Silicon Photomultipliers for Scintillation Detection Systems

by

Paul Joseph Barton

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Nuclear Engineering and Radiological Sciences)
in the University of Michigan
2011

Doctoral Committee:

Professor David K. Wehe, Chair
Professor Emeritus A. Ziyaeddin Akcasu
Professor Frederick D. Becchetti, Jr.
Research Scientist Mark D. Hammig

Acknowledgments

I would first like to thank my wife Katie, whose patience, support, and understanding have seen me through every turn of my graduate career. In a very close second is my advisor David Wehe who tackled each encounter as an educational opportunity and always left me asking the right questions. The freedom afforded me to simply *learn* has made a profound impact on my ability to tackle problems from non-traditional angles. I have come to appreciate that the supportive environment engendered by the staff, faculty and students in the department is unique and one for which I'll no doubt swiftly become nostalgic. I cannot say enough about the tremendous learning opportunities available at the Lurie Nanofabrication Facility. I learned more about the "black art" of semiconductor fabrication by positioning myself among the staff than any book or class could teach. Conversations in the middle of the night with colleagues from external disciplines provided some of the greatest mutual insight, and I owe a debt to everyone who helped teach me, and whom I helped teach. Finally, thank you to those students who preceded me. Wonho, Jason, Ben, and Haori; you each illuminated different subtleties of detector research that will nourish me throughout my career.

Table of Contents

ACKNOWLEDGMENTS .... **ii**

LIST OF FIGURES .... **vi**

LIST OF APPENDICES .... **xiii**

CHAPTER 1 INTRODUCTION TO SOLID-STATE PHOTODETECTION **1**

1.1. THE SILICON PHOTOMULTIPLIER 3
1.2. WHY SILICON? 4
1.3. MOTIVATION 5
1.4. SOLID-STATE GAIN 7
1.5. IMPACT IONIZATION 8
1.6. THE ONE-SIDED JUNCTION DIODE 9
1.7. SCINTILLATION ABSORPTION 16
1.8. DIODE EDGE BREAKDOWN 20
1.9. EDGE BREAKDOWN PREVENTION 24
1.9.A. Field Plates 24
1.9.B. Guard Rings 25
1.9.C. Junction Termination Extension 26
1.9.D. Bevel Etching 27
1.10. SUMMARY 27

CHAPTER 2 AVALANCHE DIODE FABRICATION **29**

2.1. DOPING PROCESSES 30
2.2. GETTERING 32
2.2.A. Gettering Test Structure 33
2.2.B. Gettering Process 34
2.2.C. Gettering Results 36
2.3. JUNCTION TERMINATION EXTENSION DIODES 39
2.3.A. Fabricated JTE Diode Results 42
2.3.B. Summary 49
2.4. BEVEL ISOLATION DIODES 49
2.4.A. Silicon Etching 53

2.4.B. Chemical Etchant Requirements *54*
2.4.C. TMAH Process *55*
2.4.D. Bevel Etched Diode Results *61*
2.4.E. Bevel Diodes Incorporating JTE Structure *67*
2.4.F. Summary *73*
2.5. GUARD RING DIODES 74
2.5.A. Guard Ring Results *75*
2.5.B. Breakdown Mapping *78*
2.5.C. Light Response *84*
2.5.D. Geiger Mode Behavior *85*
2.5.E. Quench Resistor Selection *86*
2.5.F. Multiple Breakdown *91*
2.5.G. Summary *93*
2.6. TRANSPARENT QUENCH RESISTORS 94
2.6.A. ITO Resistor Results *97*

CHAPTER 3 STOCHASTIC MODEL OF THE SIPM 101

3.1. SCINTILLATION DETECTION 102
3.2. SINGLE PIXEL FIRING PROBABILITY 103
3.2.A. Non-Uniform Spatial Distribution of Incoming Photons *106*
3.3. MULTIPLE PIXEL MEAN AND VARIANCE 106
3.3.A. Non-Uniform Spatial Distribution *108*
3.4. INCIDENT PHOTON FLUX ESTIMATION 108
3.4.A. Maximum Likelihood Estimator *109*
3.5. VARIANCE IN THE ESTIMATE OF INCIDENT PHOTON FLUX ESTIMATION 111
3.5.A. Cramér-Rao Lower Bound on Estimator Variance *111*
3.5.B. Functional Expansion Method *113*
3.5.C. Linearity of Logarithm Method *114*
3.6. SYSTEMATIC FEATURES 115
3.6.A. Dynamic Range *115*
3.6.B. Quantization Error *116*
3.6.C. Avalanche Excess Noise Error *117*
3.7. SIMULATION EXAMPLE 119
3.8. RECOVERY TIME 121
3.8.A. Monte Carlo Illustration *122*
3.9. NOISE MECHANISMS 124
3.9.A. Thermally Generated Dark Counts *124*
3.9.B. Afterpulsing *126*

3.9.C. Optical Crosstalk *127*
3.10. NUCLEAR COUNTER EFFECTS 128
3.11. A COMPLETE MODEL 129

CHAPTER 4 SIMULATION OF OPTICAL PROCESSES 131

4.1. ANTIREFLECTION COATINGS FOR SCINTILLATION PHOTODETECTORS 131
4.1.A. Incident Angle Distribution *132*
4.1.B. Thin-film Interference Filters *135*
4.2. EXTERNAL OPTICAL CROSSTALK 140

CHAPTER 5 CONCLUSIONS AND FUTURE WORK 142

5.1. SILICON AVALANCHE DIODE FABRICATION 142
5.2. SILICON PHOTOMULTIPLIER STATISTICS 144
5.3. FUTURE WORK 145
5.3.A. Commercial Foundry Run *145*
5.3.B. Passivation *145*
5.3.C. Reliable Ohmic Contacts *145*
5.3.D. Diffusion Masking *146*
5.3.E. VMOS *146*
5.3.F. Variation of Lateral Doping *146*
5.3.G. LOCOS Isolation *147*
5.3.H. Active Quenching and Micro-Optics *147*

APPENDICES 148

REFERENCES 162

List of Figures

Figure 1.1. Electron impact ionization coefficients in Si as a function of the reciprocal electrical field at various temperatures [Sze07]. 9

Figure 1.2. Impurity concentrations for a one-sided phosphorus (N) diffusion into a boron (P) substrate. 10

Figure 1.3. Net charge distribution (N_d + h – N_a – e) for diode at 0 V (red) and 64 V (black). 11

Figure 1.4. Net impurity concentration on the p-type side of the junction at 1.34 µm, illustrating neither an abrupt nor a linearly graded junction. The acceptor impurity concentration of 1×10^{16} is clearly visible at the right. 12

Figure 1.5. Electrostatic potential of one-sided diode at 0 V (red) and 64 V (black) applied bias. 13

Figure 1.6. Electric field for diode at 0 V (red) and 64 V (black) applied bias. 14

Figure 1.7. Local impact ionization rate in electron-hole pairs/cm^3/s at 64 V applied bias. 14

Figure 1.8. Current-voltage characteristic curve of one-sided junction diode, where the positive voltage on the cathode represents a reverse bias. The dominance of the impact ionization current (red) is apparent near the breakdown voltage. 15

Figure 1.9. Current-voltage reverse bias characteristic from Figure 1.8, on a linear scale. 15

Figure 1.10. Optical absorption coefficient for crystalline silicon [Gre95]. 17

Figure 1.11. Scintillation emission spectra (area normalized), for three common scintillators: BGO, CsI:Na, and $LaCl_3$. 18

Figure 1.12. The cumulative probability of photons being absorbed by a certain distance in crystalline silicon, for three separate scintillators. 19

Figure 1.13. Doping profile for n+/p diode with mask edge at 1.5 µm. Contour lines illustrate decades of phosphorus (n+) concentration. Diode junction depth is indicated by green-yellow fill. "CON" is the cathode contact, while the anode contact at the bottom of the substrate. 21

Figure 1.14. Electric field at 40 V applied reverse bias referenced to bottom of substrate for 2D diode (*a*) and 1D diode (*b*). Current flow lines (10% each division) are plotted from cathode to anode. 22

Figure 1.15. Current-voltage characteristic curve for 2D diode with doping profile from Figure 1.13. 23

Figure 1.16. Electroluminescence image (right) of a single-diffusion square n+/p diode operated above breakdown, as imaged by a CMOS camera on a probe station microscope. With increasing bias, this diode breaks down first at its corners, then at its edges. 23

Figure 1.17. Cross section and top view of guard ring diode concept. 25

Figure 1.18. Cross section and top view of junction termination extension concept. 26

Figure 1.19. Cross section and top view of bevel diode concept. 27

Figure 2.1. Solid solubilities for phosphorus and boron in silicon (from TSUPREM4 [TSU07]). 30

Figure 2.2. Experimental design for 8 high resistivity wafers, indicating presence and doping depth of backside polysilicon and four diffusions. 33

Figure 2.3. Structure of simple gettered one-sided diffusion diode. 34

Figure 2.4. Mask patterns for 100 µm wide diffusion, 80 µm wide metal contact, and 60 µm wide via in the oxide passivation layer. 34

Figure 2.5. Reverse bias current-voltage characteristic curve for simple through wafer p+/n diodes in high resistivity silicon. 36

Figure 2.6. Reverse bias leakage current at 20 V reverse bias for the 8 high-resistivity n-type wafers. 37

Figure 2.7. Electroluminescent images for diodes from wafers #5 (a) and #7 (b). 37

Figure 2.8. Electroluminescent images for diodes from wafers #11 (a) and #12 (b). 38

Figure 2.9. Junction termination extension structure (5 µm extension) with shallow n+ diffusion (red lines) extending past a deeper p-well (blue lines) for a junction indicated by the yellow-green fill. 40

Figure 2.10. Electric field at 20 V for diodes with varying degrees of junction termination extension (at right, in µm). Blue bars indicate the shallow diffusion mask region. 41

Figure 2.11. Electric field and current flow lines at 1 V excess bias for the structure in Figure 2.9. 42

Figure 2.12. Two micrographs illustrate 50 µm circular JTE diodes, laterally contacted (left) as in Figure 2.11 and vertically contacted (right) with a single 5 µm via to the center of the diode. The shallow and well diffusions are barely perceptible. 43

Figure 2.13. Reverse bias current-voltage characteristic for a 50 µm circular p+/n-well JTE structure on wafer #7. Extension values (in µm) are listed in the legend. 43

Figure 2.14. Reverse bias current-voltage characteristic for vertically contacted 50 µm circular JTE structure on wafer #3. Extension values (in µm) are listed in the legend. 44

Figure 2.15. Current-voltage characteristic for vertically contacted 10 µm circular JTE diode on wafer #11. Extension values (in µm) are listed in the legend. 45

Figure 2.16. Electroluminescence images for 50 µm square JTE diodes with a -2 µm extension on wafer #11 (a) and #3 (b). Edge breakdown predominates in both. 45

Figure 2.17. Evolution (left to right) of peripheral breakdown clustering of a period of one minute for a 50 µm JTE diode from wafer #11 with -1 µm extension, biased at 15 V. 46

Figure 2.18. Current-voltage characteristic and EL image for vertically contacted 50 µm JTE diode from wafer #3 with circular contact and 8 µm extension. 47

Figure 2.19. Current-voltage characteristic and EL image for vertically contacted 50 µm JTE diode from wafer #3 with 5 µm point contact at edge and 8 µm extension. 47

Figure 2.20. Electroluminescence images from n+/p-well JTE diodes from wafers #3 (top) at 15 V and #11 (bottom) at 20 V. Numbers indicate via size in microns. 48

Figure 2.21. Electroluminescence image for a vertically-contacted p+/n-well JTE diode from wafer #12 with a 8 µm extension and circular contact. 48

Figure 2.22. Doping concentration of a n+/p-well junction, where red and green fill represent boron p-well concentration, and contour lines represent phosphorus n+ concentration. The bevel and subsequent drive-in have created a reduced charge density gradient at the bevel surface. 50

Figure 2.23. Electric potential of a bevel diode junction illustrating the widened potential at the surface. 50

Figure 2.24. Electric field with slight reduction at the surface. Green central contour is region of highest electric field. 51

Figure 2.25. Current paths are illustrated for electrons beginning at 0.5 µm, distributed uniformly through the diode. The enclosed regions at the diode junction indicate regions where impact ionization will lead to sustained avalanches. 51

Figure 2.26. Etched silicon with 55° bevel edge illustrating total internal reflection of hot-carrier emissions arriving parallel to the upper surface and 35° normal to the bevel edge. 52

Figure 2.27. TMAH (25%) silicon etch at 60 °C for 10 min (2.3 µm deep). Original test feature mask (a) with 4 µm wide features in a 1500 Å *thermal* oxide. Microphotographs after etching (b) and after mask removal (c). Undercutting at corners is evident. 57

Figure 2.28. Scanning electron micrograph of 2.4 µm TMAH etch in a thermal oxide mask (removed). Scale marker is 50 µm. Note the presence of undercutting, i.e., non-square corners. 57

Figure 2.29. Scanning electron micrograph illustrating 25% TMAH wet etch and original mask design of area to be etched. Undercutting at convex corners is evident. Scale marker is 10 um. 58

Figure 2.30. Silicon etch rates (µm/min) for 25% TMAH at several temperatures as determined by a Dektak surface profiler on 16 µm wide features. 58

Figure 2.31. AFM trace of flat bevel trench transitioning to bevel edge, indicating 53.7° angle and bevel smoothness. 59

Figure 2.32. AFM image (256 x 256 points) of bevel edge (software leveled) illustrating 3 nm rms bevel smoothness. Note nm-scale in Z axis and µm-scale in X, Y axes. 59

Figure 2.33. TMAH etch (2.4 µm deep) in silicon without surfactant (left), and with 0.1% v/v Triton X-100 surfactant (right). 60

Figure 2.34. Optical micrographs illustrating poor etch rate of boron-doped silicon (left) and successful etching of highly phosphorus-doped silicon (right). 61

Figure 2.35. Structure of bevel diode with both top and bottom substrate contacts. The p+ layer is below the 1×10^{19} cm^{-3} boron etch stop threshold. 62

Figure 2.36. Photomicrograph of the 80 µm wide fabricated bevel diode with top contact to raised island. 62

Figure 2.37. Current-voltage characteristic for diode in Figure 2.36. 63

Figure 2.38. Current-voltage characteristic for diode in Figure 2.36, in series with 100 kΩ resistor, both with and without external photon stimulus. Log scale (top) and linear scale (bottom). 64

Figure 2.39. Temperature coefficient (in mV/°C) for Transys 1N4741A Zener diodes in silicon. ... 65

Figure 2.40. Temperature dependence (3.3 mV/°C) of the breakdown voltage of an 80 µm bevel diode. 66

Figure 2.41. Electroluminescence pattern from vertically contacted 80 µm bevel diode operated at 10 V reverse bias and limited to 10 mA. 67

Figure 2.42. Revised bevel diode structure (top) and series (bottom) with bevel moat etch located at increasing distances from p-well. A p+ substrate contact surrounds the majority of the diode. 68

Figure 2.43. Square bevel diodes from wafer #11, contacted laterally. Distance in µm between bevel and p-well is indicated in legend. 68

Figure 2.44. Hexagonal bevel etched diodes: p+/n-well (left) and n+/p-well (right). Not the slight asymmetry in the etch as the diode active area is longer in the vertical axis. 69

Figure 2.45. Hexagonal bevel diodes from wafer #11, contacted laterally. Distance in µm between bevel and p-well is indicated. 69

Figure 2.46. Steady state characteristic curves for *laterally* contacted n+/p-well bevel diode from wafer #3 (top) and #2 (bottom) with 0 distance between bevel and p-well. Response to light (top) indicates amplifying behavior prior to breakdown. 70

Figure 2.47. Electroluminescent images of *vertically*-contacted bevel diodes from wafer #3 at 10 V reverse bias. Numbers indicate bevel distance from p-well in µm. 72

Figure 2.48. Simulation of electric field and current flow lines for a laterally contacted bevel etch diode. Note the even distribution of current flow lines throughout the active area on the left. 72

Figure 2.49. Electroluminescence images for laterally and vertically contacted n+/p-well bevel diodes from wafers #1,3,9,11, with 0 µm bevel spacing. 73

Figure 2.50. Guard ring design utilizing four diffusions: p-well (p-), n-well (n-), p+, n+. Top view (left) and cross section (right). 75

Figure 2.51. Micrographs of 20 µm guard ring diode and top side anode and cathode contacts and probe pads. The outer non-contacted ring is an n+ diffusion for local gettering. 76

Figure 2.52. Characteristic curves for 20 µm laterally-contacted guard ring diodes on 8 different n-type wafers. 76

Figure 2.53. Characteristic curves for 20 µm laterally-contacted guard ring diodes from wafer #4 and 7 die distributed *vertically* across the wafer. 77

Figure 2.54. Characteristic curves for 20 µm laterally-contacted guard ring diodes from wafer #4 and 7 die distributed *horizontally* across the wafer. 78

