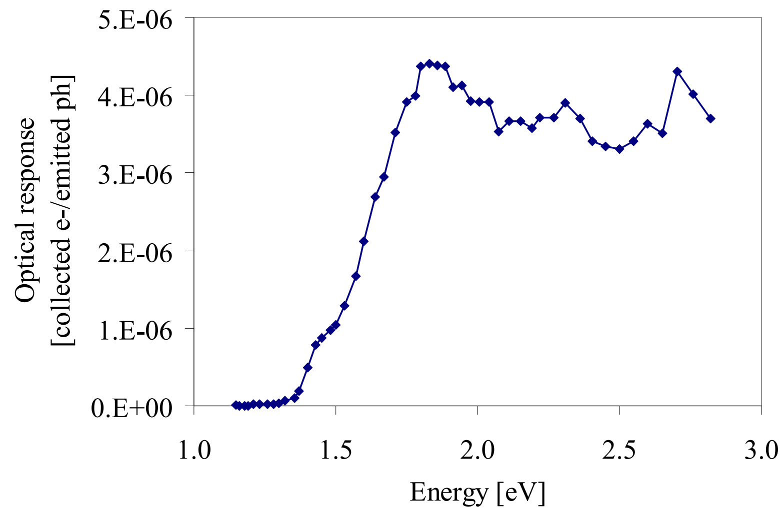


Figure 5.13 Collection efficiency of optical setup.

5.4.3. InAlAs/InGaAs Electroluminescence Yield

We followed similar steps as above in order to determine the emission yield of the InAlAs/InGaAs device. Electroluminescence of a test device was imaged using the same setup. The measured emission spectrum is shown in Figure 5.15. This spectrum must be corrected for the optical response of the system, resulting in the absolute emission spectrum, shown in Figure 5.16.

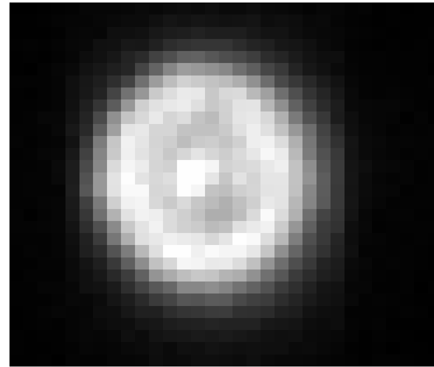


Figure 5.14 Electroluminescence from an InAlAs/InGaAs SPAD (45 μA, 100s exposure).

The measured exponential decrease in emission with photon energy is in agreement with Equation 5.1. The observed absorption band corresponds to the bandgap of the InGaAsP layer. However, the emission maximum at 1.2 eV does not correspond to the bandgap of the multiplication material (InAlAs – 1.5 eV), indicating that, most likely, recombination and emission takes place outside this region, via a more complex process than the one described above.

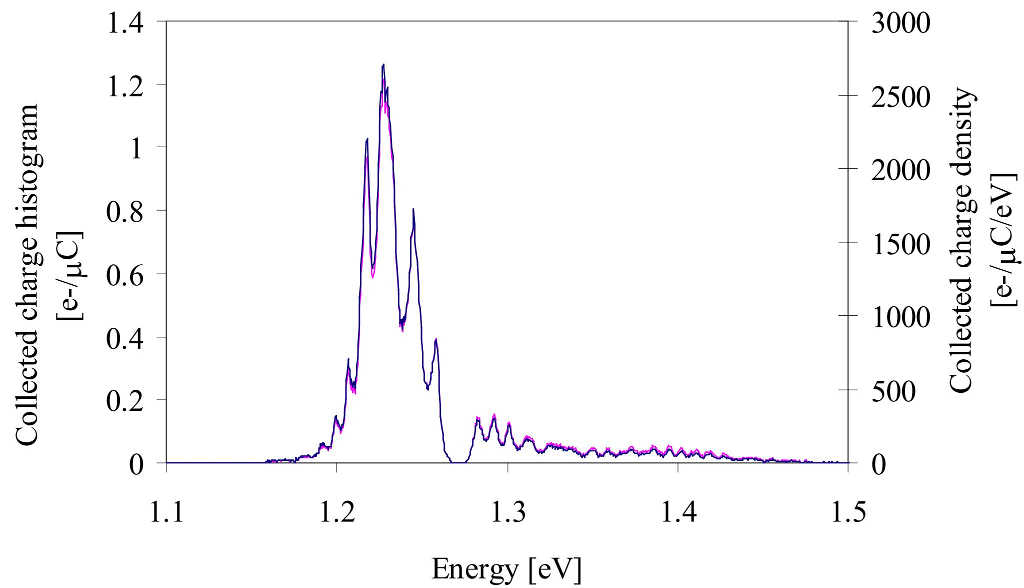


Figure 5.15 Electroluminescence spectrum of InAlAs/InGaAs SPAD.

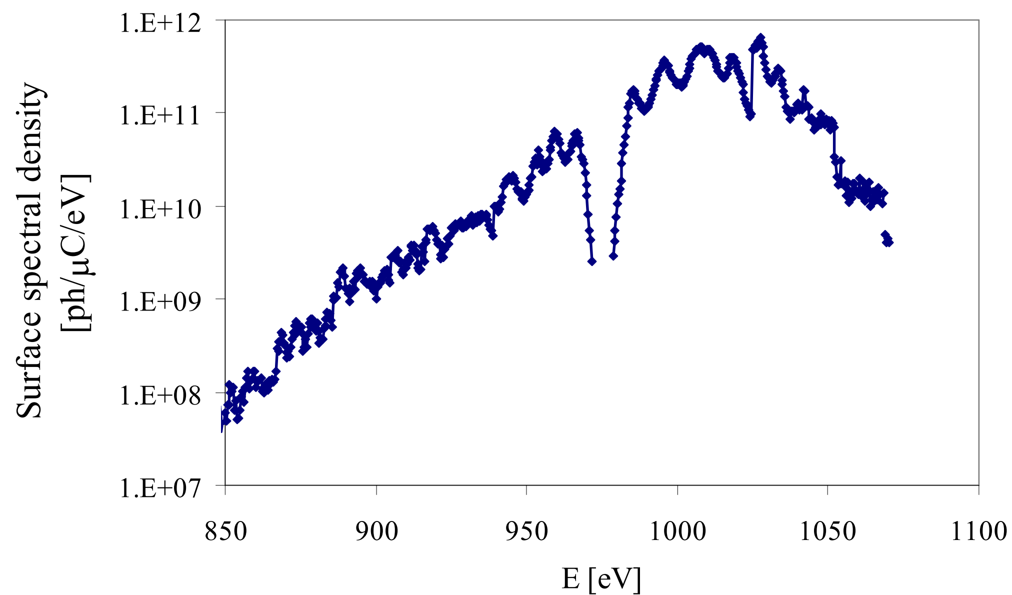


Figure 5.16 Electroluminescence surface emission from InAlAs/InGaAs SPAD.

An integration of the graph above results in the total surface emission yield – 3.5×10^{-3} photons/carrier in the absorption band of silicon.

5.4.4. InAlAs Optical Absorption

In order to determine the junction emission, we measured the absorption characteristics of the top layers above the junction. We present these results primarily for the derivation of the absorption curve of InAlAs, a material of interest in optical devices, whose absorption characteristics have not been reported to date.

In order to perform this analysis, two samples were back-etched from the SPAD wafers, labeled *s1* and *s3* in Figure 5.8, by repeated exposure to HCl and H_2SO_4. The back-thinned samples were illuminated with a halogen lamp, and imaged through the setup described above.

In traditional SPADs, carriers recombine inside the high-field region. Therefore, the absorption of the *s1* layer will be stronger than that of photons generated via electroluminescence, while the absorption of the *s3* layer will be weaker. The absorption values obtained from these test structures are plotted in Figure 5.17.

The ratio of the transmittances through *s1* and *s3* corresponds to the transmittance of a 1.05 μm layer of InAlAs (we neglect the effect of the 50 nm layer of InP). From Equation 2.37, the absorption coefficient can be calculated as:

$$\alpha = \frac{-\ln\left(1 - A(z)\right)}{z}$$

Equation 5.14

where $A(z)$ is the absorbance through a layer of width z. Results for InAlAs are shown

in Figure 5.18.