Figure 2.55. Electroluminescence images from *laterally* contacted 20 µm guard ring diodes. Wafer numbers indicated. 79

Figure 2.56. Electroluminescence images from *vertically* contacted 20 µm guard ring diodes. Wafer numbers indicated. Magnification is slightly larger than Figure 2.55. 80

Figure 2.57. Characteristic curve for vertically and laterally contacted 20 µm guard ring diode from wafer #4. 81

Figure 2.58. Doping contours at edge of guard ring diode. Blue is boron (p-type), and red is phosphorus (n-type). 82

Figure 2.59. Simulated electric field at 10 V reverse bias for both lateral and vertical contacts. Circled are the regions of greatest electric field. 83

Figure 2.60. Current-voltage characteristic curve for guard ring diode from wafer #3. An increasing photon flux is observed to result in an increased reverse bias leakage current. . 84

Figure 2.61. Current-voltage characteristic from Figure 2.60, only on a linear non-absolute-value scale (left) and magnified (right). 85

Figure 2.62. Transient voltage from a 20 µm (laterally contacted) guard ring diode (wafer #4) with an external 20 kΩ quench resistor in series, as measured by an oscilloscope. 86

Figure 2.63. Current voltage characteristic for guard ring diode from wafer #2, with and without a 20 kΩ external series quench resistor. 87

Figure 2.64. Progression of Geiger mode voltage signal from constant avalanche (200 Ω, top) towards a random telegraph signal (1000 Ω, bottom) for varying values of externally added resistance. Excess bias V_e = 0.2 V. Time scale is 1×10^{-4} s (third plot is time-expanded). 88

Figure 2.65. Progression of Geiger-mode avalanche signals with increasing quench resistance (indicated in Ohms) at V_e = 0.2 V. Time scale is 1×10^{-4} s. 90

Figure 2.66. Micrograph (left) of a 20 µm guard ring diode (wafer #12) and steady-state electroluminescence of laterally contacted diode at -15 V and 10 mA. Multiple breakdown sites are evident: at the center, in a ring, and at the tip of that ring. 91

Figure 2.67. Characteristic curve for diode depicted in Figure 2.66. Note the multiple breakdown voltages at 6 V and 12 V. 92

Figure 2.68. Electroluminescence image of poorly aligned diode indicating premature edge breakdown. 93

Figure 2.69. Sputtered ITO (2000 Å) on Si, patterned by lift off (left) and 1:7 HCl etch (right). Smallest openings are 5 µm in diameter. 96

Figure 2.70. Micrograph illustrating *incomplete* HCl etch of ITO film due to high surface tension. Colors indicate ITO thickness, similar to traditional colors for silicon dioxide thickness. 96

Figure 2.71. Transfer length method structure with 6 metal pads contacting a single strip of resistive material. The 5 resistors are each contacted by equivalent area contacts. 97

Figure 2.72. Combinatorial experimental design schematic (left) for the extraction of contact and sheet resistances for multiple materials (Al, Ti, Si, RITO, and GITO). Actual mask design shown at right. 98

Figure 2.73. Serpentine RITO resistor some ~145 squares long and 5 µm in width. Measured resistivity is 1.9 MΩ. 99

Figure 2.74. Micrograph and cross section illustrating the method for ohmically contacting ITO and oxygen-rich ITO to both silicon and aluminum. 99

Figure 2.75. Current carrying capacity of RITO film stack. Evaporation occurs at the Ti:Al interface at the far left while the GITO, RITO, and Ti-via to the PN junction at the right remain intact. 100

Figure 3.1. Diagram of random variables describing photons incident on a discretized detector. 103

Figure 3.2. The figure above illustrates the relative quantization error for devices with *M* = 20 to 1000 pixels. 117

Figure 3.3. The mean, variance, and relative FWHM (resolution) are displayed for both number of pixels fired *m* (left-hand column) and estimated number of incident photons λ_{est} (right-hand column), as a function of known incident flux (from 1 to 24436). 120

Figure 3.4. The probability distributions for the number of avalanches (from a single pixel) per scintillation event are shown for several different relative recovery times $t_{\text{recovery}}/\tau_{\text{scint}}$, as indicated by the legend entries. Mean incident photon fluxes of $\langle n_k \rangle = 5$ (left) and $\langle n_k \rangle = 1$ (right) are illustrated. 123

Figure 3.5. The mean number of avalanches from the distributions in Figure 3.4. 124

Figure 4.1. Trace of two detected photons from DETECT2000 indicating: component id, reflector finish, position (cm), age (ns), number of surfaces, and detection fate. 132

Figure 4.2. Geometry of cubic scintillator and bottom photodetector indicating one possible path of a single photon interacting at diffusely reflecting boundaries. 133

Figure 4.3. Angular distribution (ray AB from Figure 4.2) of photons refracted across the optical interface of a scintillation crystal covered by diffuse (a) and specular (b) reflectors. 134

Figure 4.4. Angular distribution (ray BC from Figure 4.2) of photons refracted across the scintillator-optical interface boundary for diffuse (a) and specular (b) crystal reflectors. .. 135

Figure 4.5. Destructive optical interference filter schematic where superposition occurs between refracted (blue) and reflected (red) rays. 136

Figure 4.6. Single layer thin-film interference filter (n_2) as a function of *relative* optical thickness for varying incident angles (at 500 nm). 137

Figure 4.7. Single layer thin-film interference filter (n_2) as a function of incident angle for varying *relative* optical thicknesses (at 500 nm). 137

Figure 4.8. Reflectivity for 500 nm scintillation photons exiting a diffusely wrapped cubic crystal (Figure 4.4) and crossing an interference filter (n = 1.8) of specified *relative* optical thickness, as a function of incident angle. 138

Figure 4.9. Integrated reflectivities from Figure 4.8 as a function of *relative* optical thickness with a minimum at 0.33. 139

Figure 4.10. Optical crosstalk mechanisms in a closed optical detection system (e.g. scintillator). 140

Figure B.1. Spreading resistance profiles performed by Solecon Labs for 10 min solid source diffusions at temperatures from 875-1175 °C. Dashed lines indicate best global fit with SUPREM simulations. 158

Figure B.2. Spreading resistance profiles for liquid bubbled phosphorus and boron sources predeposited at 800 °C for 10 min. Dashed profiles are the result of an 1100 °C drive in for 6 hr. 159

List of Appendices

APPENDIX A PROCESS AND DEVICE SIMULATION 149

APPENDIX B PROCESS FLOWS 151

Chapter 1

Introduction to Solid-state Photodetection

Several physical phenomena are known to produce electrical signals in response to energy deposition from ionizing radiation. Each method provides information about the energy of an incident particle by quantizing the energy absorbed into a discrete number of secondary charge carriers. Gaseous radiation detectors create electron-ion pairs which are made to drift inside a capacitor and induce a voltage proportional to the absorbed energy. Semiconductor detectors operate in a similar fashion where incident radiation quanta generate electron-hole pairs whose subsequent drift induces an electrical current. At very low temperatures, bolometers and superconductor detectors discretize incident energy by exciting lattice vibrations (phonons) or by breaking apart Cooper pairs, respectively. Both gas and semiconductor detectors *directly* convert incident radiation into electrical signals.

In contrast, scintillation detectors are transparent crystals which generate optical photons from lattice states excited by external radiation. The detection of these potentially few number of optical photons is then the signal that is proportional to the energy absorbed. Sensitive photodetectors are required to ensure good energy, position, and timing resolutions. In order to optimize these resolutions, the spectral sensitivity, spatial uniformity, and timing characteristics must be optimally designed or selected for a given scintillator detector system. The emission spectrum of scintillators may span several hundred nanometers, and many fast scintillators of interest peak in the blue to UV range (~400 nm or 3 eV) [Bir64]. Absorption of these shorter wavelength photons occurs at relatively shallow depths in many materials; on the order of one micron or less.

The first photodetectors sensitive enough to detect scintillation light were vacuum tubes with a thin photoelectric-converting entrance window. Photoelectrons emitted inside the tube are accelerated towards a metal dynode by a large applied voltage. Enough energy is gained in transit to liberate a small number of electrons upon impact through *secondary electron emission*. With the addition of multiple dynodes at successively higher potentials, additional gain may be obtained so that the resulting electrical charge may be up to 10^6 larger than that of the original photoelectron. These *photomultiplier tubes* (PMTs) have been the primary method of scintillation light readout for many decades. While many vacuum tube devices have given way to their semiconductor counterparts, the PMT with its low-noise amplification and capacity for larger areas, will remain a viable photodetector for the foreseeable future.

Challenging the bulky size and high voltage requirements of the PMT are solid-state devices which are identically tasked with single (photo)electron amplification. The underlying gain mechanism of the PMT is the increase in photoelectron energy and that electron's ability to liberate more than one secondary electron. The solid-state equivalent is a similar physical process termed *impact ionization* where highly energetic photoelectrons liberate secondary electrons from the lattice, which remain *inside* the physical semiconductor boundaries. After several collisions in this branching process, an avalanche of charge ensues which can be anywhere from 10 to 10^6 larger than the original photoelectron population. The challenge for single-photon sensitive solid-state photodetectors is the controlled acceleration of photoelectrons in a solid medium.

On the lower end of avalanche multiplication gain are avalanche photodiodes (APD) which are large area continuous planar diodes that respond to free electrons or holes by initiating impact ionization of some 10 to 1000 additional carriers. The evolution and extinction of this branching process is stochastically complex and exhibits a significant variance in the amount of charge generated from avalanche to avalanche. Many successful APDs have been produced and applied to a variety of applications. As we take the gain higher and higher, the APD begins to operate in a regime akin to the Geiger-Müller gas counter in which

any ionizing particle can cause the entire detector to respond with a single burst of significant charge. Indeed, APDs may be operated in the Geiger mode (GM-APD) if photon counting alone is the application. Very fast GM-APDs are in fact used at telecommunications wavelengths (1550 nm) for LADAR and laser communication [Wil10]. Many nuclear detection applications however also need the energy of incident radiation, thus precluding the use of binary detectors like the GM-APD.

1.1. The Silicon Photomultiplier

If a scintillator is placed on a GM-APD and 100 photons are simultaneously incident, the device will only respond with a single avalanche. This not-so-useful result prompts the question of how to obtain information from a binary device. Now consider 1000 GM-APDs in an array, with the 100 photons being evenly distributed across that array. The result would then be something close to 100 avalanches if the signals from all detectors were summed together. This is the concept behind the silicon photomultiplier: to create a signal proportional to the incident photon flux from a summed array of binary photodetectors. This method necessarily demands that the incoming photons are spread out over the surface of the silicon photomultiplier, and is in direct contrast to the continuous photocathode of the PMT.

The high gain of each GM-APD element (i.e., pixel) contributes to a higher intrinsic detection efficiency than a proportional mode APD, while the large burst of charge allows the use of less sensitive electronic readout circuitry. Along with this added sensitivity comes a much higher dark count rate, since any free electron or hole now has a relatively greater ability to initiate an avalanche. For this reason, the active depth of each diode must be confined to only that of the predicted photoabsorption .

The silicon photomultiplier has its origin in work done by R. McIntyre [Mci61] and R. Haitz [Hai64] in the 1960s with silicon avalanche photodiodes. The history and the majority of the salient operating features of the silicon photomultiplier are discussed in a thorough review article by Dieter Renker of the

Paul Scherrer Institute [Ren09]. The theses of Willem Kindt [Kin99] and Alexis Rochas [Roc03] each provide in-depth insight into the design, production, and characterization of Geiger-mode avalanche photodiodes in commercial CMOS processes. The current incarnations of the silicon photomultiplier have been given a wide variety of acronyms to provide for distinction between manufacturers. A few of these more or less equivalent devices are: metal-resistor-semiconductor (MRS), visible light photon counter (VLPC), solid state photomultiplier (SSPM), silicon photomultiplier (SiPM), with individual pixels also being referred to as: cell, microcell, micropixel, single photon avalanche (photo)diode (SPAD), and Geiger (mode) APD (GAPD/GM-APD).

1.2. Why Silicon?

The design of a sensitive solid-state photodetector begins with the selection of a semiconductor substrate, primarily from band gap considerations. The band gap must be sufficiently low to efficiently convert ~3 eV blue scintillation photons into photoelectrons at reasonable charge multiplication depths. However, the band gap must not be so low that the probability of thermally stimulated electrons is high, since the photodetectors of interest are essentially single *electron* detectors. Other factors like electron and hole trapping and generation sites become more relevant when considering compound semiconductors. In this thesis, the assumption is made that room temperature operation is desirable for most applications, while modest cooling (tens of degrees Celsius) may be acceptable. As a result, this precludes the use of germanium due to its low *room-temperature* band gap of 0.66 eV. While several III-V and II-VI compound semiconductors are theoretically suitable, the 1.12 eV (1100 nm equivalent) bandgap of silicon is particularly well suited to room temperature optical detection of scintillation photons. This is not to say that *solid-state* photomultipliers (SSPMs) manufactured from non-silicon materials are not viable for scintillation detection. In fact, significant progress has been made toward Geiger-mode avalanche photodiodes in 4H-SiC for the development of a UV-sensitive SSPM [Hu08].

Another argument for silicon in solid-state photomultipliers is that silicon is an indirect bandgap semiconductor, whereas compound semiconductors of interest have a direct bandgap. The efficiency with which hot-carrier visible photon emissions are generated is many orders of magnitude greater in direct bandgap materials, which makes them infinitely more suitable for use as LEDs and lasers. The indirect bandgap of silicon requires an inefficient transfer of energy through phonons to generate hot-carrier emissions. Principally, this means that direct bandgap diodes would *emit* many more optical photons above breakdown, which would lead to deleterious optical crosstalk in close-packed arrays of diodes.

An additional argument for silicon is its abundance in commercial electronic devices which has exponentially increased production, purity, and processing capabilities. The success of silicon in general is typically attributed to the availability of high purity raw materials and the ease with which a quality silicon dioxide layer can be grown and patterned for various processing steps. The considerations within this thesis are therefore restricted to the fabrication of silicon photodetectors.

1.3. Motivation

The energy, position, and timing resolutions of scintillation detection systems can all be enhanced by improving the photon collection and detection efficiencies of a given photodetector. While several institutional collaborations and commercial vendors have successfully produced silicon photomultipliers (e.g. RMD, Hamamatsu, Sensl), the fundamental limits of detection efficiency have not been well defined due to guarded intellectual property or the non-translatability of a given semiconductor process flow. It is therefore the aim of this work to improve radiation imaging capabilities by analyzing the fundamental upper limits of silicon photomultiplier photon detection efficiency from the perspectives of silicon processing and the stochastic detection process.

In order to explore the detection limits of silicon photomultipliers, we first turn our attention to the concept of an ideal detector. The ideal photodetector

exists only in theory as a lossless information probe with no noise or fluctuations. A given number of incident photons would then result in a measurement of a scalar value representing exactly that number of photons. The closest realization of this ideal is a detector with a very high detection probability (i.e., large signal) and very low noise. Many opportunities exist for photons and electrons to be stochastically removed from contribution towards a measurement of the true incident photon flux. Various independent and dependent signal and noise processes must be taken into account in order to optimize a silicon photomultiplier's signal to noise ratio (SNR), and ultimately its energy, position or timing resolutions. An outcome of this work is thus the elucidation of those stochastic processes.

In addition to theoretical statistical considerations, specific fabrication efforts have been undertaken to increase the signal, separate from the noise. At the primary optical interface with the outside world is invariably found some transparent window which acts to both passivate the surface and serve as an environmental barrier. This transparent layer can be also made to act as an antireflection coating (ARC) if the incident wavelength(s) and angle(s) are known. This thesis develops the concept of an ideal ARC for scintillation detectors in Chapter 4. Any opaque integrated components on the surface of the silicon photomultiplier will undoubtedly reduce the quantum efficiency with which photons enter the detector. For this reason, transparency at the end of Chapter 2 we investigate the production of thin-film transparent conductors like ITO to increase quench resistor. If the incident wavelength spectrum is known (i.e., scintillation emission spectrum), then the photoabsorption depth is also known, and the photodetection depth sensitivity can be specifically tailored to maximize detection of all incoming photoelectrons (or photoholes). Specific structures are analyzed for several scintillators. Because the silicon photomultiplier is a discrete array of diodes, any area between the diodes is potentially unresponsive to incoming photons. The silicon photomultiplier signal can also be improved by decreasing the pixel pitch. This would require a careful analysis of the efficacy of various pixel isolation techniques in order to minimize the pixel pitch. Further

improvements in dead area minimization can be realized with innovative chip scale optical packaging and through-wafer via technology.