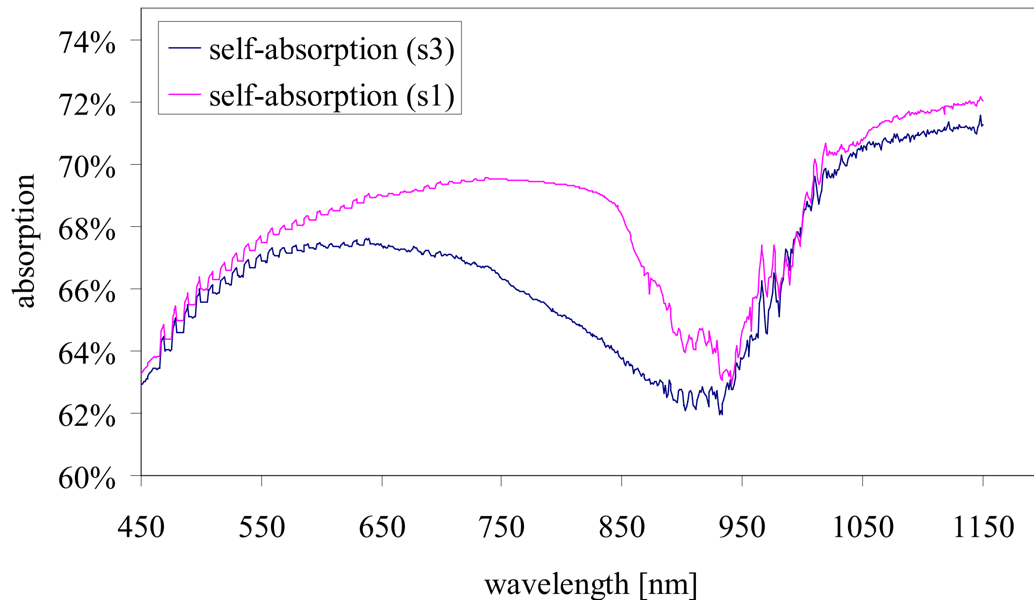


Figure 5.17 Absorbance of layer s1 and s3 corresponding to lower- and upper-bounds of junction electroluminescence yields.

The operation of the present device is somewhat different that standard passively- or actively-quenched SPADs. Avalanche quenching is achieved via charge accumulation resulting from band discontinuities of the top layers of the device. As such, recombination seems to take place in one of the top layers rather than in the depletion regions. This can explain the location of the emission peak at 1.2 eV rather than the bandgap of InAlAs, which is 1.5 eV. This mechanism is still under investigation.

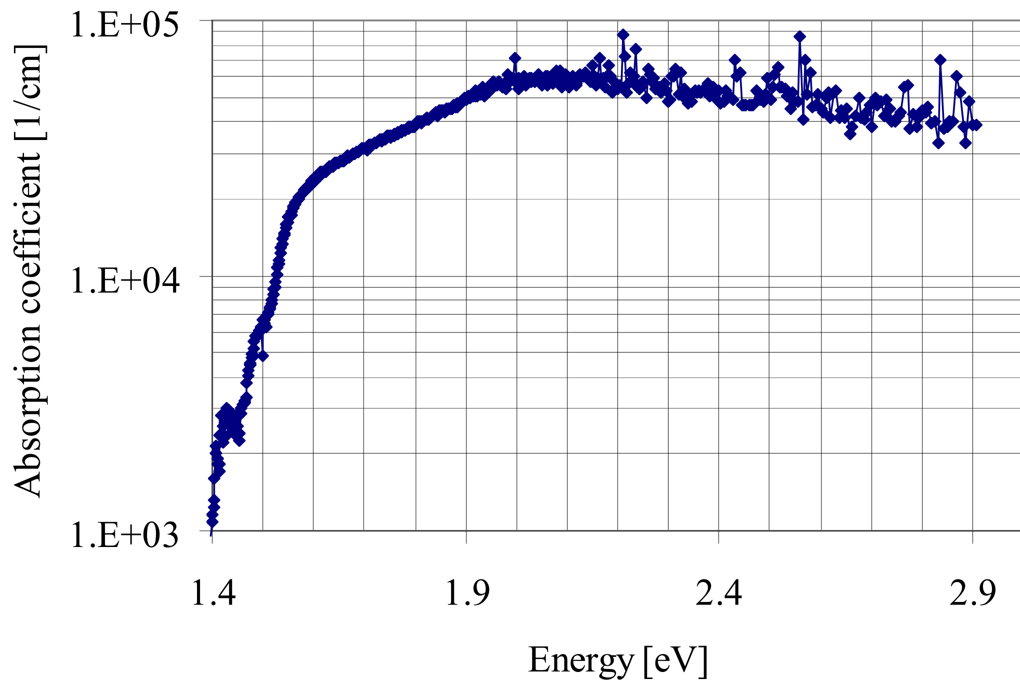


Figure 5.18: Measured absorption coefficient of InAlAs.

5.5. Conclusions

We proposed and analyzed a novel method for single-photon upconversion. The benefits of the proposed scheme are fourfold: it offers wire-free, parasitic-free, and massively parallel interconnection between the IR-SPAD and the CMOS silicon SPAD, enabling large-scale array integration of III-V SPADs; the elimination of bumps allows for high pixel density and high fill factors; the low-temperature glass-to-glass bonding processes replaces large-array flip-chip bonding which would be otherwise required and greatly simplifies the manufacturing of the hybrid devices; and it reduces avalanche charge, resulting in lower power and reduced afterpulsing.

In order to achieve upconversion, this method utilizes a natural byproduct of the avalanche process, specifically the spectral component of the electroluminescent photons which is higher than the bandgap of the absorbing material of the detector. We demonstrated analytically that a silicon SPAD with a spectral response similar to commercially-available devices, can detect these secondary photons with a 91%-97% detection probability, resulting in a high-efficiency low-power upconversion scheme. Requirements for the amount of avalanche charge and electroluminescence yield necessary for efficient upconversion were quantified.

Acknowledgement:

This chapter, in part, is a reprint of the material as it appears in the following publication:

H. Finkelstein, M. Gross, YH Lo, and S. Esener, "Analysis of Hot-Carrier Luminescence for Infrared Single-Photon Up-Conversion and Readout", to appear in *IEEE Journal of Selected Topics in Quantum Electronics – Single Photon Counting: Detectors and Applications*, 2007.

6. CONCLUSION

6.1. Dissertation Summary and Original Contributions

This dissertation presented a novel implementation of a single-photon avalanche diode, which is fully compatible with commercial deep-submicron technologies. Whereas all prior surface SPADs utilized a variation of a diffused guard-ring, the present work explored a new and efficient technique for planarizing the p-n junction at the center of the SPAD and of isolating it from adjacent structures. As shown in Chapter 1, in FCS and related applications, a reduction in SPAD dead time is equivalent to an improvement in detection efficiency. This forms an added incentive for SPAD miniaturization.

We developed an analytical model for the operation of the SPAD, and simulated its operation, from an electrostatic and an electrical perspective. We also demonstrated some of the benefits of SPAD integration in deep-submicron technologies by designing novel ultrafast circuitry in an area-efficient manner. The new SPAD exhibits the highest fill factors and the fastest operation of any SPAD to date.

Finally, we introduced and analyzed a new and efficient single-photon upconversion scheme and presented preliminary experimental results confirming its feasibility.

The original contributions of this work are summarized below:

- Theoretical analysis of the trade-offs related to device miniaturization, including noise – speed trade-off, as well as analysis of the applications that stand to benefit from miniaturization.
- Invention of the first SPAD device in a generic commercial deep-submicron CMOS technology, based on a new shallow-trench isolation concept for junction planarization and device isolation.
- Development of a simulation environment for electrostatic and electrical simulation of the device and its peripheral circuitry.
- Design of a novel self-timed circuits for active quenching and recharging, aimed at an optimal trade-off between device speed and afterpulsing.
- Invention and analysis of a dual-junction SPAD, based on two junctions per pixel, for cross-correlation experiments.
- Invention and simulation of a CMOS circuit for the quenching and processing of the dual-junction SPAD's outputs.
- Development of a complete characterization environment for single-photon detection, including detection efficiency, output signal statistics, spectral response, dead time, and cross-talk. This includes the development of several efficient new methods for experimentally measuring auto- and cross-correlation functions in real-time, for large pulse populations.
- Characterization of passive- and active-quenched SPADs, proof of planarity, and experimental measurement of the key metrics of the SPAD devices, and analysis of the experimental data, showing a good fit with analytical and simulated results.

- Comparison of device performance to a "traditional", diffused guard-ring SPAD.
- Invention of a new single-photon upconversion technique based on hot-carrier recombination-induced electroluminescence, including theoretical modeling and feasibility analysis.
- Experimental measurement of electroluminescence yield of a near-infrared InGaAs/InAlAs avalanche diode. This is the only semiconductor other than silicon whose hot-carrier recombination-induced electroluminescence yield has been measured.
- First experimental determination of the absorption curve of InAlAs.

6.2. Outlook

The outlook for single-photon detection can be divided into several device categories:

1. Highly-sensitive low-jitter single-element SPADs: These detectors are targeted for non-imaging fluorescence lifetime experiments. Custom processes seem to be optimal for such devices, because large active areas are desirable for ease of optical alignment. Detection efficiency and low jitter are of utmost importance, for the detection of fluorescence from single molecules, and processing can be performed off-chip, even at the expense of higher cost and wiring complexity. Ultra-low-noise is highly desired for these applications, so that non-time-gated

operation is possible. Surface, custom-processed SPADs are likely to continue to dominate this category.

2. Medium-sensitivity ultra-low jitter single-element SPADs: These devices are optimal for quantum-key distribution systems, in conjunction with a frequency-upconversion scheme. Because time-gating is possible in these applications, ultra-low jitter is highly-desirable as it limits the gate width, thereby setting the noise of the receiver. Quantum key distribution protocols, such as BB84, can achieve low quantum bit-error rates even with detection efficiencies as low as 30% over tens of kilometers [2]. Data rates in these detectors are limited by the dead time of the devices, so fast devices are highly desirable. The STI-bounded device described in this work is a good candidate for this type of application.
3. Medium-sensitivity low-jitter highly-scalable SPADs: A mega-pixel single-photon avalanche-diode array can revolutionize several fields. In FLIM, such devices will make it possible to monitor the movement of several molecules in parallel with high resolution, while simultaneously monitoring subtle environmental changes affecting their fluorescence lifetimes. In an unrelated application, 3D images can be obtained by measuring the times-of-arrivals of actively-illuminated targets with high-timing precision. High-resolution time-domain optical tomography can also greatly benefit from high spatial and temporal resolution SPAD arrays. All the above applications can operate in a non-photon-limited mode, i.e., where each exposure contains more than one photon, thereby greatly increasing the detection probability, even with moderately-sensitive SPADs. In order to achieve megapixel

arrays, individual pixels must be highly area-efficient and have high fill factors. This category is the most promising for the STI-bounded SPADs as discussed below.

4. High-sensitivity low-jitter highly-scalable SPADs: This is of course the best of all worlds, but does not seem to be feasible due to inherent physical trade-offs which were discussed in this work. Specifically, high-sensitivity requires wide depletion regions, but these are only available in low-doped “old” technologies, resulting in large pixels, making megapixel integration not realistic.

A first step in the implementation of megapixel SPAD arrays is the design of small-area pixels. A megapixel array of 20 μm^2 pixels, as were demonstrated in this work with 80% array area utilization will occupy 25 mm^2, but with a 625 μm^2 pixel, as was the state-of-the-art [71], 780 mm^2 will be required, making it unrealistic.

A next and necessary step for the evolution of a megapixel SPAD imager is the development of architectures for the processing of the array’s data and the reduction of the data bandwidth which needs to be output from the chip. In order to understand this requirements, assume a megapixel SPAD array, where each SPAD only fires during 0.1% of the cycles (sparse signal, such as is common in dilute biological samples), and with a pulsed-laser excitation with a 20 MHz pulse-repetition rate. Assuming we would like to have a 10 ps timing resolution within a 1 µs time-gate, we will need approximately:

$$n_{bits} = \ln\left(\frac{t_{frame}}{t_{precision}}\right) = \ln\left(\frac{1\times10^{-6}}{10\times10^{-12}}\right) \approx 12\,\text{bits}$$

Equation 6.1

per detected photon. The data bandwidth produced by such an array will be:

$$BW = \eta_{firing} \times n_{bits} \times N_{pixels} \times R_{pulses} = 0.1\% \times 12 \times 1024 \times 1024 \times 20 \times 10^{6} \approx 250\ \text{Gb/sec}$$

Equation 6.2

which is an unrealistically high amount of data. This means that some amount of processing and data reduction must be performed on-chip, using radically different architectures than those used by traditional image sensors.

Data bandwidth can also be reduced by setting limitations on the imager, such as only 1 pixel can fire per column per exposure; reduced timing precision at maximal spatial resolution; or reduced spatial resolution at maximal timing precision.

The design, simulation and characterization of these architectures is a subject of future research, as outlined below.

6.3. Future Research

This dissertation summarized research on various aspects of single-photon detection, including device design and simulation, circuit design, characterization and integration in a frequency upconversion hybrid device. This research should be seen as a first step in the implementation of imaging systems utilizing the many advantages resulting from SPAD integration in advanced CMOS technologies. Below, we will

outline some of the exciting opportunities for innovation arising from the work described here.

On the device-design level, it is desirable to investigate and locate the source of the excessive afterpulses. It is likely that these may arise from the oxide interface. Implementing the variable-delay active-recharge circuitry described in section 3.4.2.3 can result in optimized device operation, with acceptable noise levels, higher detection efficiencies and high fill factors.

The future availability of high-frequency packages, such as the QFN (Quad Flat No-lead), should allow for accurate measurement of the device jitter. This can be performed using the setup previously described by Cova [63]. Alternately, the deep submicron CMOS process creates opportunities for performing built-in jitter testing [165].

Several deep-submicron CMOS technologies, especially DRAM processes, offer deep-trench isolation. These structures may make it possible to planarize the deep N-well / p-substrate junction and utilize it for detection in longer wavelengths. The polysilicon lining of these trenches may be a mixed blessing – they may prevent charge trapping and afterpulsing, but may also result in parasitic MOS structures being formed between pixels. All of these may be topics of future research.

The prospect of a dual-junction SPAD which was proposed and analyzed in this work offers a real and significant benefit to a number of applications. The manufacturing and characterization of the proposed device and peripheral circuitry is a promising avenue. Future steps may include optimization of the device operation by

selecting junctions which are better separated spectrally, or by making use of the color filters available in commercial CMOS image sensor technologies, to reduce spectral cross-talk between the planar shallow junction, and the deep curved ring.

The benefits of a high-fill factor fully-CMOS compatible SPAD can only be realized if some or most of the signal processing is done on-chip. The first stage in processing the SPAD pulses is converting their times-of-arrival to digital format, using time-to-digital converters (TDC) [166]. For TCSPC and especially FLIM applications, the TDC must have sub-10 ps precision and must be small enough to be scaled for parallel photon detection. The design and integration of such converters is a major challenge to the successful integration of CMOS SPADs.

As discussed in the previous sub-section, the prospect of a megapixel SPAD array can only be realized if a new architecture paradigm can be derived and implemented, which will significantly reduce the IO bandwidth of the imager, and which can process the large amounts of data generated in real-time and in an area- and power-efficient manner. Preliminary work has recently been published on this topic [167] but there is much room for improvements.

Chapter 5 of this dissertation presented the concept of single-photon frequency-upconversion based on hot-carrier electroluminescence. Having shown that the emission yield is sufficiently high for efficient upconversion, future work should provide a proof-of-concept and, eventually result in a working NIR high-resolution imager. First, single-photon upconversion can be characterized using free-space optics. Next, silicon and III-V (e.g., InGaAs/InAlAs) dies should be fused together on

their glass layers, after developing the proper alignment and annealing steps. Lastly, arrays of these detectors can be manufactured, with the processing and control being preformed on the silicon SPAD die.

REFERENCES

[1] X. Michalet, O. H. W. Siegmund, J. V. Vallerga, P. Jelinsky, J. E. Millaud, and S. Weiss, "Detectors for single-molecule fluorescence imaging and spectroscopy," *Journal of Modern Optics*, vol. 54, pp. 239 - 281, 2007.