For a complete treatment of silicon photomultiplier SNR, methods have also been investigated with which to lower the intrinsic noise. Due to the limited number of available photons, simply increasing the signal will reduce the relative counting errors which broaden spectral peaks. Energy spectrum features can also be broadened by the ever-present thermally generated dark counts. Depending on the electronic integration time or filter window width, a random number of dark counts will invariably be added to the measured signal. Specific fabrication efforts for reducing these dark counts will be discussed. Afterpulses from the delayed release of electron trap states are another noise feature which require reduction. Increased operating temperatures, cleaner processing, shorter integration times, and smaller avalanches can each improve the noise component arising from afterpulsing. The high gain of Geiger-mode APDs unfortunately comes at the price of additional hot-carrier emissions as visible photons. These photons can trigger adjacent pixels and are an additional source of noise. Fewer hot-carrier emissions will be produced with smaller avalanches. In addition to designing devices with lower hot-carrier emission rates, specific optical isolation structures, like bevelling, can be incorporated to reduce the probability of the triggering probability of these photons.

1.4. Solid-state Gain

Direct measurement of single electrons in solid media at non-cryogenic temperatures is difficult due to the superimposed noise current from random thermal motion of electrons. Some isolated amplification scheme is therefore required to distinguish between the generation of individual electrons. Because what is measured must be fundamentally some flow of electrons, a certain threshold charge is required to positively detect the presence of one photoelectron. Amplification of a single photoelectron requires that electron to be accelerated above some threshold velocity where enough energy can be gained to liberate

additional electrons, thereby creating an ensemble of charge and thus a *measurable* signal.

In the PMT, the photoelectron leaving the photocathode is accelerated in a vacuum over some distance by an electric field. That primary electron then gains enough energy during its flight to strike a metal dynode and cause secondary electrons to be emitted towards a subsequent dynode. This electron avalanche photomultiplication process is very immune to noise, because the probability of spontaneous electron emission from the dynodes is relatively low.

Likewise, in solid-state devices such as the APD or SiPM, a photoelectron, or any electron, in the bulk of a semiconductor is accelerated by the application of an electric field until it gains sufficient energy to liberate bound electrons from the crystalline lattice. These liberated electrons are also accelerated and contribute towards the generation of an easily measurable avalanche of charge.

1.5. Impact Ionization

The theory for impact ionization [McK54, Goe64] has its roots in Townsend's avalanche theory [Loe55] for gas multiplication detectors. The probability that an electron accelerating inside a crystal ionizes a valence band electron on impact is governed by the electric field in which it is accelerated. This probability is characterized by coefficients which define the average number of ionizing collisions per unit length (Figure 1.1). The origin of these coefficients is found in measurements from planar silicon p-n junctions [Bat60, Chy60, Van70, Mae90].

To illustrate, the ionization rate for electrons being accelerated in a field strength of 2×10^5 V/cm in silicon at 300 K is 200 cm^{-1}. This rate indicates that an electron in that field would ionize an average of 0.02 valance band electrons for every micron traveled. Increasing the electric field to 3×10^5 V/cm would increase that electron production rate by two orders of magnitude, to 2 e$^-$/µm. For practical silicon diode depletion widths on the order of microns, this sets the critical electric field required for junction breakdown via impact ionization at $E_{cr} \approx 3\times10^5$ V/cm. The actual values of ionization rates have largely been derived from avalanche structures such as the ones outlined in this work, and therefore,

variation is to be expected for coefficients derived from one type of silicon or another as fabricated from lab to lab.

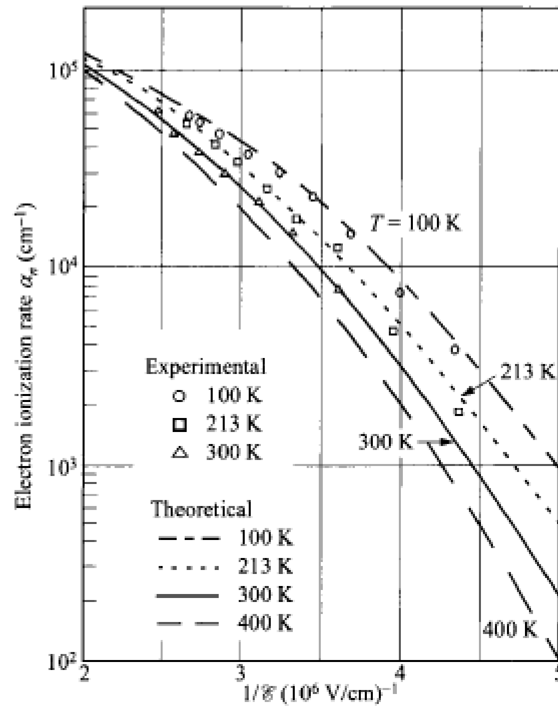


Figure 1.1. Electron impact ionization coefficients in Si as a function of the reciprocal electrical field at various temperatures [Sze07].

While the ionization rate is a smooth function of the electric field, the electric field itself is seldom spatially uniform in realized avalanche diodes. Therefore, the *local* impact ionization rates may vary significantly over the surface and depth of fabricated diodes. This nonuniformity leads away from textbook abrupt junction approximations and is a direct consequence of the method with which dopants are introduced. The significance is that a determination of total electron multiplication requires the varying ionization rates of photoelectrons (or holes) to be integrated over the entire path of acceleration, termed the ionization integral. This requires an exact knowledge of the electric field.

1.6. The One-Sided Junction Diode

In order to further illustrate the behavior of impact ionization, a simple one-sided junction diode is presented in Figure 1.2, wherein a large concentration of phosphorus donor atoms (n+) has been diffused from the cathode side into a

substrate loaded with boron (p-type). This simple diode serves to demonstrate the first requirement of a single-photon-sensitive silicon photodetector, namely a high electric field. This section also serves to illustrate the limitations of analytical formulas for even simple realized diodes. The simulations contained herein are developed from advanced models and codes by the Synopsis corporation. TSUPREM-4 is used to simulation the doping and diffusion processes, while MEDICI is used to simulate the subsequent electrostatics.

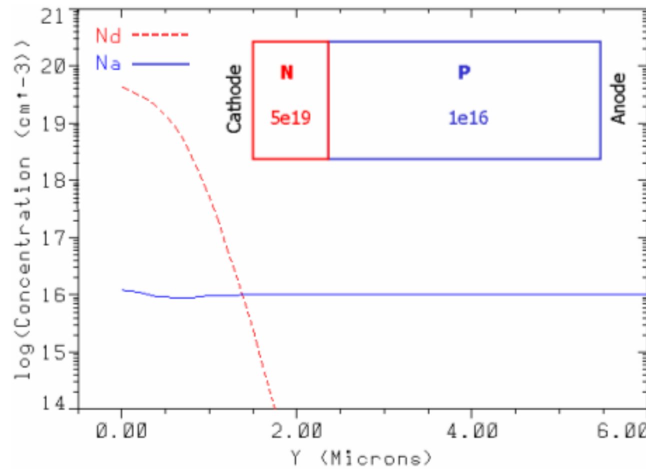


Figure 1.2. Impurity concentrations for a one-sided phosphorus (N) diffusion into a boron (P) substrate.

In silicon processing, the dominant acceptor and donor impurities are boron and phosphorus respectively, while the slower-diffusing donors arsenic and antimony are also employed for various process-specific reasons. It is given here that the n-type cathode side is doped with a surface concentration of 5×10^{19} cm^{-3} atoms of phosphorus, while the substrate is uniformly doped with 1×10^{16} cm^{-3} atoms of boron. While introductory semiconductor physics books tout diffusion profiles as conforming to either Gaussian or complementary error functions, the dynamics of real diffusion processes are typically more complex and can only be accurately simulated numerically, if then. For instance, the diffusion simulation in Figure 1.2 includes coupled solutions of standard diffusion equations with models of point defect-assisted diffusion from lattice vacancies and interstitials.

With knowledge of the doping profile, the net charge distribution can be determined, which will lead directly to the built-in (i.e., 0 V applied bias)

potential and built-in electric field. As illustrated in Figure 1.3, a positive charge occurs on the left where n-type region electrons *diffusing* across the junction leave behind positively charged phosphorus atoms. Similarly, a negative charge is found on the right side from the absence of holes. The net charge density is roughly the sum of the donor impurity and hole concentrations minus the sum of the acceptor impurity and electron concentrations. Because the donor concentration far outweighs the acceptor concentration, $N_d/(N_a+N_d)=99.98\%$ of the depletion region will be contained within the lightly doped p-type side, as observed in Figure 1.3 at 64 V bias.

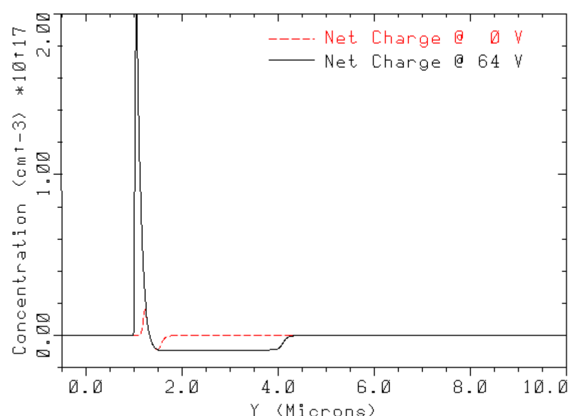


Figure 1.3. Net charge distribution (N_d + h – N_a – e) for diode at 0 V (red) and 64 V (black).

The depletion width for a one-sided abrupt junction is defined by [Sze07]

$$W_D = \sqrt{\frac{2\varepsilon_s}{qN}(\psi_{bi}-V)}\,, \tag{1.1}$$

where *N* represents the lowest doping concentration and *V* is the applied bias. The approximate built-in potential ψ_{bi} is calculated as [Sze07]

$$\psi_{bi} \approx \frac{kT}{q}\ln\left(\frac{N_D N_A}{n_i^2}\right) = 0.94\ \mathrm{V}\,. \tag{1.2}$$

At zero applied bias, the calculated depletion width is 0.35 μm, and at 64 V, the width is predicted to be 2.89 μm. The *simulated* 64 V bias depletion width is approximately 3.2 μm and matches reasonably well with the depletion approximation. If instead, the junction is represented as a linear grading of impurities, then the depletion width is predicted by [Sze07]

$$W_D = \left(\frac{12\varepsilon_s \left(\psi_{bi} - V \right)}{qa} \right)^{1/3}, \tag{1.3}$$

where a is the impurity gradient and is 1.2×10^{20} cm^{-4} in this particular case. Because a graphical determination of a is not inherently obvious from Figure 1.4, it is solved iteratively using the relation $a = N_d / W_d$. This revised formula yields a built-in depletion width of 0.85 μm and a 64 V depletion width of 3.5 μm, which is a slight overestimate compared to the simulation. Looking at the actual impurity gradient, we see that neither the abrupt junction depletion approximation nor the linearly graded junction theory applies particularly well to this simple one-sided diode. As more complicated devices are proposed, this simple diode should serve as a reminder of the importance of numerical simulations and physical validation.

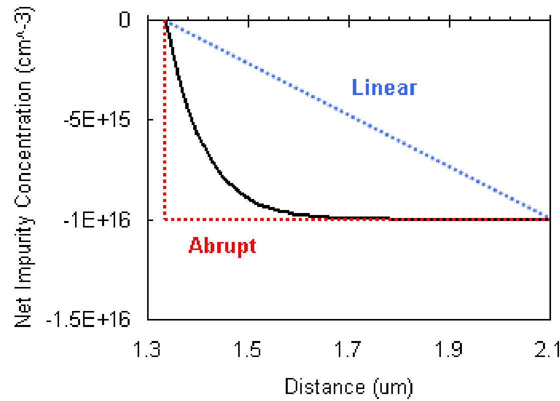


Figure 1.4. Net impurity concentration on the p-type side of the junction at 1.34 μm, illustrating neither an abrupt nor a linearly graded junction. The acceptor impurity concentration of 1×10^{16} is clearly visible at the right.

Having accounted for all charge carriers in the diode, we turn our focus to the electrostatic consequences of any charge gradients. Poisson's equation provides the built-in electrostatic potential (Figure 1.5) through double integration of the net charge density

$$\nabla^2 \psi = -\frac{\rho}{\varepsilon}. \tag{1.4}$$

The regions of charge neutrality are identified in Figure 1.5 as those with flat electrostatic potential. In the absence of an applied bias, the approximate built-in voltage is predicted to be 0.94 V and is close to the simulated value of 0.88 V.

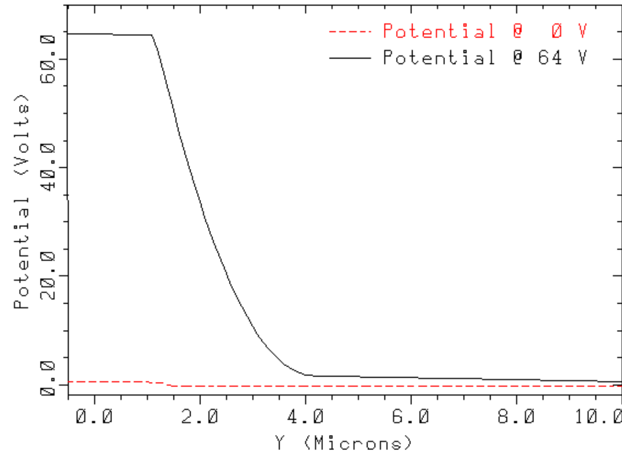


Figure 1.5. Electrostatic potential of one-sided diode at 0 V (red) and 64 V (black) applied bias.

It is important at this point to recall the distinction between electrostatic / electrical potential and *electronic* potential, as these terms have evolved from differing points of reference in regards to the polarity of a fundamental charge unit. Having obtained the built-in *electrostatic* potential in Figure 1.5, the electric field is next computed, the quantity of interest for determining the efficiency of electron multiplication. The electric field is simply the derivative of the potential

$$E = -\nabla \psi \, . \tag{1.5}$$

From Figure 1.5, a very small built-in electric field is observed at zero bias. Increasing the bias above the breakdown voltage of 59 V generates an electric field greater than the $\sim 3 \times 10^5$ V/cm needed for the junction to breakdown via impact ionization.

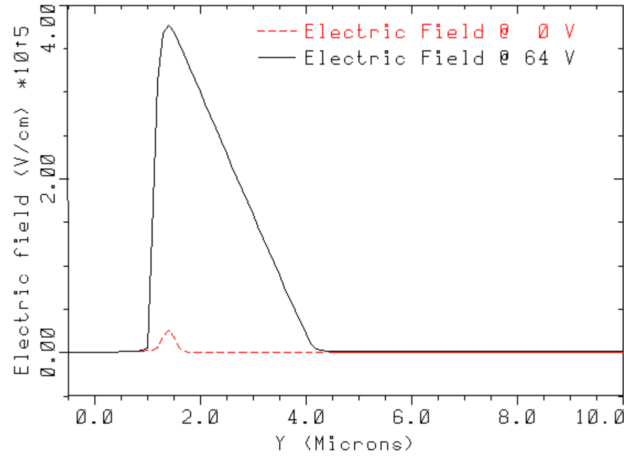


Figure 1.6. Electric field for diode at 0 V (red) and 64 V (black) applied bias.

Although the high electric field region spans the entire depletion width (3 µm at 64 V), the region of impact ionization is confined to within a fairly small window around the junction (Figure 1.7). This is again due to the strong dependence of the electron and hole impact ionization coefficients on the electric field.

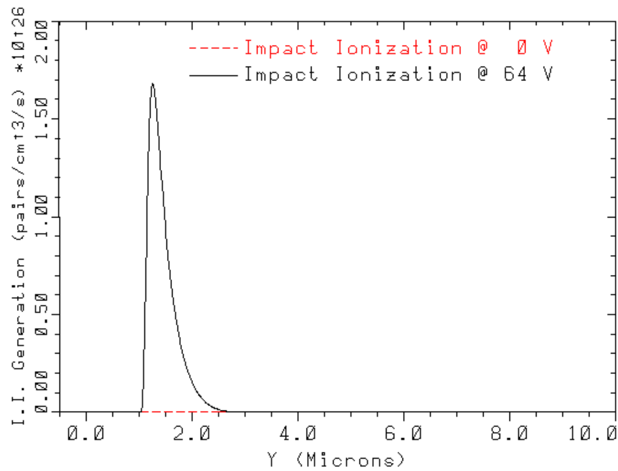


Figure 1.7. Local impact ionization rate in electron-hole pairs/cm^3/s at 64 V applied bias.

Because the junction is located some distance into the substrate (1.34 µm), the region of impact ionization is strictly confined to that same depth, between 1.0 µm and 2.5 µm below the surface. Therefore any photoelectrons (or "photoholes") created above or below this region must drift towards the junction in order to become detected through avalanche initiation. This should also serve to illustrate the relative sensitivity of impact ionization to changes in the applied bias.

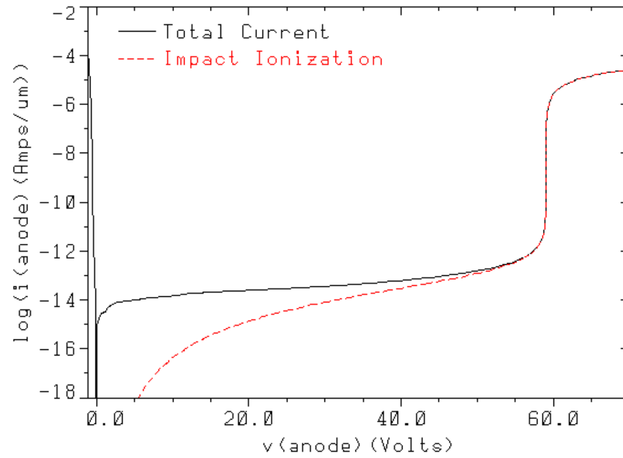


Figure 1.8. Current-voltage characteristic curve of one-sided junction diode, where the positive voltage on the cathode represents a reverse bias. The dominance of the impact ionization current (red) is apparent near the breakdown voltage.

Without impact ionization, the reverse bias current (Figure 1.8) would continue at low levels towards much higher voltages. In fact, this is how an avalanche photodiode operates in Geiger mode. In the momentary absence of free electrons to cause impact ionization, the I-V characteristic would appear to have a low current blocking feature above 59 V (i.e., no breakdown). However, in the presence of free carriers, as in this non-transient simulation, the current always rises swiftly at the breakdown voltage. The gradual onset of impact ionization is observed as a soft knee in the curve near breakdown. The eventual reduction of current above breakdown is a result of the diode resistance which limits the current flow. On a linear scale, the I-V curve in Figure 1.8 appears as Figure 1.9.

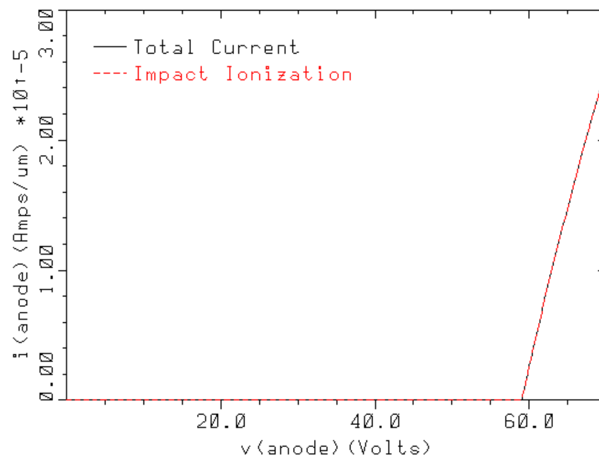


Figure 1.9. Current-voltage reverse bias characteristic from Figure 1.8, on a linear scale.

The inverse slope of the current above breakdown yields a diode *resistivity* of

$$R = \frac{\Delta V}{\Delta I} = \frac{11\ \text{V}}{2.6\times 10^{-5}\ \ \text{A}/\mu\text{m}} = 423\ \text{k}\Omega\cdot\mu\text{m} \tag{1.6}$$

for this two dimensional simulation of a “one dimensional” diode that is 1 μm wide and 15 μm deep. This is reasonable since the resistivity of p-type silicon doped to 1×10^{16} cm^{-3} is approximately 14 kΩ-μm, making the resistivity of the original 1 by 15 μm p-type substrate 210 kΩ-μm.

It might seem that this one-dimensional diode could perhaps function as an avalanche photodiode, which would be an accurate assessment if this fictional 1D diode could be realized. However three dimensional devices often perform differently than their simplified one dimensional counterparts. There are many additional processing and device requirements that must be addressed before a functional avalanche photodiode can be fabricated. The process of diffusing dopants into silicon involves a complex interplay between vacancies, interstitials, lattice defects and electric potentials, for instance. Therefore, simple Gaussian and complementary error function depth profiles can be inadequate to describe achievable profiles. A determination must next be made as to where the high field region is to be located and thus which diffusion profiles are most desirable.

1.7. Scintillation Absorption

In order to tailor the electric field for scintillation detection, we must determine the location of scintillation photon absorption in silicon. The depth at which photons are absorbed is a strong function of the photon energy (or wavelength), as indicated by the absorption coefficient $\alpha(\lambda)$ in Figure 1.10 for optical photons in crystalline silicon.

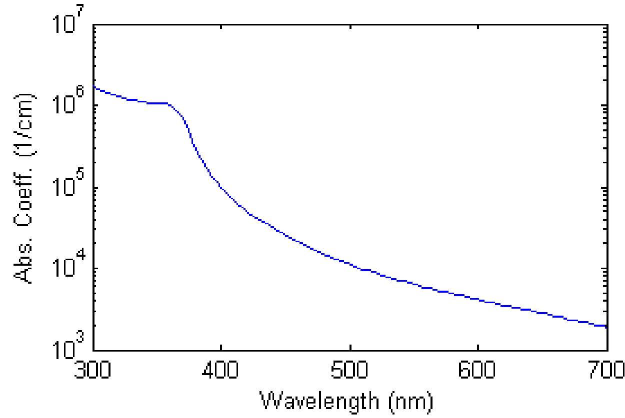


Figure 1.10. Optical absorption coefficient for crystalline silicon [Gre95].

Incidentally, the absorption coefficient is directly related to the extinction coefficient *k* of the complex index of refraction ($n + ik$) as

$$\alpha = \frac{4\pi k}{\lambda}. \tag{1.7}$$

The sharp rise in absorptivity below 400 nm is due to the electronic band structure of crystalline silicon. Photons with energies less than the 1.12 eV indirect bandgap (>1100 nm) require *phonon* assistance, but photons with energies greater than 3.4 eV (<360 nm) can transfer their energy *directly* within the same wavevector and do not require phonons for change in direction (i.e., momentum). This also explains the minimum ionization energy in silicon (3.62 eV) for higher energy particles like gamma rays and charged particles. Near-UV scintillation photons just above this ionization energy have a much higher probability of being stopped very near the surface.

The choice of optimal electric field location and width should then depend on the particular scintillation emission spectrum. Three such spectra are displayed in Figure 1.11 from the commonly used scintillators BGO, CsI:Na, and $LaCl_3$.

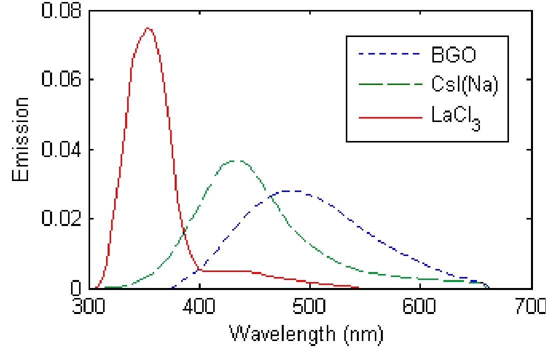


Figure 1.11. Scintillation emission spectra (area normalized), for three common scintillators: BGO, CsI:Na, and $LaCl_3$.

Given that the emissions from many scintillators peak in the 350-500 nm range, an ideal electric field would be expected to be confined to within the first few microns of the surface. Just how deep the junction should be located depends on the depth at which photons are actually absorbed. The probabilistic depth can be obtained by weighting the exponential absorption probabilities by the emission probabilities at a given wavelength, essentially combining data from Figure 1.10 and Figure 1.11. The probability of a monoenergetic photon being absorbed between the surface and a depth x is simply

$$P(x;\lambda) = 1 - e^{-a(\lambda)x} . \tag{1.8}$$

For the three scintillators noted above, the cumulative probability of photons being absorbed from the surface down to a certain depth is obtained by integrating the absorption probability over all λ and then integrating from 0 depth to some depth x. These scintillator-specific cumulative probabilities are illustrated in Figure 1.12.

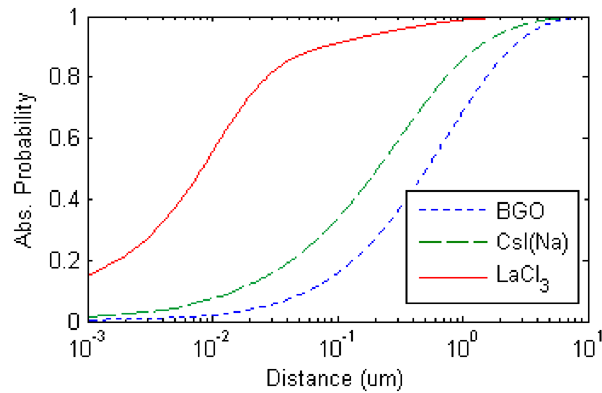


Figure 1.12. The cumulative probability of photons being absorbed by a certain distance in crystalline silicon, for three separate scintillators.

Note that 50% of BGO photons are absorbed within 500 nm, while 50% of $LaCl_3$ photons are absorbed within just the first 9 nm. Clearly, for these blue-UV scintillators, the electric field need be no deeper than several microns to be in the vicinity of absorption. The risk in creating deeper and wider junctions is the increased sensitivity to a larger volume of thermally generated electrons which increases the dark noise. However, processing challenges become much more complex when attempting to create ultra-shallow junctions. The increased surface- and interface-generated currents in shallow junction technologies are often detrimental, and their avoidance requires a significant investment in process research.

Another significant conclusion concerning photodetector design for scintillators can be drawn from Figure 1.12. It should be noted that even for scintillators like BGO (λ_{max} = 500 nm), nearly 100% of photons are absorbed within the first 5-10 µm. Consequently, the actual thickness of silicon required to realize total optical absorption need only be this thin. There of course are practical processing concerns limiting the minimum thickness of wafers during fabrication to ~300 µm. However, final devices may be thinned significantly if useful to the application. For instance multiple crystal arrays could be stacked with interfacing thinned SiPM's to reduced the radiation-sensitive dead region.

One possible method for creating shallow, more UV-sensitive junctions is to simply remove the "dead" layer of silicon immediately above the junction. Wet chemical, dry plasma, and gaseous XeF_2 etching may each have utility in this

regard, however the exact surface science of the etching process must be known to ensure a contamination-free and reasonably well passivated surface. Even without this dead layer removal, a careful selection of junction depth and depletion widths should be made, accompanied by knowledge of the incident photon spectrum.

1.8. Diode Edge Breakdown

The spatial confinement of diodes and other planar semiconductor devices is achieved through masked diffusion of donor and acceptor atoms. For silicon, a thermally grown oxide of several thousand angstroms is sufficient to block the diffusion of boron and phosphorus at typical temperatures of ~1000 °C. For instance, fabrication of a n^+/p diode like the 1D structure previously discussed might begin with the opening of a window in an oxide layer through which n-type atoms (typically phosphorus) can diffuse into the underlying silicon. The higher the doping concentration in either the diffusion or substrate, the greater the charge density gradient at the n^+/p junction, which leads to a larger built-in electric field. A larger built-in electric field requires a lower reverse bias to achieve the critical electric field necessary for impact ionization (~3×10^5 V/cm).

In physically realized diodes, the concentration of diffused dopants falls off gradually at the diode edges and corners, due to the isotropic diffusion process, as illustrated in Figure 1.13. The oxide diffusion mask is used to initially restrict n+ impurities laterally to within 0 and 1.5 μm. The dopants then diffuse down into the bulk of the wafer, but also diffuse laterally, creating a curved junction.

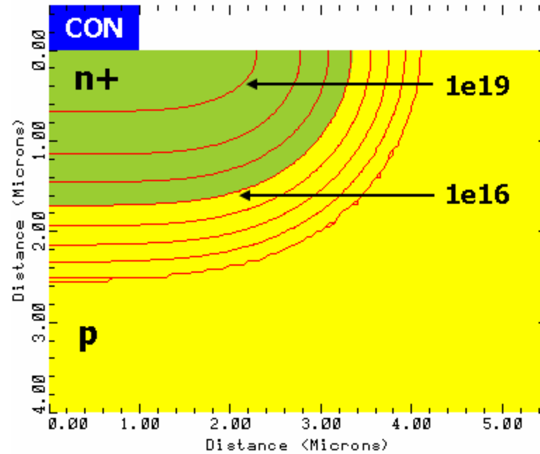


Figure 1.13. Doping profile for n+/p diode with mask edge at 1.5 µm. Contour lines illustrate decades of phosphorus (n+) concentration. Diode junction depth is indicated by green-yellow fill. “CON” is the cathode contact, while the anode contact at the bottom of the substrate.

This junction curvature at edges and corners represents a locally higher gradient in charge density, and thus produces a larger electric field. Therefore, a single diffused junction diode will break down preferentially at its corners, then at its edges, followed by the remaining planar area [Sze66]. The diode illustrated in Figure 1.13 borrows identical doping conditions from the 1D diode considered in Section 1.6. That is, the 1D diode doping of Figure 1.2 has exactly the doping depicted in Figure 1.13 along the ($x = 0$, y) cross section.

Examining next the resulting electric field (Figure 1.14a), we note that at 40 V applied bias, the maximum electric field, ~3×10^5 V/cm, is of the order necessary for probable impact ionization. However, the high field region is not very wide, therefore a greater electric field may be required to allow electrons to accelerate over an appreciable distance in order to trigger a sustained avalanche (i.e., breakdown). When compared to the 1D case in Figure 1.14b, the high field region of the 2D diode is clearly larger at 40 V. The current, as indicated by the 10% flow lines, is observed to travel in very straight lines as it passes through the depletion width. Only after passing the end of the lightly doped region does the current change direction, attracted by the anode potential at the bottom of the substrate. The flow lines are also observed to crowd near the high field region as the diode reaches breakdown.

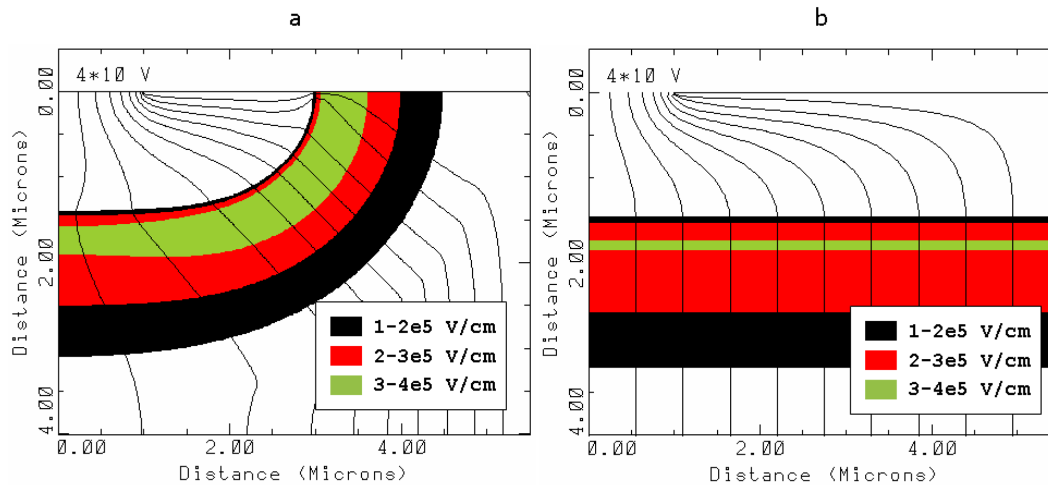


Figure 1.14. Electric field at 40 V applied reverse bias referenced to bottom of substrate for 2D diode (*a*) and 1D diode (*b*). Current flow lines (10% each division) are plotted from cathode to anode.

The high electric field which extends to the surface is also worthy of notice. The simulation of charge transport at the oxide-silicon interface is highly dependent on surface passivation and many other morphological parameters. Therefore, high electric fields at the surface will most likely lead to unintended noise consequences unless proper characterization data is drawn from real devices.

Prediction of the breakdown voltage is achieved by numerically simulating the impact ionization process and observing the reverse bias current-voltage characteristic curve, as shown in Figure 1.15. Note that at 40 V, the contribution from impact ionization is already the dominant component of the total current. A breakdown voltage of 48 V is apparent from the multi-decade rise in current indicating runaway impact ionization.

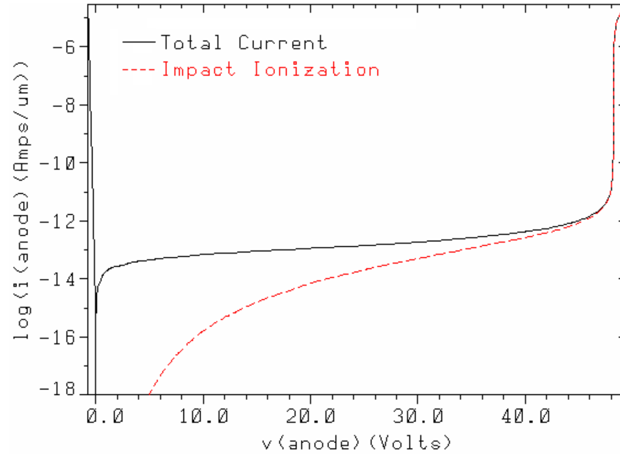


Figure 1.15. Current-voltage characteristic curve for 2D diode with doping profile from Figure 1.13.

When compared with the 1D diode breakdown voltage (Figure 1.8) of 59 V, there is a marked difference even though the bulk diffusion profiles are equivalent. The 2D masked diffusion process creates a diode which breaks down at its edge a full 11 V before breaking down across its length.