[2] R. T. Thew, S. Tanzilli, L. Krainer, S. C. Zeller, A. Rochas, I. Rech, S. Cova, H. Zbinden, and N. Gisin, "Low jitter up-conversion detectors for telecom wavelength GHz QKD," *New Journal of Physics*, vol. 8, pp. 32, 2006.

[3] D. Rosenberg, S. W. Nam, P. A. Hiskett, C. G. Peterson, R. J. Hughes, J. E. Nordholt, A. E. Lita, and A. J. Miller, "Quantum key distribution at telecom wavelengths with noise-free detectors," *Applied Physics Letters*, vol. 88, pp. 021108 , 2006.

[4] B. F. Aull, A. H. Loomis, D. J. Young, A. Stern, B. J. Felton, P. J. Daniels, D. J. Landers, L. Retherford, D. D. Rathman, R. M. Heinrichs, R. M. Marino, D. G. Fouche, M. A. Albota, R. E. Hatch, G. S. Rowe, D. G. Kocher, J. G. Mooney, M. E. O'Brien, B. E. Player, B. C. Willard, Z. L. Liau, and J. J. Zayhowski, "Three-dimensional imaging with arrays of Geiger-mode avalanche photodiodes," *Proceedings of the SPIE*, vol. 5353, pp. 105-116, 2004.

[5] D. K. Joseph, T. J. Huppert, M. A. Franceschini, and D. A. Boas, "Diffuse optical tomography system to image brain activation with improved spatial resolution and validation with functional magnetic resonance imaging," *Applied Optics*, vol. 45, pp. 8142-8151, 2006.

[6] C. Niclass, A. Rochas, P. A. Besse, and E. Charbon, "Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 1847-1854, 2005.

[7] D. A. Hicks, "Deep industrial dynamics shaping next-generation semiconductor manufacturing," presented at Semiconductor Manufacturing Conference Proceedings, 1997 IEEE International Symposium on, 1997.

[8] J. S. Dunn, D. C. Ahlgren, D. D. Coolbaugh, N. B. Feilchenfeld, G. Freeman, D. R. Greenberg, R. A. Groves, F. J. Guarin, Y. Hammad, A. J. Joseph, L. D. Lanzerotti, S. A. St Onge, B. A. Orner, J. S. Rieh, K. J. Stein, S. H. Voldman, P. C. Wang, M. J. Zierak, S. Subbanna, D. L. Harame, D. A. Herman, Jr., and B. S. Meyerson, "Foundation of RF CMOS and SiGe BiCMOS technologies," *IBM Journal of Research and Development*, vol. 47, pp. 101-38, 2003.

[9] S. Mendis, S. E. Kemeny, and E. R. Fossum, "CMOS active pixel image sensor," *IEEE Transactions on Electron Devices*, vol. 41, pp. 452-453, 1994.

[10] P. Andersson, "Long-range three-dimensional imaging using range-gated laser radar images," *Optical Engineering*, vol. 45, pp. 034301, 2006.

[11] M. A. Albota, R. M. Heinrichs, D. G. Kocher, D. G. Fouche, B. E. Player, M. E. O'Brien, B. F. Aull, J. J. Zayhowski, J. Mooney, B. C. Willard, and R. R. Carlson, "Three-dimensional imaging laser radar with a photon-counting avalanche photodiode array and microchip laser," *Applied Optics*, vol. 41, pp. 7671-7678, 2002.

[12] R. M. Marino and W. R. J. Davis, "Jigsaw: a foliage-penetrating 3D imaging laser radar system.," *Lincoln Laboratory Journal*, vol. 15, pp. 23-36, 2005.

[13] G. S. Buller, R. D. Harkins, A. McCarthy, P. A. Hiskett, G. R. MacKinnon, G. R. Smith, R. Sung, A. M. Wallace, R. A. Lamb, K. D. Ridley, and J. G. Rarity, "Multiple wavelength time-of-flight sensor based on time-correlated single-photon counting," *Review of Scientific Instruments*, vol. 76, pp. 083112, 2005.

[14] W. E. Moerner and D. P. Fromm, "Methods of single-molecule fluorescence spectroscopy and microscopy," *Review of Scientific Instruments*, vol. 74, pp. 3597-3619, 2003.

[15] K. Bacia and P. Schwille, "A dynamic view of cellular processes by in vivo fluorescence auto- and cross-correlation spectroscopy," *Methods*, vol. 29, pp. 74-85, 2003.

[16] T. Waizenegger, R. Fischer, and R. Brock, "Quantification of cellular uptake of small molecules," Carl Zeiss, 2007.

[17] J. L. Swift, R. Heuff, and D. T. Cramb, "A two-photon excitation fluorescence cross-correlation assay for a model ligand-receptor binding system using quantum dots," *Biophysical Journal*, vol. 90, pp. 1396-1410, 2006.

[18] X. Michalet, S. Weiss, and M. Jager, "Single-molecule fluorescence studies of protein folding and conformational dynamics," *Chemical Reviews*, vol. 106, pp. 1785-1813, 2006.

[19] S. Ruttinger, R. Macdonald, B. Kramer, F. Koberling, M. Roos, and E. Hildt, "Accurate single-pair Forster resonant energy transfer through combination of pulsed interleaved excitation, time correlated single-photon counting, and fluorescence correlation spectroscopy," *Journal of Biomedical Optics*, vol. 11, pp. 024012, 2006.

[20] H. Yang, G. B. Luo, P. Karnchanaphanurach, T. M. Louie, I. Rech, S. Cova, L. Y. Xun, and X. S. Xie, "Protein conformational dynamics probed by single-molecule electron transfer," *Science*, vol. 302, pp. 262-266, 2003.

[21] A. Leon-Garcia, Probability and Random Processes for Electrical Engineering, 2 ed. Toronto: Addison-Wesley, 1994.

[22] S. Maiti, U. Haupts, and W. W. Webb, "Fluorescence correlation spectroscopy: Diagnostics for sparse molecules," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, pp. 11753-11757, 1997.

[23] S. T. Hess, S. H. Huang, A. A. Heikal, and W. W. Webb, "Biological and chemical applications of fluorescence correlation spectroscopy: A review," *Biochemistry*, vol. 41, pp. 697-705, 2002.

[24] O. Krichevsky and G. Bonnet, "Fluorescence correlation spectroscopy: the technique and its applications," *Reports on Progress in Physics*, vol. 65, pp. 251-297, 2002.

[25] M. Sauer, B. Angerer, W. Ankenbauer, Z. Foldes-Papp, F. Gobel, K. T. Han, R. Rigler, A. Schulz, J. Wolfrum, and C. Zander, "Single molecule DNA sequencing in submicrometer channels: state of the art and future prospects," *Journal of Biotechnology*, vol. 86, pp. 181-201, 2001.

[26] R. C. Habbersett and J. H. Jett, "An analytical system based on a compact flow cytometer for DNA fragment sizing and single-molecule detection," *Cytometry Part A*, vol. 60A, pp. 125-134, 2004.

[27] L. Onsager, "Reciprocal relations in irreversible processes.," *Physical Review*, vol. 37, pp. 405 LP - 426, 1931.

[28] W. Becker, A. Bergmann, M. A. Hink, K. Konig, K. Benndorf, and C. Biskup, "Fluorescence lifetime imaging by time-correlated single-photon counting," *Microscopy Research and Technique*, vol. 63, pp. 58-66, 2004.

[29] M. Tramier, I. Gautier, T. Piolot, S. Ravalet, K. Kemnitz, J. Coppey, C. Durieux, V. Mignotte, and M. Coppey-Moisan, "Picosecond-hetero-FRET microscopy to probe protein-protein interactions in live cells," *Biophysical Journal*, vol. 83, pp. 3570-3577, 2002.

[30] P. Schwille, F. J. MeyerAlmes, and R. Rigler, "Dual-color fluorescence cross-correlation spectroscopy for multicomponent diffusional analysis in solution," *Biophysical Journal*, vol. 72, pp. 1878-1886, 1997.

[31] M. Leutenegger, H. Blom, J. Widengren, C. Eggeling, M. Gosch, R. A. Leitgeb, and T. Lasser, "Dual-color total internal reflection fluorescence cross-

correlation spectroscopy," *Journal of Biomedical Optics*, vol. 11, pp. 040502, 2006.

[32] A. Periasamy, P. Wodnicki, X. F. Wang, S. Kwon, G. W. Gordon, and B. Herman, "Time-resolved fluorescence lifetime imaging microscopy using a picosecond pulsed tunable dye laser system," *Review of Scientific Instruments*, vol. 67, pp. 3722-3731, 1996.

[33] H. J. Lin, P. Herman, and J. R. Lakowicz, "Fluorescence lifetime-resolved pH imaging of living cells," *Cytometry Part A*, vol. 52A, pp. 77-89, 2003.

[34] W. Becker, K. Benndorf, A. Bergmann, C. Biskup, K. Konig, U. Tirplapur, and T. Zimmer, "FRET measurements by TCSPC laser scanning microscopy," in *Proceedings of the SPIE - The International Society for Optical Engineering, USA.*, vol. 4431: SPIE-Int. Soc. Opt. Eng., 2001, pp. 94-98.

[35] V. Ntziachristos, A. G. Yodh, M. Schnall, and B. Chance, "Concurrent MRI and diffuse optical tomography of breast after indocyanine green enhancement," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, pp. 2767-2772, 2000.

[36] W. Becker, The bh TCSPC handbook, 2nd edition ed. Berlin: Becker & Hickl GmbH, 2006.

[37] R. Cubeddu, A. Pifferi, P. Taroni, A. Torricelli, and G. Valentini, "Noninvasive absorption and scattering spectroscopy of bulk diffusive media: An application to the optical characterization of human breast," *Applied Physics Letters*, vol. 74, pp. 874-876, 1999.

[38] A. P. Gibson, J. C. Hebden, and S. R. Arridge, "Recent advances in diffuse optical imaging," *Physics in Medicine and Biology*, vol. 50, pp. R1-R43, 2005.

[39] Hamamatsu, "Photomultiplier tubes: basics and applications," Third ed. http://sales.hamamatsu.com/assets/applications/ETD/pmt_handbook_complete.pdf, 2007.

[40] T. Ohnuki, X. Michalet, A. Tripathi, S. Weiss, and K. Arisaka, "Development of an ultrafast single photon counting imager for single molecule imaging," presented at Ultrasensitive and Single-Molecule Detection Technologies, Boston, 2006.

[41] H. Kume, K. Koyama, K. Nakatsugawa, S. Suzuki, and D. Fatlowitz, "Ultrafast Microchannel Plate Photomultipliers," *Applied Optics*, vol. 27, pp. 1170-1178, 1988.

[42] X. Michalet, O. H. W. Siegmund, J. V. Vallerga, P. Jelinsky, J. E. Millaud, and S. Weiss, "Photon-counting H33D detector for biological fluorescence imaging," *Nuclear Instruments & Methods in Physics Research Section A-Accelerators Spectrometers Detectors and Associated Equipment*, vol. 567, pp. 133-136, 2006.

[43] A. J. Kerman, E. A. Dauler, B. S. Robinson, R. Barron, D. O. Caplan, M. L. Stevens, J. J. Carney, S. A. Hamilton, W. E. Keicher, J. K. W. Yang, K. Rosfjord, V. Anant, and K. Bergren, "Superconducting nanowire photon-counting detectors for optical communications," *Lincoln Laboratory Journal*, vol. 16, pp. 217-24, 2006.

[44] Hamamatsu, "Concepts in digital imaging technology." http://learn.hamamatsu.com/articles/ccdanatomy.html, 2007.

[45] B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*. New York: Wiley, 1991.

[46] M. S. Robbins and B. J. Hadwen, "The noise performance of electron multiplying charge-coupled devices," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1227-1232, 2003.

[47] Andor, "ICCD detectors," http://www.lot-oriel.com/site/site_down/cc_notesiccd_deen.pdf, Ed., 2003.

[48] Q. Y. Fang, T. Papaioannou, J. A. Jo, R. Vaitha, K. Shastry, and L. Marcu, "Time-domain laser-induced fluorescence spectroscopy apparatus for clinical diagnostics," *Review of Scientific Instruments*, vol. 75, pp. 151-162, 2004.

[49] W. Enloe, R. Sheldon, L. Reed, and A. Amith, "An electron-bombarded CCD image intensifier with a GaAs photocathode," presented at Proceedings of the SPIE - The International Society for Optical Engineering, vol.1655, 1992, pp. 41-9. USA.

[50] S. Cova, A. Longoni, and A. Andreoni, "Towards picosecond resolution with single-photon avalanche-diodes," *Review of Scientific Instruments*, vol. 52, pp. 408-412, 1981.

[51] K. G. Mckay and A. G. Chynoweth, "Optical studies of avalanche breakdown in silicon," *Physical Review*, vol. 99, pp. 1648-1648, 1955.

[52] S. L. Miller, "Avalanche breakdown in Germanium," *Physical Review*, vol. 99, pp. 1234-1241, 1955.

[53] A. G. Chynoweth, "Ionization rates for electrons and holes in silicon," *Physical Review*, vol. 109, pp. 1537-1540, 1958.

[54] R. H. Haitz and A. Goetzberger, "Avalanche noise study in microplasmas and uniform junctions," *Solid-State Electronics*, vol. 6, pp. 678-&, 1963.

[55] J. L. Moll and R. Vanoverstraeten, "Charge multiplication in sSilicon p-n junctions," *Solid-State Electronics*, vol. 6, pp. 147-157, 1963.

[56] S. M. Sze and G. Gibbons, "Effect of junction curvature on breakdown voltage in semiconductors," *Solid-State Electronics*, vol. 9, pp. 831-&, 1966.

[57] R. H. Haitz, "Model for electrical behavior of microplasma," *Journal of Applied Physics*, vol. 35, pp. 1370-&, 1964.

[58] H. Dautet, P. Deschamps, B. Dion, A. D. Macgregor, D. Macsween, R. J. Mcintyre, C. Trottier, and P. P. Webb, "Photon-counting techniques with silicon avalanche photodiodes," *Applied Optics*, vol. 32, pp. 3894-3900, 1993.

[59] F. Zappa, S. Tisa, S. Cova, P. Maccagnani, D. Bonaccini Calia, G. Bonanno, M. Belluso, R. Saletti, and R. Roncella, "Pushing technologies: single-photon avalanche diode arrays," presented at Advancements in Adaptive Optics, 2004.

[60] B. F. Aull, A. H. Loomis, D. J. Young, R. M. Heinrichs, B. J. Felton, P. J. Daniels, and D. J. Landers, "Geiger-mode avalanche photodiodes for three-dimensional imaging," *Lincoln Laboratory Journal*, vol. 13, pp. 335-50, 2002.

[61] B. Aull, "3D imaging with Geiger-mode avalanche photodiodes," *Optics & Photonics News*, vol. 16, pp. 42-6, 2005.

[62] A. Goetzberger, R. M. Scarlett, R. H. Haitz, and B. Mcdonald, "Avalanche effects in silicon p-n junctions .2. Structurally perfect junctions," *Journal of Applied Physics*, vol. 34, pp. 1591-1600, 1963.

[63] S. Cova, M. Ghioni, A. Lotito, I. Rech, and F. Zappa, "Evolution and prospects for single-photon avalanche diodes and quenching circuits," *Journal of Modern Optics*, vol. 51, pp. 1267-1288, 2004.

[64] A. Lacaita, M. Ghioni, and S. Cova, "Double epitaxy improves single-photon avalanche-diode performance," *Electronics Letters*, vol. 25, pp. 841-843, 1989.

[65] D. M. Taylor, J. C. Jackson, A. P. Morrison, A. Mathewson, and J. G. Rarity, "Characterization of novel active area silicon avalanche photodiodes operating in the Geiger mode," *Journal of Modern Optics*, vol. 51, pp. 1323-1332, 2004.

[66] A. Spinelli, M. A. Ghioni, S. D. Cova, and L. M. Davis, "Avalanche detector with ultraclean response for time-resolved photon counting," *IEEE Journal of Quantum Electronics*, vol. 34, pp. 817-821, 1998.