In Section 2.2.A, a simple single-diffusion diode structure is presented which when biased above breakdown emits photons due to recombination of hot-carriers. These hot-carrier emissions allow localization of regions with presumably the greatest electric field.

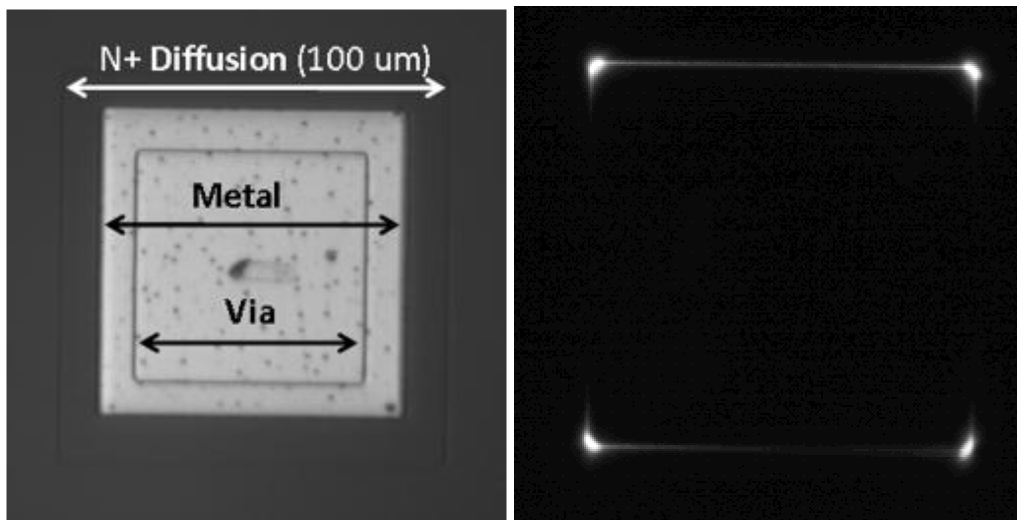


Figure 1.16. Electroluminescence image (right) of a single-diffusion square n+/p diode operated above breakdown, as imaged by a CMOS camera on a probe station microscope. With increasing bias, this diode breaks down first at its corners, then at its edges.

One can clearly see from Figure 1.16 that the brightest regions are the corners of the diode, while the edges appear more faint. If we were to continue increasing the bias on this diode, we might eventually see the central area begin to glow, however the majority of the current flow would still be confined to the corners and edges. At sufficiently large reverse biases, heating from the intense avalanche current would destroy the contacts or short the rectifying junction. The addition of a quench resistor limits this current to a standard amount in working devices.

If operated as a single photoelectron detector in Geiger mode (above breakdown), there would be so many more thermally generated electrons causing impact ionization at the edges or corners, that any photoelectrons in the planar bulk would go undetected. However, if one were able to focus all incoming photons onto the corner or edge of such a single-diffused junction, the diode could be operated as a Geiger-mode avalanche photodiode, albeit one of very poor design. If the primary requirement of a single photon sensitive avalanche photodiode is a high electric field, then the second requirement is for that electric field to be both confined and uniform.

1.9. Edge Breakdown Prevention

In order to create a diode with spatially uniform avalanche initiation probability, the electric field must be reduced at the diode's edge and made constant over the majority of the surface. While this region serves as the "active" area of the diode, the choice of edge breakdown prevention structure can allow for (photo)electrons or holes to initiate avalanches by drifting in from "non-active" regions. The enhanced electric field at the edge is derived from a relatively large potential difference over some relatively small distance [Sze66]. Several structures are next introduced which are capable of reducing planar diode edge breakdown by altering this local potential difference.

1.9.A. Field Plates

Field plates are metal layers located above the diode which exhibit some electrostatic force on the charge carriers below [Con72] and have long been used

to tailor electric fields at the edges of large-featured semiconductor power [Bal08] and detector devices [Lut07]. However, they are usually separately biased or left floating and often consume a not insignificant amount of real estate. Because the diodes in this work are necessarily small (tens of μm) and closely packed (pitch ≈ width), little room is left for high voltage field plate structures. In order to control very shallow fields, much higher voltages would also be needed on field plate structures, and breakdown of passivation oxides and interlayer dielectrics would become problematic. For these reasons, we have chosen not to examine field plates and other metal overlying structures for the termination of planar avalanche diode edge breakdown.

1.9.B. Guard Rings

Since the large electric field at the periphery of the diode is due to a high concentration of some single diffusion, one could imagine some structure which would gradually lessen this concentration, thus reducing the peripheral electric field. In this context, we define guard rings to mean additional diffusion into silicon, rather than the metal "guard rings" which are an extension of the overlying metal field plate theory just mentioned.

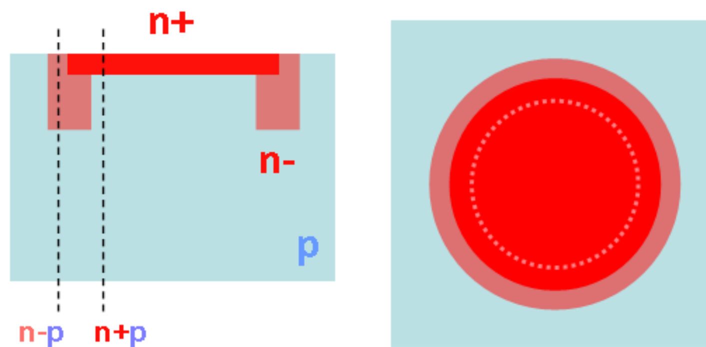


Figure 1.17. Cross section and top view of guard ring diode concept.

For instance, given the p-type substrate in Figure 1.17, a single-diffusion n+/p diode would first break down at its periphery, where the lateral diffusion edge curves up towards the surface. By introducing another diffusion of lesser doping, the n- guard ring, the peripheral diode is altered to one of a lowered charge density gradient. Effectively, the planar diode edge is transformed from an n+/p

diode into an n-/p diode, and the high electric field is contained within the bulk at the expense of silicon real estate and a lowered active (high electric field) area. Depending on the difference in doping between the primary diffusion and the additional guard ring diffusion, the diode periphery may break down at voltages very close to that of the "active" primary diode, thus preventing the use of large excess biases.

1.9.C. Junction Termination Extension

Instead of adding lower doped regions to the outside of the diode, a higher doping concentration can be introduced into the bulk of the diode with the same type as the substrate. This again achieves the effect of creating two types of diodes. The original n+/p diode at the periphery, and the new n+/p+ or n+/p-well diode at the center.

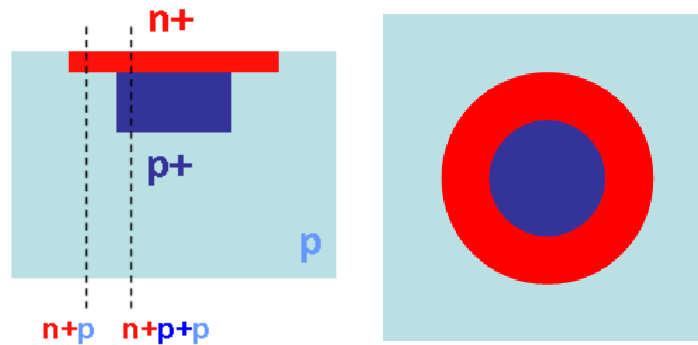


Figure 1.18. Cross section and top view of junction termination extension concept.

The effects of junction curvature are thus nullified since the edge of the p+ well decreases laterally in concentration, and the n+ concentration decreases vertically at the junction periphery. Thus, no peak electric field region is created at this location, and breakdown will be confined to within the bulk of the diode. It should be mentioned that the term *junction termination extension* also applies in the literature to successively smaller diffusions surrounding the original junction, which also has the effect of reducing the electric field. For the sake of silicon real estate consumption, this alternative interpretation is not considered as a viable method.

1.9.D. Bevel Etching

The concept of bevel etch or moat termination (see Section 2.4.A) is slightly different from the methods mentioned previously. Instead of controlling the doping concentration, the three-dimensional geometry of the diode itself is altered. A unique property of the face centered cubic diamond lattice structure is that certain chemicals can be used to selectively etch in certain lattice directions. Etchants like KOH, EDP, and TMAH each have a (100):(111) etch selectivity ratio of 10 to 10^4. Depending on the orientation in which the silicon boule is sawn, the plane which is coplanar to the surface can be selected, common orientations being (100) and (111). When subjected to the previously mentioned wet etchants, near 54° and 90° sidewalls can be achieved for (100) and (111) wafers, respectively.

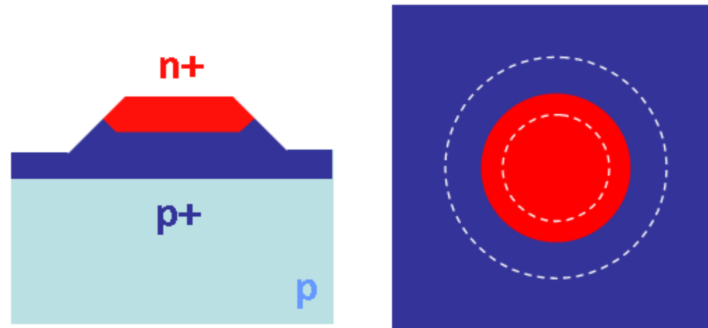


Figure 1.19. Cross section and top view of bevel diode concept.

The sloped sidewall reduces the electric field at the surface by increasing the effective depletion width at the surface due to the relative imbalance of ionized dopants [Dav64]. Additional diffusions after bevelling can add further curvature to the depletion region through the out-diffusion of dopants at the surface, thus reducing the electric field even further.

1.10. Summary

This chapter sets the stage for the results of specific fabrication efforts in the following chapter. First, motivation was provided for the enhancement of scintillation photodetectors based on the improved energy, position, and timing resolutions expected. The process of impact ionization was determined to be

necessary for the detection of single free charge carriers. The creation of regions of high impact ionization rates were examined through a fictional one-dimensional diode in which the doping, net charge, potential, and electric field were illustrated. The consequence of variable wavelength scintillation emission leads to varying depths of photon absorption. However, the vast majority of scintillators emit photons which are preferentially absorbed at depths much less than the thickness of a standard 0.5 mm silicon wafer. The total probability of absorption for different scintillators should inform the selection of appropriate doping profiles and electric field locations for *matched* photodetectors.

A more realistic two-dimensional simulation of junction electrostatics illustrated the phenomenon of premature peripheral breakdown. Electroluminescence images from real diodes confirmed this well-known fact. In an attempt to mitigate this effect and produce a uniform electric field, and thus uniform probability of detection, several edge breakdown prevention structures were presented. Field plates commonly found in power devices and large semiconductor detectors were deemed unreasonable given the size constraints of small diodes in a SiPM. Diffused guard rings provide a gradual decline in doping at the diode periphery, and thus lessen the electric field. A junction termination extension structure was determined to reduce the edge breakdown by increasing the electric field at the center of the diode. Finally, bevel etching was theorized to reduce edge breakdown through the net charge imbalance due to the altered bevel geometry. In Chapter 2, we examine several initial fabrication efforts towards the realization of these structures.

Chapter 2

Avalanche Diode Fabrication

The majority of existing single-photon avalanche photodiode and silicon photomultiplier devices have been fabricated using well characterized CMOS processes or custom foundries dedicated to clean electronics processing. In contrast, this work attempts to make use of existing technologies within the Lurie Nanofabrication Facility at the University of Michigan, a laboratory dedicated primarily to micro electrical-mechanical systems (MEMS) research. The requirements for MEMS processing are substantially different from those for clean detector or CMOS processing, and while successful CMOS processes have been implemented in this lab, they are not necessarily guaranteed or well-maintained. With these infrastructure concerns in mind, the first step in clean processing of detectors in a shared facility is to effectively firewall one's process and perform tool and process tests sufficient to characterize the desired level of cleanliness. Done correctly, this is an extremely costly proposition, and therefore, one strives for balance between repeatable process outcomes and excessive testing.

The first section in this chapter is devoted to the reduction of leakage current through gettering processes which aim to remove contamination introduced as a result of shared tools. Subsequently, various edge breakdown prevention structures are fabricated and analyzed for their capacity to tailor the electric field and provide for a suitable Geiger-mode structure. Finally, a novel method for the fabrication of transparent quench resistors is described.

2.1. Doping Processes

While ion implantation remains the primary method for the controlled introduction of dopants, there are several potential drawbacks to this technology. The interaction of highly energetic ions with a well-ordered semiconductor lattice by their very nature leads to crystalline defects. The implanted ion in addition will necessarily find itself some distance from a silicon lattice node. Therefore, it will not contribute as an ideal donor or acceptor, as-implanted. The degree to which donor or acceptor atoms are incorporated into an ordered semiconductor lattice is referred to as the level of impurity *activation*. Both lattice damage reduction and the electrical activation of dopant atoms can be achieved to varying degrees by high temperature annealing. Any remaining defects in lattice periodicity can lead to noise in single photoelectron sensitive devices. For this reason, we have attempted to circumvent these potential issues by generating diffused junctions with legacy furnace diffusion technology. This removes the damaging implantation aspect, but the displacement of silicon atoms remains necessary to make room for the impurity atoms. One additional consideration is the solid solubility of dopant atoms in silicon. As the diffusion temperature increases, silicon can accept a larger concentration of non-precipitated dopants (Figure 2.1).

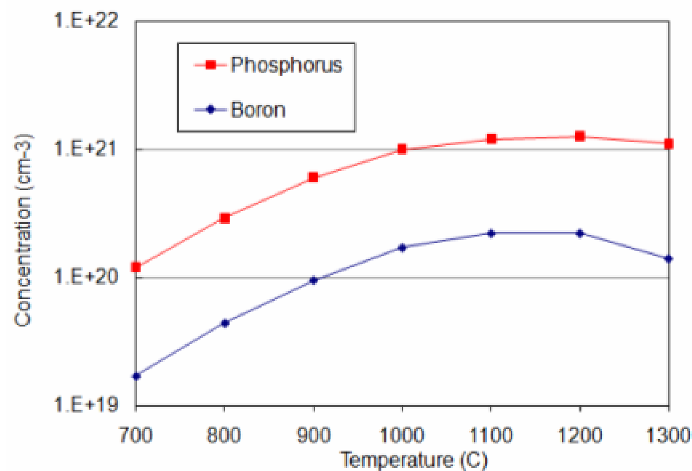


Figure 2.1. Solid solubilities for phosphorus and boron in silicon (from TSUPREM4 [TSU07]).

There is also a separate temperature-dependent limit to the electrically active concentration of dopant atoms in silicon, even with furnace diffusion. Reduced

doping concentrations assist in the total activation of dopant atoms and in the elimination of interstitials or precipitates which would lead to excess generation of charge carriers and generation of dark noise from SiPM pixels.

While ion implantation offers a very atomically selective method for introducing dopants, there are a number of published opinions which either hypothesize or directly observe the negative impact of implantation-induced defects [Key01]. Rochas points out [Roc03] that high-energy implantation damage, particularly at the wafer surface, may only be partially repaired by thermal annealing. He further indicates this phenomenon as a potential problem for UV photons which are preferentially absorbed in this potentially damaged region. This would cause both a higher thermal generation rate, as well as a reduced photoelectron lifetime. Thus, residual implantation damage may remove the possibility for a photoelectron to drift or diffuse into the depletion region. Kindt [Kin99] attributed poor dark noise characteristics of fabricated avalanche diodes to the possible incomplete dissolution of implantation damage. He noted that damage located nearer to the junction creates traps in an area with a larger electric field. These traps then contribute to trap-assisted tunneling whereby a larger population of electrons are free to initiate breakdown. Spieler cautions [Spi05] of "low-temperature" processes which do not fully activate ion implanted dopants. He mentions the implantation-generated interstitials, which can exist or migrate beyond doped regions and create defect zones. It is noted that doping by diffusion ensures a high level of activation. The doping gradients at the lateral extents of ion implanted regions are also sharper, and thus require more care in preventing unwanted peripheral breakdown.

The open question perhaps is to what degree lattice defects and dopant inactivation can be tolerated by single photoelectron detectors. The answer will at the very least depend on the exact spatial distribution of the electric field in each realized diode. The counterargument against traditional diffusion doping is that additional impurities in the dopant sources may actually contribute more trap states through contamination atoms than would ion implantation through lattice defects and inactivated dopants. However, given the above concerns regarding ion

implantation, the methods in this work represent a departure from industry standard processing in favor of furnace diffusion. This presents the work at hand with both the opportunity and the requirement of removing any additional atomic contaminants which would degrade the band gap quality and lead to higher dark counts, thus limiting device size.

2.2. Gettering

Gettering is defined as the process by which unwanted impurities are removed from important areas of a semiconductor substrate. Harmful impurities to charge carrier lifetime and junction quality are those atoms which would disrupt the periodicity of a perfect semiconductor lattice. Interstitials, vacancies, and lattice damage or disorder can each lead to reduced carrier lifetimes and enhanced generation currents.