[67] E. Sciacca, S. Lombardo, M. Mazzillo, G. Condorelli, D. Sanfilippo, A. Contissa, M. Belluso, F. Torrisi, S. Billotta, A. Campisi, L. Cosentino, A. Piazza, G. Fallica, P. Finocchiaro, F. Musumeci, S. Privitera, S. Tudisco, G. Bonanno, and E. Rimini, "Arrays of Geiger mode avalanche photodiodes," *IEEE Photonics Technology Letters*, vol. 18, pp. 1633-1635, 2006.

[68] A. Rochas, A. R. Pauchard, P. A. Besse, D. Pantic, Z. Prijic, and R. S. Popovic, "Low-noise silicon avalanche photodiodes fabricated in conventional CMOS technologies," *IEEE Transactions on Electron Devices*, vol. 49, pp. 387-394, 2002.

[69] A. Rochas, M. Gani, B. Furrer, P. A. Besse, R. S. Popovic, G. Ribordy, and N. Gisin, "Single photon detector fabricated in a complementary metal-oxide-semiconductor high-voltage technology," *Review of Scientific Instruments*, vol. 74, pp. 3263-3270, 2003.

[70] A. Rochas, "Single photon avalanche diodes in CMOS technology," EPFL, 2003.

[71] C. Niclass, M. Sergio, and E. Charbon, "A single photon avalanche diode array fabricated in deep-submicron CMOS technology," presented at 2006 Design, Automation and Test in Europe (IEEE Cat. No. 06EX1285C). IEEE. 2006, pp. 6. Piscataway, NJ, USA., 2006.

[72] C. Niclass, A. Rochas, P. A. Besse, R. Popovic, and E. Charbon, "A 4 mu s integration time imager based on CMOS single photon avalanche diode technology," *Sensors and Actuators a-Physical*, vol. 130, pp. 273-281, 2006.

[73] C. Niclass, A. Rochas, P. A. Besse, and E. Charbon, "Toward a 3-D camera based on single photon avalanche diodes," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 10, pp. 796-802, 2004.

[74] C. Niclass and E. Charbon, "A single photon detector array with 64x64 resolution and millimetric depth accuracy for 3D imaging," presented at 2005 IEEE International Solid-State Circuits Conference (IEEE Cat. No. 05CH37636). IEEE. Part Vol. 1, 2005, pp. 364-604 Vol. 1. Piscataway, NJ, USA., 2005.

[75] S. Tisa, A. Tosi, and F. Zappa, "Fully-integrated CMOS single photon counter," *Optics Express*, vol. 15, pp. 2873-2887, 2007.

[76] C. J. Stapels, W. G. Lawrence, F. L. Augustine, and J. F. Christian, "Characterization of a CMOS Geiger photodiode pixel," *IEEE Transactions on Electron Devices*, vol. 53, pp. 631-635, 2006.

[77] E. Charbon, "CMOS single photon detectors." http://si.epfl.ch/webdav/site/si/shared/QuantumDevices_Charbon.pdf, 2007.

[78] A. S. Tremsin, J. V. Vallerga, O. H. W. Siegmund, C. P. Beetz, and R. W. Boerstler, "The latest developments of high gain Si microchannel plates," presented at Future EUV/UV and Visible Space Astrophysics Missions and Instrumentation, Hawaii, 2002.

[79] A. Gulinatti, P. Maccagnani, I. Rech, M. Ghioni, and S. Cova, "35 ps time resolution at room temperature with large area single photon avalanche diodes," *Electronics Letters*, vol. 41, pp. 272-274, 2005.

[80] H. Luo, D. Ban, H. C. Liu, A. J. Springthorpe, Z. R. Wasilewski, M. Buchanan, and R. Glew, "1.5 μm to 0.87 μm optical upconversion using wafer fusion technology," *Journal of Vacuum Science & Technology A*, vol. 22, pp. 788-791, 2004.

[81] E. Diamanti, H. Takesue, C. Langrock, M. M. Fejer, and Y. Yamamoto, "100 km differential phase shift quantum key distribution experiment with low jitter up-conversion detectors," *Optics Express*, vol. 14, pp. 13073-13082, 2006.

[82] M. A. Albota and F. N. C. Wong, "Efficient single-photon counting at 1.55 μm by means of frequency upconversion," *Optics Letters*, vol. 29, pp. 1449-1451, 2004.

[83] A. P. VanDevender and P. G. Kwiat, "Quantum transduction via frequency upconversion (Invited)," *Journal of the Optical Society of America B-Optical Physics*, vol. 24, pp. 295-299, 2007.

[84] C. Langrock, E. Diamanti, R. V. Roussev, Y. Yamamoto, M. M. Fejer, and H. Takesue, "Highly efficient single-photon detection at communication wavelengths by use of upconversion in reverse-proton-exchanged periodically poled LiNbO3 waveguides," *Optics Letters*, vol. 30, pp. 1725-1727, 2005.

[85] S. M. Sze, *Physics of Semiconductor Devices*, 2 ed. Canada, 2004.

[86] G. A. M. Hurkx, "On the modeling of tunnelling currents in reverse-biased p-n-junctions," *Solid-State Electronics*, vol. 32, pp. 665-668, 1989.

[87] J. J. Liou, "Modeling the tunneling current in reverse-biased p/n junctions," *Solid-State Electronics*, vol. 33, pp. 971-972, 1990.

[88] W. Fulop, "Calculation of avalanche breakdown voltages of silicon p-n junctions," *Solid-State Electronics*, vol. 10, pp. 39-43, 1967.

[89] W. N. Grant, "Electron and hole ionization rates in epitaxial silicon at high electric-fields," *Solid-State Electronics*, vol. 16, pp. 1189-1203, 1973.

[90] S. M. Sze and G. Gibbons, "Avalanche breakdown voltages of abrupt and linearly graded p-n junctions in Ge Si GaAs and GaP," *Applied Physics Letters*, vol. 8, pp. 111-&, 1966.

[91] C. Y. Chang and S. M. Sze, *ULSI technology*. New York: McGraw-Hill, 1996.

[92] R. J. Locker and G. C. Huth, "A New ionizing radiation detection concept which employs semiconductor avalanche amplification and tunnel diode element," *Applied Physics Letters*, vol. 9, pp. 227-&, 1966.

[93] E. Gramsch, M. Szawlowski, S. Zhang, and M. Madden, "Fast, high-density avalanche photodiode-array," *IEEE Transactions on Nuclear Science*, vol. 41, pp. 762-766, 1994.

[94] J. P. Donnelly, K. A. McIntosh, D. C. Oakley, A. Napoleone, S. H. Groves, S. Vernon, L. J. Mahoney, K. Molvar, J. Mahan, J. C. Aversa, E. K. Duerr, Z. L. Liau, B. F. Aull, and D. C. Shaver, "1-mum Geiger-mode detector development," presented at Proceedings of the SPIE, vol.5791, no.1, 2005, pp. 286-92. USA., 2005.

[95] H. Yilmaz, "Optimization and surface-charge sensitivity of high-voltage blocking structures with shallow junctions," *IEEE Transactions on Electron Devices*, vol. 38, pp. 1666-1675, 1991.

[96] K. Ranjan, A. Bhardwaj, Namrata, S. Chatterji, A. K. Srivastava, and R. K. Shivpuri, "Analysis and optimal design of Si microstrip detector with overhanging metal electrode," *Semiconductor Science and Technology*, vol. 16, pp. 635-639, 2001.

[97] R. H. Haitz, A. Goetzberger, R. M. Scarlett, and W. Shockley, "Avalanche effects in silicon p-n junctions .1. Localized photomultiplication studies on microplasmas," *Journal of Applied Physics*, vol. 34, pp. 1581-&, 1963.

[98] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Applied Optics*, vol. 35, pp. 1956-1976, 1996.

[99] W. G. Oldham, Samuelso.Rr, and Antognet.P, "Triggering Phenomena in Avalanche-Diodes," *IEEE Transactions on Electron Devices*, vol. ED19, pp. 1056-&, 1972.

[100] W. Becker, *TCSPC Handbook*, 2 ed. Berlin: Becker & Hickl GmbH, 2006.

[101] S. Felekyan, R. Kuhnemuth, V. Kudryavtsev, C. Sandhagen, W. Becker, and C. A. M. Seidel, "Full correlation from picoseconds to seconds by time-resolved and time-correlated single photon detection," *Review of Scientific Instruments*, vol. 76, pp. 083104-083104-14 , 2005.