Many gettering processes rely on the high mobility of harmful ionic contaminants. Just how high the mobility is depends on the temperature and species of interest. Many metals and salts are present in common laboratory conditions, and many of these atoms are capable of traversing the depth of a standard wafer if not further. Various forms of physical trapping sites can act to sequester impurities during the many high temperature steps required for planar junction processing [Pol88]. Early methods employed a wafer backside which was roughened or damaged by ion bombardment. More modern methods seek to create gettering *sinks* in unimportant areas of the wafer, thereby leaving a more pure silicon lattice in important device areas. One reason for the relative lack of information on gettering arises from implicit or unplanned gettering which happens as a natural result of many high temperature process steps. This makes inter-lab or inter-institution comparisons difficult unless standardized and costly testing is performed at each process step. Information from industrial gettering developments is also unlikely to be shared since successful gettering techniques can dramatically affect the bottom line of profit-driven organizations. Because devices in this work are processed in our own fabrication facility, we are free (or

required) to test various gettering methods to improve the band gap quality at the avalanching silicon junction.

2.2.A. Gettering Test Structure

Following previous developments [Hol89 Log95], we conclude that a polysilicon backside, doped with phosphorus, will act as a successful getter for diodes in this work. Other groups have achieved success with phosphorus gettering sites on the front side as well [Sci03] and claim that gettering sites must be located very close to the active junction. One possible explanation for the need for this necessary proximity may be the short anneal times accompanying ion implanted doping. We therefore have laid out a compact experimental design (Figure 2.2) to test the effects of front and backside phosphorus doping as well as the presence of a backside polysilicon layer. These eight wafers will be referred to throughout the remainder of this thesis as wafers #1-8. Eight additional ~1 Ω-cm prime CZ p-type wafers were identically processed and will be referred to as wafers #9-16.

n-	p	poly	back dope	n-well	n+	p-well	p+
1	9		n-well, n+				
2	10		n-well, n+			deep	
3	11		n-well, n+		deep		
4	12		n-well, n+	deep			
5	13		n+	(none)	deep	deep	
6	14	(none)	n+	(none)			
7	15	(none)	n+				
8	16	(none)	n+	deep			

Figure 2.2. Experimental design for 8 high resistivity wafers, indicating presence and doping depth of backside polysilicon and four diffusions.

A very simple structure for testing gettering efficacy in a silicon planar process is the one-sided junction diode illustrated in Figure 2.3. The reverse bias diode leakage current of these p+/n- diodes should be a direct indicator of the junction quality and therefore an indicator of gettering effectiveness. For instance, we might naively expect wafer #4 to perform the best since it contains a highly doped backside as well as phosphorus doped features on the front side.

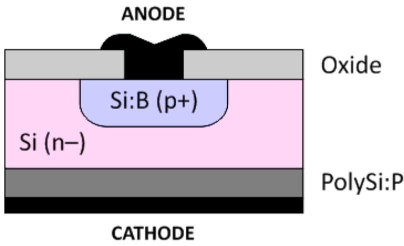


Figure 2.3. Structure of simple gettered one-sided diffusion diode.

The wafer thickness is 550 μm and the p+ (or p-well) region is 100 μm wide. A metal pad overlaps a via (i.e., window) in the 150 nm passivation oxide (Figure 2.4).

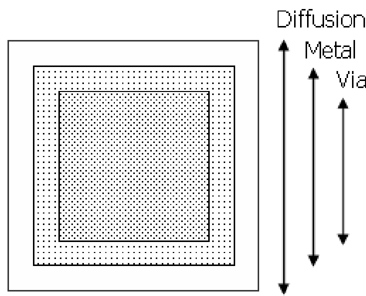


Figure 2.4. Mask patterns for 100 μm wide diffusion, 80 μm wide metal contact, and 60 μm wide via in the oxide passivation layer.

Both highly doped p+ and deeper low-doped p-well diodes are created over the entire surface of the wafer to ensure that die-to-die process variations are accounted for.

2.2.B. Gettering Process

The process for testing gettering functionality begins with the selection of very pure silicon wafers. In this test, the wafers supplied by Silicon Quest International, Inc. were four inch, n-type, (100), double side polished, high resistivity (>1e4 Ω-cm), float zone silicon wafers. For reference, the cost of each wafer was approximately 90 U.S. dollars. Upon receipt of wafers, an initial pre-furnace clean (PFC) was performed, followed by a wet trichloroethane (TCA) oxidation. The chlorine atoms in the TCA are known to attract metal ionic contaminants, just as the chlorine in $HCl:H_2O_2:H_2O$ is used to precipitate metal ions during the second step of the PFC.

For the phosphorus-doped polysilicon backside wafer, photoresist is applied to the front side of the wafer, and the thermal oxide is selectively etched off the back side. After photoresist stripping and another PFC, a layer of low-stress polysilicon is deposited on both sides of the wafer via low temperature, low pressure chemical vapor deposition (LPCVD). By controlling the ambient pressure and temperature, the dissociated silane gas atoms (SiH_4) are deposited onto the wafer with a small grain structure. The minimization of polysilicon grains reduces residual stress later when it will be found on only one side of the wafer. Small polysilicon grains also act as excellent gettering sites for highly mobile impurities diffusing during the remaining high temperature steps of the process. The polysilicon on the front side of the wafer is removed down to the original thermal oxide via reactive ion etching in a STS Pegasus tool. Because of the multi-user nature of the facility, pre- and post-etch plasma chamber cleaning procedures were developed with the tool engineer. A precautionary PFC is also performed *after* etching. The resulting wafer then comprises thermal oxide on the top and polysilicon on the bottom.

Deep phosphorus diffusion of the polysilicon is achieved by performing another PFC and doping the wafer in a $POCl_3$ furnace for 10 minutes at 900 °C. During this process, a thin layer of phosphorus doped oxide is built up on the wafer surface and acts as a constant source of dopants at the solid solubility of silicon ($\sim 5\times10^{20}$ cm^{-3}). After removing this layer, a 3 hour diffusion at 1100 °C drives the phosphorous into the polysilicon and partially into the overlying silicon. During this high temperature process, many mobile ionic contaminants are supposed to be trapped by either the polysilicon grains or by the excess of phosphorus atoms. Beyond this point, these *gettered* wafers are processed in a similar manner to the non-gettered control wafers.

All oxides are stripped, and a 5000 Å LPCVD oxide is deposited at 600 °C as a diffusion mask for the one-sided diode diffusion. A high concentration of boron is then diffused through the oxide window. Another masking oxide is deposited and selectively etched from the backside so that a high phosphorus concentration can be introduced at the backside contact, similar to the initial $POCl_3$ diffusion.

This step is essential for forming an ohmic contact to high resistivity silicon. A final passivation oxide (LPCVD) is deposited through which contact windows are etched. Metal contacts, 5000 Å thick, consist of room-temperature sputtered Ti/TiN/Ti/Al:Si(1%), patterned by a photoresist liftoff method. A backside contact of Ti/Al:Si(1%) is also sputtered, and all wafers are subsequently annealed at 400 °C for 30 min in a forming gas (10% H_2 in N_2) ambient.

2.2.C. Gettering Results

The diodes are left on the wafer (i.e., not diced), since dicing has the potential for inducing stress and causing additional leakage currents. A dark probe station is used to test the reverse bias leakage current of diodes from each wafer (Figure 2.5). Contact is made to the entire backside by holding the wafer to a conductive chuck with a slight vacuum and contacting the chuck with a tungsten probe tip. Front side contact to the diode metal pad is established with another 45° tungsten probe tip. Triaxial cables connect the probe arms to a Keithley 4200 Semiconductor Characterization System for current-voltage testing.

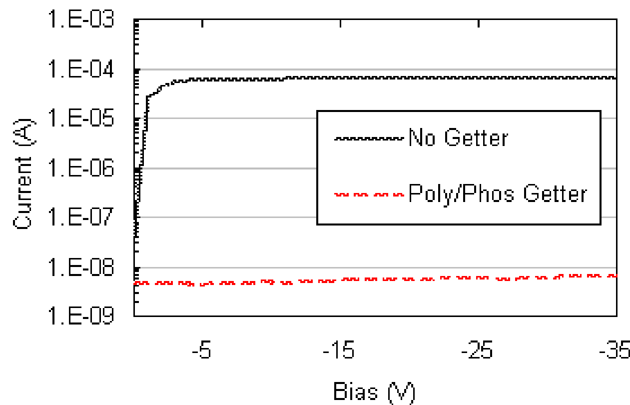


Figure 2.5. Reverse bias current-voltage characteristic curve for simple through wafer p+/n diodes in high resistivity silicon.

The leakage current from the gettered wafer diode (#4) is approximately four orders of magnitude less than the identical structure from the non-gettered wafer (#6). A reverse bias current of 5 nA is achieved between the 78.5 cm^2 backside contact and the 1×10^{-4} cm^2 front side contact. The contact metallization and passivation may play a large role in the absolute value of this leakage current density, but the key parameter of interest is the presence of a doped polysilicon

backside. For this reason, all wafers were given the exact same passivation and metallization structure and were processed simultaneously to rule out effects from tool and process variations over time.

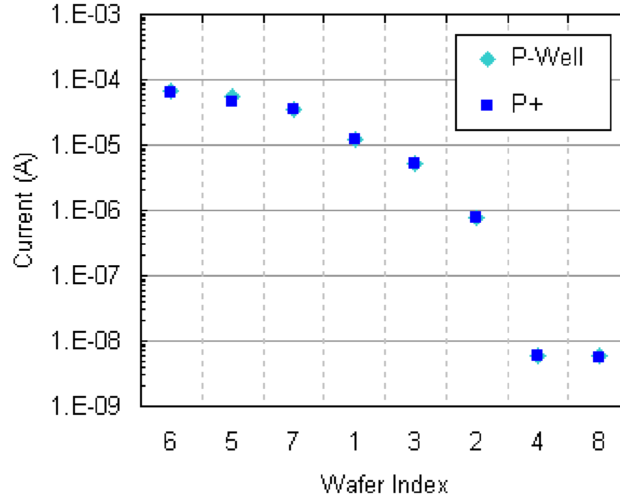


Figure 2.6. Reverse bias leakage current at 20 V reverse bias for the 8 high-resistivity n-type wafers.

When observing the leakage currents from all diodes (Figure 2.6), the wafers with deep n-wells (#4 and #8) are seen to have the lowest leakage current which may indicate that the effect is more one of doping and contacts than of bulk leakage. However, the four wafers with phosphorus-doped polysilicon backsides (#1-4) each exhibit lower leakage currents. We can at the very least conclude that a heavily-doped polysilicon backside is not detrimental to diode leakage current. It should noted that the leakage current from the p-type wafer diodes (#9-16) were in a much larger range of 10 μA to 1 mA.

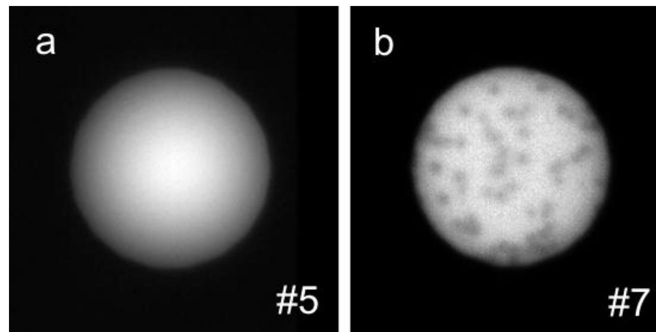


Figure 2.7. Electroluminescent images for diodes from wafers #5 (a) and #7 (b).

Upon examination of the reverse bias electroluminescence patterns from diodes of several wafers, some additional conclusions can be drawn. Figure 2.7 indicates the location of hot-carrier emissions for diodes from wafers #5 and #7. As hot (i.e., energetic or accelerated) carriers recombine, the extra energy is emitted as an optical photon which indicates the location of recombination. Local dark spots indicate regions where no energetic recombination takes place, and therefore one would expect a lower electric field. Wafer #5 employs an *undoped* polysilicon backside, while wafer #7 has no special backside treatment. One might conclude that the presence of the polysilicon acts successfully as an impurity getter. One additional difference is that wafer #5 has no front-side n-well features (only n+), while wafer #7 does have front side n-well features. Thus one might additionally conclude *for this process* that even with front side n-well and n+ features, a polysilicon backside (even undoped), is beneficial to electric field uniformity.

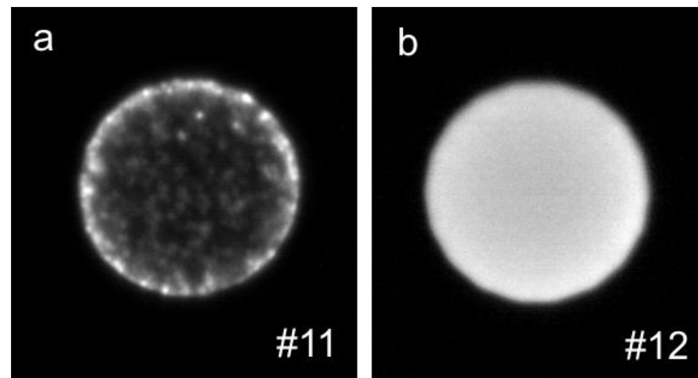


Figure 2.8. Electroluminescent images for diodes from wafers #11 (a) and #12 (b).

If instead of starting with high-resistivity silicon, we use “prime” grade, Czochralski (CZ) grade silicon of resistivity 1-10 Ω-cm, where there are approximately 10^{15} cm^{-3} more dopant atoms and many more potential impurities. Eight additional wafers were processed (#9-16) within the same gettering experimental design framework. From Figure 2.8 it is evident that the electric field is much more uniform in wafer #12 than in wafer #11. The improvement can be attributed from either a deeper n-well or a more shallow n+. The shallower n+ comes from a 50 °C reduction in doping temperature (to 850 °C), while the deeper n-well is the result of a higher temperature drive-in. One possible conclusion is

that the lower temperature n+ creates fewer phosphorus precipitates, and thus a more uniform electric field. Another plausible conclusion is that the increased n-well drive-in temperature effectively getters impurities from the active area of the diode.

Even with the potential for ionic contamination from the deep reactive ion etcher which strips the polysilicon off one side, a fundamental improvement is observed with the addition of a phosphorus-doped polysilicon backside. It is expected that specific efforts in contact and passivation development would also reduce the surface leakage measured within this gettering trial.

2.3. Junction Termination Extension Diodes

The junction termination extension (JTE) diode structure mentioned in Chapter 1 has been realized in several forms which illustrate the response to various diode parameters such as size, shape, extension, and contact location. Before analyzing these fabricated structures, we first take a closer look at the theory behind this edge breakdown prevention scheme.

A representative doping for an n+/p-well JTE diode is illustrated in Figure 2.9, where the junction is indicated by the yellow-green transition. Since deeper well diffusions diffuse isotropically from the original predeposition location, the degree of extension stated will be in reference to the original mask location. The structure illustrated represents a diode with a junction depth of 0.3 μm and an extension of 4 μm (x-axis: [8, 12] μm). The exact extension for any given diode will depend on the exact depth and doping of the well, therefore extension values should only be considered in a relative sense among diodes from identical processes.

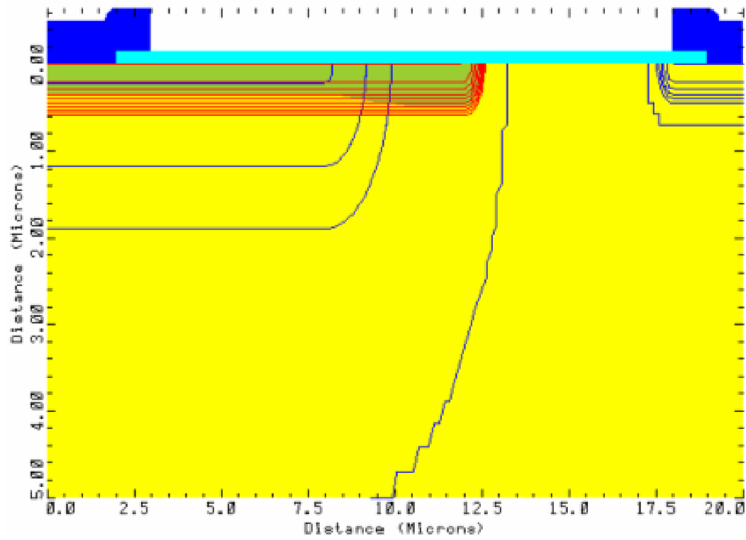


Figure 2.9. Junction termination extension structure (5 μm extension) with shallow n+ diffusion (red lines) extending past a deeper p-well (blue lines) for a junction indicated by the yellow-green fill.

Ohmic contact is made to the diode via the metal pad at the left, while ohmic contact to the p-type substrate is made at the right with the assistance of a shallow p+ diffusion. This represents a lateral contact scheme in which the current flow is confined to within several microns of the surface. A vertical contact scheme can also be implemented in which backside contact is made to the substrate and current flows through the entire wafer.