[102] H. Finkelstein, M. J. Hsu, and S. Esener, "An ultrafast Geiger-mode single-photon avalanche diode in 0.18-µm CMOS technology," presented at Advanced Photon Counting Techniques, Boston, MA, 2006.

[103] H. Finkelstein, M. J. Hsu, and S. C. Esener, "STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology," *IEEE Electron Device Letters*, vol. 27, pp. 887-889, 2006.

[104] A. Rochas, G. Ribordy, B. Furrer, P. A. Besse, and R. S. Popovic, "First passively-quenched single photon counting avalanche photodiode element integrated in a conventional CMOS process with 32ns dead time," presented at SPIE-Int. Soc. Opt. Eng. Proceedings of the SPIE - The International Society for Optical Engineering, vol.4833, 2002, pp. 107-15. USA.

[105] S. Cova, A. Longoni, and G. Ripamonti, "Active-quenching and gating circuits for single-photon avalanche-diodes (SPADs)," *IEEE Transactions on Nuclear Science*, vol. 29, pp. 599-601, 1982.

[106] A. Gallivanoni, I. Rech, D. Resnati, M. Ghioni, and S. Cova, "Monolithic active quenching and picosecond timing circuit suitable for large-area single-photon avalanche diodes," *Optics Express*, vol. 14, pp. 5021-5030, 2006.

[107] I. Rech, G. B. Luo, M. Ghioni, H. Yang, X. L. S. Xie, and S. Cova, "Photon-timing detector module for single-molecule spectroscopy with 60-ps resolution," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 10, pp. 788-795, 2004.

[108] A. Rochas, P. A. Besse, and R. S. Popovic, "Actively recharged single photon counting avalanche photodiode integrated in an industrial CMOS process," *Sensors and Actuators a-Physical*, vol. 110, pp. 124-129, 2004.

[109] F. Zappa, A. Gulinatti, P. Maccagnani, S. Tisa, and S. Cova, "SPADA: Single-photon avalanche diode arrays," *IEEE Photonics Technology Letters*, vol. 17, pp. 657-659, 2005.

[110] D. Bisello, Y. Gotra, V. Jejer, V. Kushpil, N. Malakhov, A. Paccagnella, Z. Sadygov, I. Stavitsky, and E. Tsyganov, "Silicon avalanche detectors with negative feedback as detectors for high energy physics," *Nuclear Instruments & Methods in Physics Research Section a-Accelerators Spectrometers Detectors and Associated Equipment*, vol. 367, pp. 212-214, 1995.

[111] V. Golovin and V. Saveliev, "Novel type of avalanche photodetector with Geiger mode operation," *Nucl. Instr. Meth. Phys. Res. A*, vol. 518, pp. 560-564, 2004.

[112] O. S. Heavens, *Optical properties of thin solid films*. London: Butterworths scientific publications, 1955.

[113] R. F. Wolffenbuttel, "Integrated silicon color sensors," Delft University, 1998.

[114] IBM-Corporation, "CMOS7RF design manual." Essex Junction, VT: Mixed Signal Technology Development Department, IBM Microelectronics Division, 2005.

[115] "Optical properties of silicon," Virginia Semiconductor Inc.

[116] P. Ossieur, X. Z. Qiu, J. Bauwelinck, and J. Vandewege, "Sensitivity penalty calculation for burst-mode receivers using avalanche photodiodes," *Journal of Lightwave Technology*, vol. 21, pp. 2565-2575, 2003.

[117] S. M. Sze and W. Shockley, "Unit-cube expression for space-charge resistance," *Bell System Technical Journal*, vol. 46, pp. 837-+, 1967.

[118] w. Shockley and W. T. Read, "Statistics of the recombinations of holes and electrons," *Physical Review*, vol. 87, pp. 835-842, 1952.

[119] G. A. M. Hurkx, D. B. M. Klaassen, and M. P. G. Knuvers, "A new recombination model for device simulation including tunneling," *IEEE Transactions on Electron Devices*, vol. 39, pp. 331-338, 1992.

[120] W. J. Kindt and H. W. van Zeijl, "Modeling and fabrication of Geiger mode avalanche photodiodes," *IEEE Transactions on Nuclear Science*, vol. 45, pp. 715-719, 1998.

[121] S. Cova, A. Lacaita, and G. Ripamonti, "Trapping phenomena in avalanche photodiodes on nanosecond scale," *IEEE Electron Device Letters*, vol. 12, pp. 685-687, 1991.

[122] K. E. Jensen, P. I. Hopman, E. K. Duerr, E. A. Dauler, J. P. Donnelly, S. H. Groves, L. J. Mahoney, K. A. McIntosh, K. M. Molvar, A. Napoleone, D. C. Oakley, S. Verghese, C. J. Vineis, and R. D. Younger, "Afterpulsing in Geiger-mode avalanche photodiodes for 1.06 μm wavelength," *Applied Physics Letters*, vol. 88, pp. 133503, 2006.

[123] M. Ghioni, A. Giuduce, S. Cova, and F. Zappa, "High-rate quantum key distribution at short wavelength: performance analysis and evaluation of

silicon single photon avalanche diodes," *Journal of Modern Optics*, vol. 50, pp. 2251-2269, 2003.

[124] A. Lacaita, S. Cova, A. Spinelli, and F. Zappa, "Photon-assisted avalanche spreading in reach-through photodiodes," *Applied Physics Letters*, vol. 62, pp. 606-608, 1993.

[125] A. Lacaita and M. Mastrapasqua, "Strong dependence of time resolution on detector diameter in single photon avalanche-diodes," *Electronics Letters*, vol. 26, pp. 2053-2054, 1990.

[126] A. Lacaita, M. Mastrapasqua, M. Ghioni, and S. Vanoli, "Observation of avalanche propagation by multiplication assisted diffusion in p-n-junctions," *Applied Physics Letters*, vol. 57, pp. 489-491, 1990.

[127] I. Rech, I. Labanca, M. Ghioni, and S. Cova, "Modified single photon counting modules for optimal timing performance," *Review of Scientific Instruments*, vol. 77, pp. 033104-033104-5, 2006.

[128] J. Philip and K. Carlsson, "Theoretical investigation of the signal-to-noise ratio in fluorescence lifetime imaging," *Journal of the Optical Society of America a-Optics Image Science and Vision*, vol. 20, pp. 368-379, 2003.

[129] A. L. Lacaita, F. Zappa, S. Bigliardi, and M. Manfredi, "On the Bremsstrahlung origin of hot-carrier-induced photons in silicon devices," *IEEE Trans. Electron Dev.*, vol. 40, pp. 577-582, 1993.

[130] H. Finkelstein, M. Gross, Y.-H. Lo, and S. C. Esener, "Analysis of hot-carrier luminescence for infrared single-photon up-conversion and readout," *IEEE Journal of Special Topics in Quantum Electronics - Single Photon Detection*, vol. 13, 2007.

[131] S. Yamada and M. Kitao, "Recombination radiation as possible mechanism of light-emission from reverse-biased p-n-junctions under breakdown condition," *Japanese Journal of Applied Physics Part 1*, vol. 32, pp. 4555-4559, 1993.

[132] A. G. Chynoweth and K. G. Mckay, "Photon emission from avalanche breakdown in silicon," *Physical Review*, vol. 102, pp. 369-376, 1956.

[133] F. Zappa, M. Ghioni, S. Cova, L. Varisco, B. Sinnis, A. Morrison, and A. Mathewson, "Integrated array of avalanche photodiodes for single-photon counting," presented at ESSDERC '97. Proceedings of the 27th European Solid-State Device Research Conference. Editions Frontieres. 1997, pp. 600-3. Paris, France.

[134] R. J. McIntyre, "A new look at impact ionization - Part I: A theory of gain, noise, breakdown probability, and frequency response," *IEEE Transactions on Electron Devices*, vol. 46, pp. 1623-1631, 1999.

[135] *CMOS7RF (CMRF7SF) Design Manual*. Essex Junction, VT: Analog and Mixed Signal Technology Development, IBM Microelectronics Division, 2005.