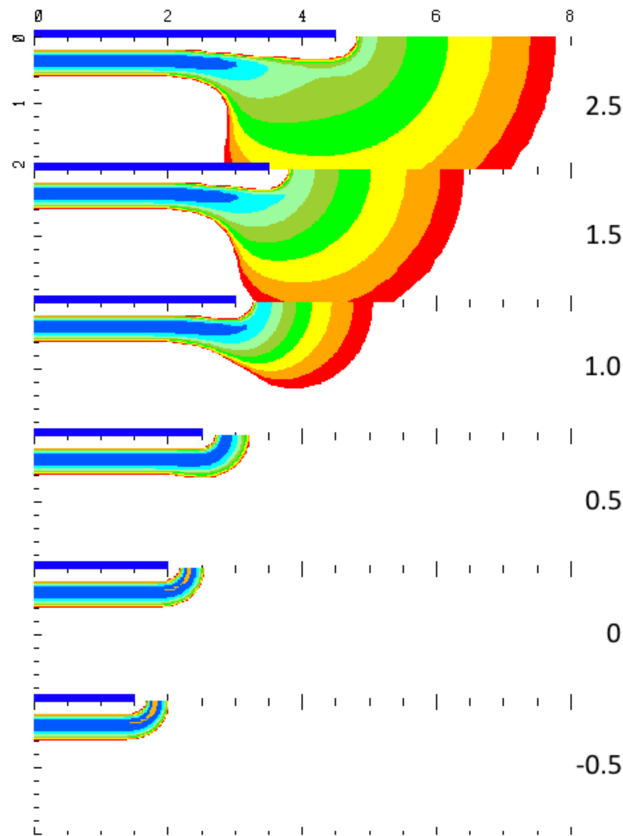


Figure 2.10. Electric field at 20 V for diodes with varying degrees of junction termination extension (at right, in µm). Blue bars indicate the shallow diffusion mask region.

Having established the doping structure in Figure 2.9, we turn our attention to the resultant electric field at a sufficiently high bias (Figure 2.10). At a junction extension of -0.5 µm, we see that the electric field is peaking at the periphery. Essentially, the shallow n+ diffusion is entirely contained within the deeper p-well. To the shallow diffusion, the effective substrate simply appears as p-type, and edge breakdown results are similar to those from the single-diffusion of Section 1.8. As the shallow junctions extend laterally further beyond the well, the peripheral electric field begins to subside. However, there still remains a significant electric field curving up towards the surface. In addition, a not unsubstantial electric field is driven ~isotropically away from the edge for each increment of junction extension. This places a limit on the pitch with which neighboring diodes can be placed and still remain electrically isolated. The added electric field also creates a drift component for photons absorbed outside of the

central active region. Some balance must be struck in order to determine the optimal degree of extension.

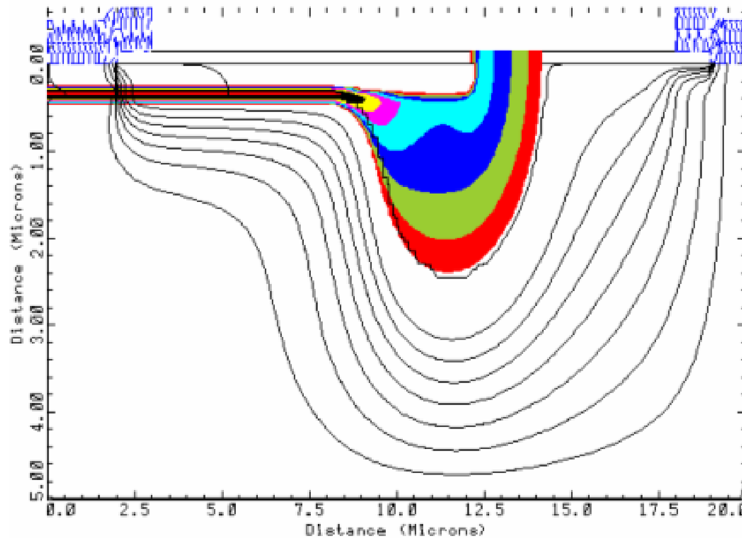


Figure 2.11. Electric field and current flow lines at 1 V excess bias for the structure in Figure 2.9.

Returning to the structure in Figure 2.9, we observe the electric field and the flow of current at 1 V excess bias in Figure 2.11. Notice that the path of least resistance from anode to cathode is through the central active region and around the electric field bulge. If electrons were not attracted across the high field region, then they would travel between contacts directly along the surface with a decrease in rectifying behavior.

2.3.A. Fabricated JTE Diode Results

Fabrication of JTE diodes follows the various processes and device wafers defined in Section 2.2. Recall that the run consists of eight high-resistivity n-type wafers (#1-8) and eight low-resistivity p-type wafers (#9-16). Diodes were fabricated with extensions from -2 to 8 μm in 1 μm increments, with widths of 10, 20, 50, and 100 in three different shapes: circular, square, and hexagonal. In addition, two different styles of contact were provided. The first employs a simple 5 μm via contact to the center of the diode and a backside contact for a vertically biased structure (Figure 2.12 at right). The second contact scheme is lateral, wherein the substrate contact is made in an almost closed ring around the diode and diode contact is established with another ring (Figure 2.12 at left) touching

the diffusions. This second scheme can also be used in a vertical fashion, so that the difference between the two vertical schemes is that one employs a point contact, while the other, an annulus.

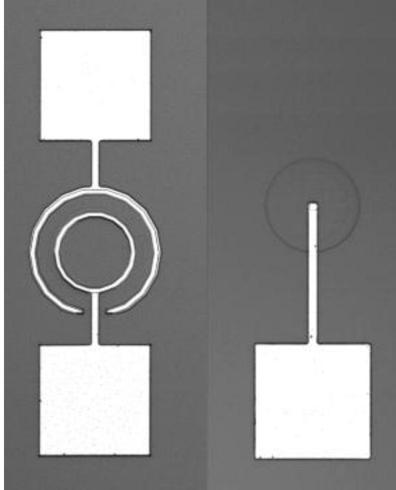


Figure 2.12. Two micrographs illustrate 50 µm circular JTE diodes, laterally contacted (left) as in Figure 2.11 and vertically contacted (right) with a single 5 µm via to the center of the diode. The shallow and well diffusions are barely perceptible.

The shallow and deep diffusions are barely perceptible, as the only visual evidence from a furnace diffusion is the shadow from the trough of etched silicon dioxide which is selectively grown during the ~900 °C, ~10 min diffusion.

First examining the *vertically* contacted diodes, we observe the effect of the extension on a 50 µm circular diode from wafer #7 (Figure 2.13).

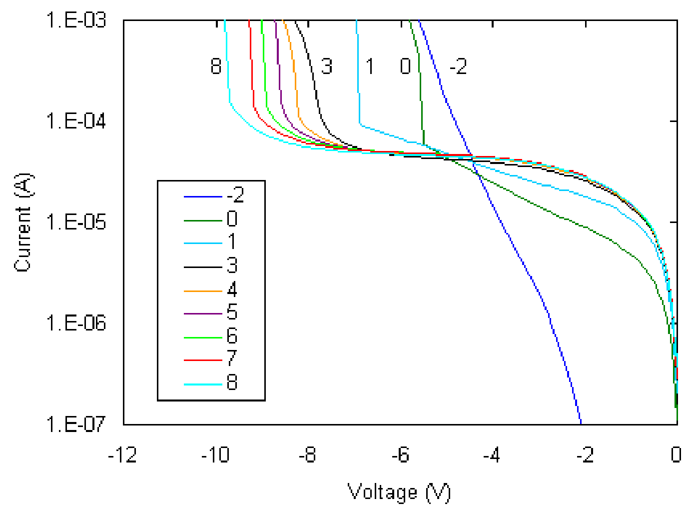


Figure 2.13. Reverse bias current-voltage characteristic for a 50 µm circular p+/n-well JTE structure on wafer #7. Extension values (in µm) are listed in the legend.

Note that the current for an extension of -2 µm begins at a very low leakage. Considering the diode structure in its entirety, a p+ diffusion is entirely contained within an n-well which sits on a high resistivity n- substrate. This provides excellent isolation at lower biases, however edge breakdown soon sets in as the current increases exponentially. As the extension is increased to 0 µm, the reverse bias current increases significantly due to the new p+/n- diode formed at the periphery. In addition, a clearly defined breakdown voltage begins to appear at 5.5 V. The IV curves behave similarly at 3 µm extension and greater with a shift in the apparent breakdown voltage. There is also a slight increase in the curvature just prior to breakdown as the extension is maximized. This may be an indication that the bulge in the electric field is incorporating more free carriers from regions surrounding the central junction. Similar results are obtained for diodes from wafer #3 (Figure 2.14). The difference between wafers #3 and #7 is that wafer #3 contains a gettering layer, where #7 does not. The presence of this gettering layer is observed to drive down the reverse bias leakage current from 50 µA to 1 µA. The reason for the relatively soft breakdown curves is the large series resistance presented by the high-resistivity substrate through which the current must flow in the vertical contact scheme. It is unclear why the curves in Figure 2.13 also do not exhibit a strong series resistive component.

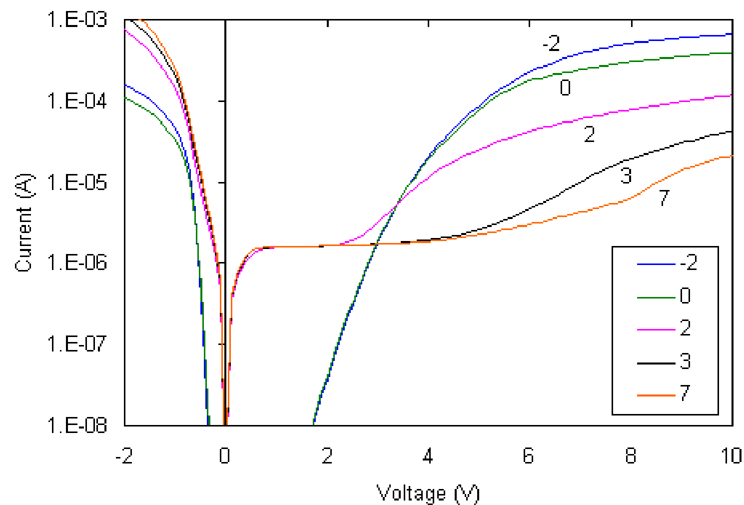


Figure 2.14. Reverse bias current-voltage characteristic for vertically contacted 50 µm circular JTE structure on wafer #3. Extension values (in µm) are listed in the legend.

A similar trend of increasing breakdown voltage can also be observed in the low-resistivity p-type wafers (#9-16). For instance, a 10 μm circular JTE diode on wafer #11 (Figure 2.15) finds its ideal junction termination extension value somewhere between 0 and 1 μm. This extension is the location at which the breakdown voltage stabilizes at ~7.5 V. Any additional extension (up to 8 μm) does not appear to significantly alter the current-voltage response.

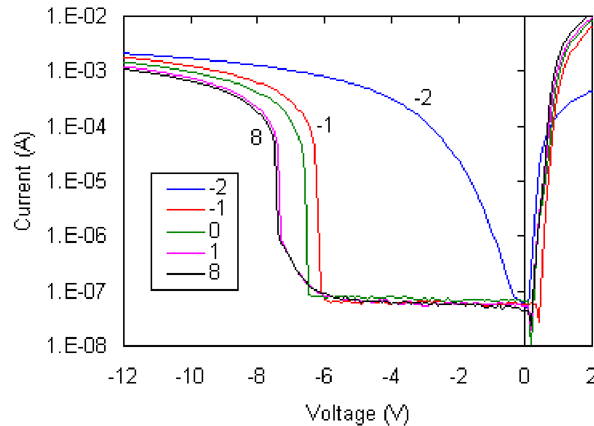


Figure 2.15. Current-voltage characteristic for vertically contacted 10 μm circular JTE diode on wafer #11. Extension values (in μm) are listed in the legend.

The premature breakdown in the -2 μm extension case is due to edge breakdown of the n+ which has been fully contained by the p-well. That is, the n+/p-well diode functions just as if the n+ region were located on an entire wafer of p-well-type doping.

Evidence of this edge breakdown can be seen in the EL images for a similar square diode with a -2 μm extension (Figure 2.16).

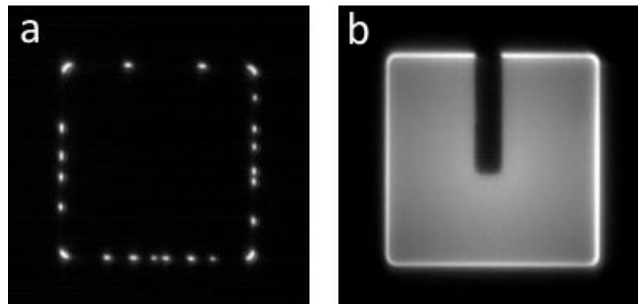


Figure 2.16. Electroluminescence images for 50 μm square JTE diodes with a -2 μm extension on wafer #11 (a) and #3 (b). Edge breakdown predominates in both.

Note first that while the bulk of the wafer #3 diode appears reasonably bright in Figure 2.16b, the periphery appears brightest, thus localizing the region of primary breakdown. The perhaps peculiar discontinuous breakdown for the wafer #11 diode (Figure 2.16a) was observed to evolve over a period of seconds from an originally continuous peripheral glow. One indicative feature is the prominent glow from the four corners, where the three-dimensional electric field is predicted to be greatest. The mechanism for the clustering of the remainder of the edges remains to be identified but may be influenced by overbiasing of the diode. The time evolution of this clustering is shown in Figure 2.17 for a similar -1 μm extension JTE diode, also from wafer #11. After one minute, the pattern of peripheral breakdown was sufficiently stabilized.

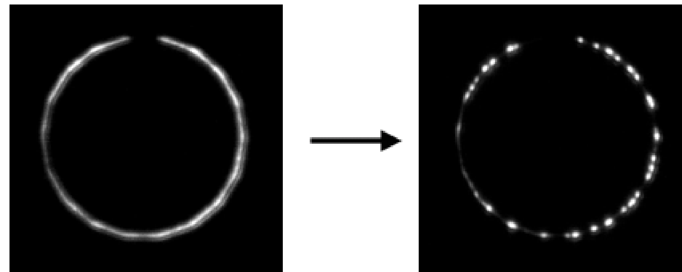


Figure 2.17. Evolution (left to right) of peripheral breakdown clustering of a period of one minute for a 50 μm JTE diode from wafer #11 with -1 μm extension, biased at 15 V.

Having illustrated the edge breakdown for JTE diodes with negative extension, we next focus on the characteristics of JTE diodes with positive extension. The breakdown characteristic and EL image is shown in Figure 2.18 for a JTE diode with circular contact this time. That is, instead of a point contact with a small via, an annular contact surrounds the diode. Note the steep breakdown voltage at -13 V, where tunneling is presumed to not contribute to breakdown. The EL image indicates a particularly smooth and well-confined breakdown. Were it not for the relatively high leakage current, this diode would present an ideal structure with which to construct a silicon photomultiplier. Indeed, no Geiger-mode behavior was observed from this diode. That is, no avalanche behavior was observed, presumably due to the excessive leakage current. Various radial locations for this circular ring contact also were tested, and identical results were obtained throughout.

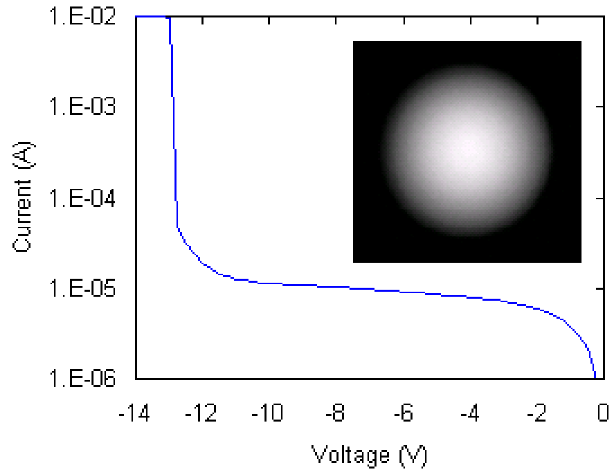


Figure 2.18. Current-voltage characteristic and EL image for vertically contacted 50 µm JTE diode from wafer #3 with circular contact and 8 µm extension.

Similar results are obtained when the diode is contacted instead at a single 5 µm point at the periphery (Figure 2.19). Notice that even though the contact is located at the edge, that the diode breaks down preferentially nearer its center.