[136] M. Gosch, A. Serov, T. Anhut, T. Lasser, A. Rochas, P. A. Besse, R. S. Popovic, H. Blom, and R. Rigler, "Parallel single molecule detection with a fully integrated single-photon 2X2 CMOS detector array," *Journal of Biomedical Optics*, vol. 9, pp. 913-921, 2004.

[137] F. Zappa, A. Lotito, A. C. Giudice, S. Cova, and M. Ghioni, "Monolithic active-quenching and active-reset circuit for single-photon avalanche detectors," *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 1298-1301, 2003.

[138] X. Michalet, T. D. Lacoste, and S. Weiss, "Ultrahigh-resolution colocalization of spectrally separable point-like fluorescent probes," *Methods*, vol. 25, pp. 87-102, 2001.

[139] Z. C. Wu, E. T. Arakawa, J. R. Jimenez, and L. J. Schowalter, "Optical-Properties of Epitaxial Cosi2/Si and Cosi2 Particles in Si from 0.062 to 2.76 eV," *Journal of Applied Physics*, vol. 71, pp. 5601-5605, 1992.

[140] A. L. Beck, G. Karve, S. Wang, J. Ming, X. Guo, and J. C. Campbell, "Geiger mode operation of ultraviolet 4H-SiC avalanche photodiodes," *IEEE Photonics Technology Letters*, vol. 17, pp. 1507-1509, 2005.

[141] R. H. Haitz, "Variation of junction breakdown voltage by charge trapping," *Physical Review*, vol. 138, pp. A260 LP - A267, 1965.

[142] L. Q. Li and L. M. Davis, "Single-photon avalanche-diode for single-molecule detection," *Review of Scientific Instruments*, vol. 64, pp. 1524-1529, 1993.

[143] S. Cova, A. Lacaita, M. Ghioni, G. Ripamonti, and T. A. Louis, "20-ps timing resolution with single-photon avalanche-diodes," *Review of Scientific Instruments*, vol. 60, pp. 1104-1110, 1989.

[144] F. Zappa, M. Ghioni, S. Cova, C. Samori, and A. C. Giudice, "An integrated active-quenching circuit for single-photon avalanche diodes," *IEEE Transactions on Instrumentation and Measurement*, vol. 49, pp. 1167-1175, 2000.

[145] D. S. Bethune, W. P. Risk, and G. W. Pabst, "A high-performance integrated single-photon detector for telecom wavelengths," *Journal of Modern Optics*, vol. 51, pp. 1359-1368, 2004.

[146] T. Maruyama, F. Narusawa, M. Kudo, M. Tanaka, Y. Saito, and A. Nomura, "Development of a near-infrared photon-counting system using an InGaAs avalanche photodiode," *Optical Engineering*, vol. 41, pp. 395-402, 2002.

[147] G. Ribordy, N. Gisin, O. Guinnard, D. Stucki, M. Wegmuller, and H. Zginden, "Photon counting at telecom wavelengths with commercial InGaAs/InP avalanche photodiodes: current performance," *Journal of Modern Optics*, vol. 51, pp. 1381-1398, 2004.

[148] S. E. Steen, M. K. McManus, and D. G. Manzer, "Timing high-speed microprocessor circuits using picosecond imaging circuit analysis," presented at High-Speed Imaging and Sequence Analysis III, 2001.

[149] A. J. Annunziata, A. Frydman, M. O. Reese, L. Frunzio, M. Rooks, and D. E. Prober, "Superconducting niobium nanowire single photon detectors," presented at Advanced Photon Counting Techniques, Boston, MA, 2006.

[150] S. Pellegrini, R. E. Warburton, L. J. J. Tan, J. S. Ng, A. B. Krysa, K. Groom, J. P. R. David, S. Cova, M. J. Robertson, and G. S. Buller, "Design and performance of an InGaAs-InP single-photon avalanche diode detector," *IEEE Journal of Quantum Electronics*, vol. 42, pp. 397-403, 2006.

[151] Y. Kang, Y. H. Lo, M. Bitter, S. Kristjansson, Z. Pan, and A. Pauchard, "InGaAs-on-Si single photon avalanche photodetectors," *Applied Physics Letters*, vol. 85, pp. 1668-1670, 2004.

[152] P. A. Hiskett, G. S. Buller, A. Y. Loudon, J. M. Smith, I. Gontijo, A. C. Walker, P. D. Townsend, and M. J. Robertson, "Performance and design of InGaAs/InP photodiodes for single-photon counting at 1.55 μm," *Applied Optics*, vol. 39, pp. 6818-6829, 2000.

[153] K. Zhao, A. Zhang, K. Y., and Y.-H. Lo, "InGaAs/InP MOS single photon detector," presented at Digest of the LEOS Summer Topical Meetings, 2006.

[154] J. P. Donnelly, E. K. Duerr, K. A. McIntosh, E. A. Dauler, D. C. Oakley, S. H. Groves, C. J. Vineis, L. J. Mahoney, K. M. Molvar, P. I. Hopman, K. E. Jensen, G. M. Smith, S. Verghese, and D. C. Shaver, "Design considerations for 1.06-μm InGaAsP-InP Geiger-mode avalanche photodiodes," *IEEE Journal of Quantum Electronics*, vol. 42, pp. 797-809, 2006.

[155] J. Wei, S. M. L. Nai, C. K. S. Wong, Z. Sun, and L. C. Lee, "Low temperature glass-to-glass wafer bonding," *IEEE Transactions on Advanced Packaging*, vol. 26, pp. 289-294, 2003.

[156] R. Newman, "Visible light from a silicon p-n junction," *Physical Review*, vol. 100, pp. 700-703, 1955.

[157] M. Lahbabi, A. Ahaitouf, M. Fliyou, E. Abarkan, J. P. Charles, A. Bath, A. Hoffmann, S. E. Kerns, and D. V. Kerns, "Analysis of electroluminescence spectra of silicon and gallium arsenide p-n junctions in avalanche breakdown," *Journal of Applied Physics*, vol. 95, pp. 1822-1828, 2004.

[158] C. Kurtsiefer, P. Zarda, S. Mayer, and H. Weinfurter, "The breakdown flash of silicon avalanche photodiodes-back door for eavesdropper attacks?," *Journal of Modern Optics*, vol. 48, pp. 2039-2047, 2001.

[159] A. L. Lacaita, F. Zappa, S. Bigliardi, and M. Manfredi, "On the Bremsstrahlung Origin of Hot-Carrier-Induced Photons in Silicon Devices," *IEEE Transactions on Electron Devices*, vol. 40, pp. 577-582, 1993.

[160] C. Y. Huang, J. L. Hwang, C. H. Y, and Y. K. Tai, "Counting efficiency of nuclear multiplate camera," *Chinese Journal of Physics*, vol. 1, pp. 33-38, 1963.

[161] A. Rochas, A.Pauchard, L.Monat, A.Matteo, P.Trinkler, R.Thew, and R.Ribordy, "Ultra-compact CMOS single photon detector," presented at Advanced Photon Counting Techniques, Boston, MA, 2006.

[162] H. Burkhard, H. E. Dinges, and E. Kuphal, "Dielectric functions and optical parameters of Si, Ge, GaP, GaAs, GaSb, InP, InAs, and InSb from 1.5 to 6.0 eV," *J. Appl. Phys.*, vol. 53, pp. 655-662, 1982.

[163] Princeton_Instruments, "Pixis 400 data sheet," 2007.

[164] G. E. Stillman, V. M. Robbins, and N. Tabatabaie, "III-V compound semiconductor-devices - Optical detectors," *IEEE Transactions on Electron Devices*, vol. 31, pp. 1643-1655, 1984.

[165] H. Finkelstein and S. Esener, "Shallow-trench-isolation (STI)-bounded single-photon CMOS photodetector." U.S.: University of California, 2006.

[166] J. Jansson, A. Mantyniemi, and J. Kostamovaara, "A delay line based CMOS time digitizer IC with 13 ps single-shot precision," presented at Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on, 2005.

[167] C. Niclass, M. Sergio, and E. Charbon, "A CMOS 64x48 single photon avalanche diode array with event-driven readout," presented at ESSCIRC 2006. Proceedings of the 32nd European Solid-State Circuits Conference (IEEE Cat. No. 06EX1347). IEEE. 2006, pp. 556-9. Piscataway, NJ, USA.