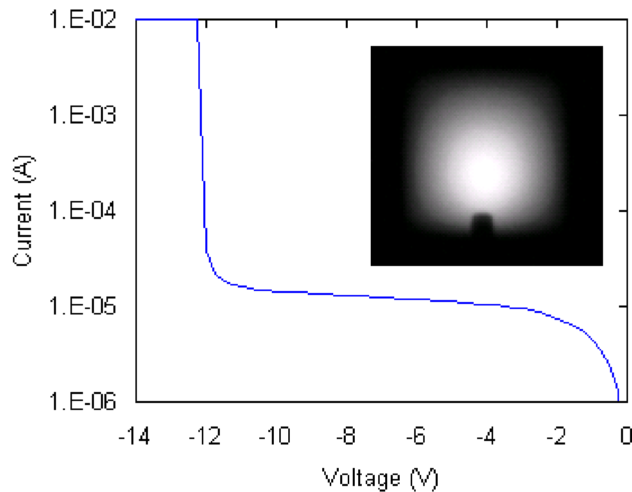


Figure 2.19. Current-voltage characteristic and EL image for vertically contacted 50 µm JTE diode from wafer #3 with 5 µm point contact at edge and 8 µm extension.

The contact hole size and location with which a vertically biased JTE diode is contacted may have an impact on the electric field. However, as illustrated in the electroluminescence images in Figure 2.20 for n+/p-well diodes on both (n-) and p-type wafers (#3 and #11), the via size of 3 to 5 µm and via location do not appear to alter the electric field.

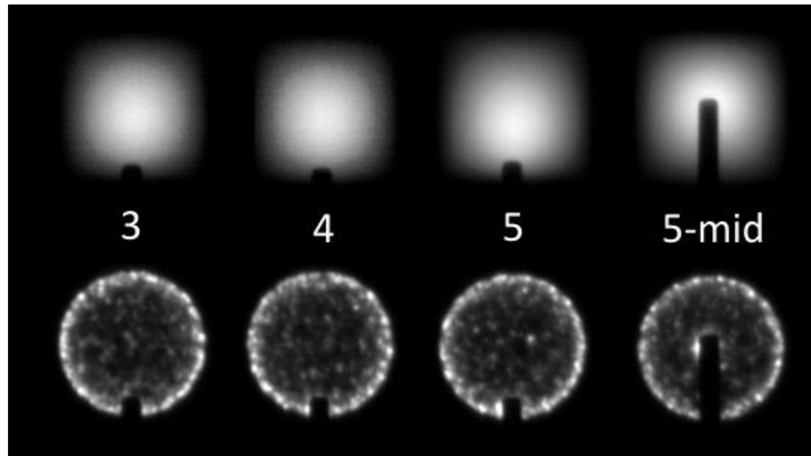


Figure 2.20. Electroluminescence images from n+/p-well JTE diodes from wafers #3 (top) at 15 V and #11 (bottom) at 20 V. Numbers indicate via size in microns.

It appears that edge breakdown is significant if not dominant in the p-type wafer diodes, which may be a consequence of the interaction between relatively high doping concentrations and the higher substrate impurity concentration. This may also be a consequence of an unsuccessfully gettered wafer or diode area.

It is interesting to note that the JTE diodes from wafer #11 (bottom of Figure 2.20) exhibit significantly less uniform breakdown than the diodes from the high-resistivity n-type wafer #3. However, when observing a similar JTE diode from wafer #12 in Figure 2.21, a uniform glow is observed. The suspected difference between the "speckled" and the uniform EL patterns is a result of the greater gettering temperatures and concentrations applied to wafer #12.

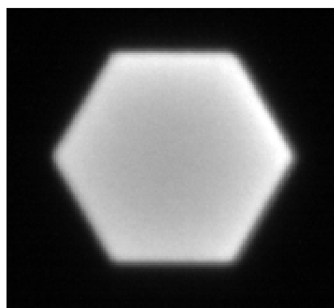


Figure 2.21. Electroluminescence image for a vertically-contacted p+/n-well JTE diode from wafer #12 with a 8 μm extension and circular contact.

A similar argument was also presented previously in reference to Figure 2.8. It is important to understand that a smooth, *well-confined* EL breakdown pattern can be obtained from diodes on both high- and low-resistivity n-type and p-type

wafers. This however is not a guarantee of good Geiger-mode or avalanche performance. In fact it seems unlikely that any amount of gettering can improve the CZ-grade wafers to the point of their high resistivity float zone counterparts.

2.3.B. Summary

The fabricated test diodes presented in this section point toward the successful application of the junction termination extension technique for reducing or eliminating premature edge breakdown. Because close-packing of these diodes is an eventual goal, the dead (or less active) region comprising the junction extension should be limited as much as possible. After determining the minimum extension, other factors must be considered such as excessive series resistance from charge transport through high-resistivity silicon. This series resistance does not allow avalanche current to return to a quiescent state before the next avalanche, and so no Geiger-mode behavior can be observed. Another parameter to consider is the reverse bias leakage current. If the leakage current is due to surface leakage, then it may only play a minor role in the generation of spurious avalanches. If however the leakage current contains a significant contribution from bulk generation sites, then no avalanche behavior would be evident due to the large density of free carriers in the active region. While the JTE technique proves useful in preventing edge breakdown, it is not suitable for use as a single-photon sensitive structure without being located in an extremely pure lattice with enhanced conductivity. Such a material is found in the epitaxial layers deposited for commercial CMOS processes and was not studied herein due to the incompatibility with multiple furnace doping steps.

2.4. Bevel Isolation Diodes

One intriguing method for edge breakdown prevention is beveling at the diode edge. If a planar diode were simply to be bevel etched at its periphery, the high electric field region would extend directly into the beveled surface and the passivation layer. This arrangement would lead to a significant thermally generated dark current because of the large density of lattice defects at the

boundary, the beveled edge. If however, the charge density gradient could be spread out at the boundary, then the electric field could be reduced. The bevel is used to effectively widen the diode junction at the surface, thus reducing the electric field.

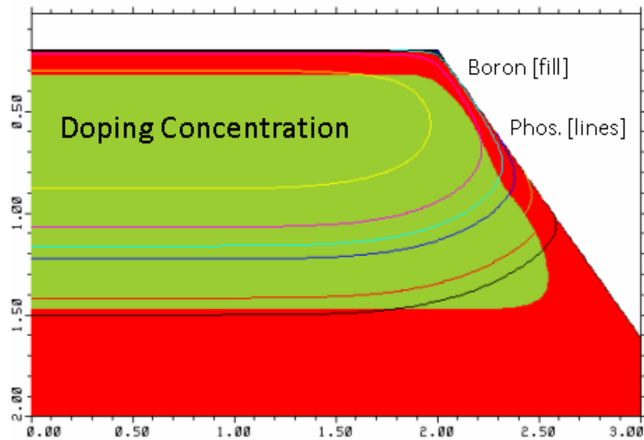


Figure 2.22. Doping concentration of a n+/p-well junction, where red and green fill represent boron p-well concentration, and contour lines represent phosphorus n+ concentration. The bevel and subsequent drive-in have created a reduced charge density gradient at the bevel surface.

The process flow is relatively simple, wherein a doubly-diffused junction is created and bevel etched. A subsequent drive-in pushes dopants down and out of the beveled surface (Figure 2.22). This is the mechanism which alters the electric field and is hoped to prevent edge breakdown. The electric potential created by this beveling process is shown in Figure 2.23, where the widened potential at the bevel edge results from the drive-in.

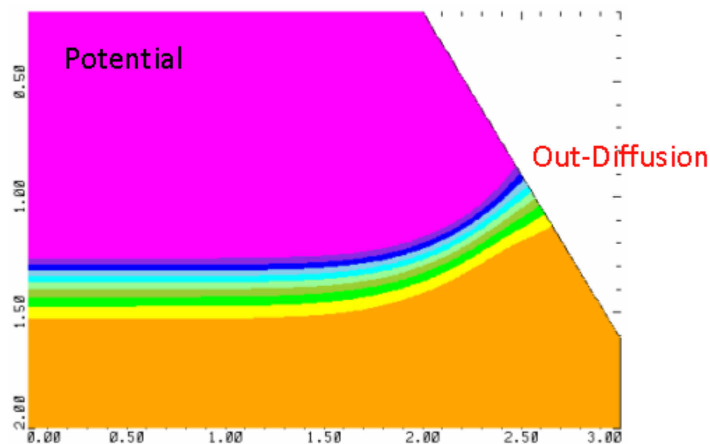


Figure 2.23. Electric potential of a bevel diode junction illustrating the widened potential at the surface.

The direct consequence of the widened electric potential is a locally reduced electric field as in Figure 2.24, although in this example the field is still quite high at the surface.

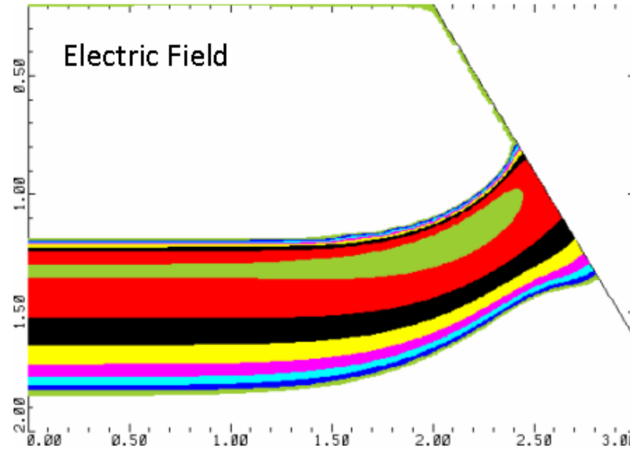


Figure 2.24. Electric field with slight reduction at the surface. Green central contour is region of highest electric field.

A careful selection of doping and drive-in parameters must be made in order to reduce as much as possible the surface electric field.

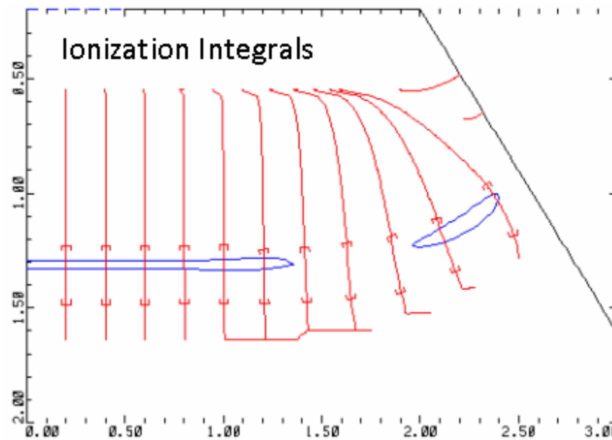


Figure 2.25. Current paths are illustrated for electrons beginning at 0.5 µm, distributed uniformly through the diode. The enclosed regions at the diode junction indicate regions where impact ionization will lead to sustained avalanches.

In order to illustrate the features of this simulated diode, we can take several imaginary electrons and start them at a depth of ~0.35 µm (0.55 µm in Figure 2.25). The n+/p-well diode accelerates these electrons toward the junction, or in several cases, towards the bevel edge. The bracketed regions along the current

flow lines indicate electric fields greater than 3×10^5 V/cm. In Figure 2.25, any flow line crossing a circled region (ionization integral) represents the certainty that that electron will cause a sustained avalanche through impact ionization. Although there still is a large electric field near the surface, the actual bevel surface has an ionization integral less than one, and the avalanche probability will be reduced at the surface. This does not however preclude the partial amplification of electrons originating at the surface from generation sites there. Again, fine-tuning of process parameters must be made in order to realize this diode.

One additional and perhaps significant benefit of beveling diode edges is the reduction in optical crosstalk when arrays of diodes are placed in close proximity to one another. The bevel can act as an optical barrier to prevent hot-carrier emissions [Chy56] from landing in and triggering neighboring diodes. The bevel of a silicon diode will most likely be in contact with a passivation oxide of index of refraction 1.46. The majority of hot-carrier emissions (500 to 1200 nm [Aki98]) emitted from the bulk of the diode, which would cause optical crosstalk, will reach the diode edge at an angle of ~35° normal to the bevel surface. This is due to the 55° angle between selectively etched (100) and (111) planes in crystalline silicon.

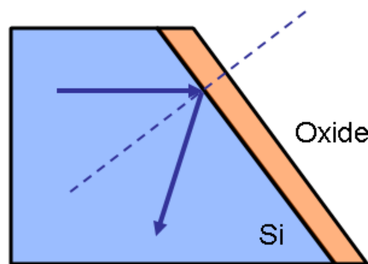


Figure 2.26. Etched silicon with 55° bevel edge illustrating total internal reflection of hot-carrier emissions arriving parallel to the upper surface and 35° normal to the bevel edge.

Snell's law indicates that the critical angles for total internal reflection are [20, 25]° for n_{Si} = [4.3, 3.5] at [500, 1200] nm. Thus the majority of hot-carrier emissions will be internally reflected at $90°-2(35°)=20°$ from vertical towards the bottom of the substrate. No optically reflective or opaque barrier should be

needed on top of the oxide, and the bevelled edge need only be as deep as the junction depletion depth.

Another possible benefit is the predicted packing fraction or geometric efficiency obtainable with bevel isolated diodes. If very little silicon real estate is needed for inter-diode electrical and optical isolation, then the probability of detection for an array of diodes can be significantly improved. We focus next on how to translate this bevel etching theory into realized diodes.

2.4.A. Silicon Etching

The selective removal of silicon can be achieved by solid, liquid, or gaseous methods. Solid methods entail mechanical abrasion by fine particles akin to sandblasting, however there is very little application for this gross method. Wet etching via liquid (typically alkaline) chemicals is historically the most popular method for etching silicon. Dry etching via gasses involves a high energy plasma to effectively remove silicon. One exception is XeF_2, a gas which selectively etches silicon without the need for additional energy. The cross-sectional profiles of dry etched silicon are often extremely anisotropic, with nearly vertical sidewalls resulting from the directionality of ion bombardment. Of primary concern for clean silicon processing is the energetic UV and ion damage caused by the high energy plasma. There are also very specific passivation and contamination concerns which depend highly on the tool, its configuration and use history.

Wet etching of silicon typically occurs anisotropically due to the simple crystalline structure of silicon. This anisotropy coincides with the specific lattice planes of crystalline silicon. The chemical and electrochemical theory underlying this complex process is still being developed, but can best be understood from an atomistic point of view. Lattice planes with a higher number of bonds tend to etch more slowly. Atoms on a (111) surface for instance are tightly bound to three atoms, whereas each atom on a (100) surface is only linked to two neighbors. This leads to pyramidal formations on (100) aligned wafers as depicted in Figure 2.29.

The reader is referred to a recent review paper [Gos07] for a proper understanding of the dynamics of pit nucleation and step propagation.

All etching processes require adequate masking to define useful planar features. The ability of a mask to withstand a given etch is represented by its *selectivity*, defined as the ratio of etch rates between the mask and the underlying material. Selectivities inform the thickness of mask required to etch a given depth into the silicon. For instance, etching through a 500 μm wafer requires an etch selectivity of >500 if the mask were to be limited to 1 μm thick. Many dry etches processes are selected for their good selectivity to easily applied photoresist. Other wet and dry etch chemistries require a more dense mask, such as silicon nitride or oxide. The method of mask application can also have a dramatic impact on the mask etch rate. For example, low temperature CVD oxides tend to be less dense and less ordered and therefore etch more quickly than a thermally grown oxide. Sputtered oxides and nitrides etch faster still. Metals may also be used as masking materials, however they are not considered here due to their potential for bandgap contamination.

Because silicon wet etching is so axis-specific, alignment of planar features must coincide with crystal orientation or feature asymmetries will evolve. Wafers are all manufactured with *flats* ground into their edges which locate the crystallographic axes. The wafer flat grinding tolerance over a batch of wavers is typically 1°, but can be as high as several degrees. A premium may be paid for sub-0.1° tolerance flats using x-ray crystallography. Alternatively, the very first step of the process can be a set of alignment marks followed by a wet etch from which each wafer can be optically aligned to ~0.025° [Cha05]. Simple wafer flat alignment has been used in this work due to the simplicity of features and shallow etch depth.

2.4.B. Chemical Etchant Requirements

The requirement for a low-contamination etchant necessitates the use of high grade chemicals. There are many industries which may use identical chemicals, and a premium may be paid for varying degrees or methods of purification. For

microfabrication, “electronic grade” chemicals are desired which represent both small particulate sizes as well as low absolute concentration of specific impurities (e.g. Na). Any impurity metal or salt can be detrimental to bandgap quality for avalanche diodes in ultra-pure silicon. Therefore, a solution is desired with extremely low concentrations (ppb) of impurities such as Na, K, Cr, Fe, Zn. The historical etchant has been potassium hydroxide (KOH) but is ill-suited to CMOS-clean device processing because of the potential for potassium contamination of devices, furnaces, etc. Ethylene-diamene-pyrocatechol (EDP) is another viable etchant with improved performance, but is extremely hazardous and is also not CMOS-compatible.

The CMOS-clean silicon etchant of choice is presently tetramethylammonium hydroxide (TMAH), containing only carbon, oxygen, hydrogen, and nitrogen as $C_4H_{13}NO$. Given the widespread use of TMAH in the electronic display industry, economies of scale have dramatically increased the availability of ultra-high purity TMAH in recent years. The current widely accepted standard method for low-level impurity analysis is inductively coupled plasma mass spectrometry (ICPMS). In this technique, a sample is aerosolized into an argon stream which is then ignited as a plasma. The plasma atomizes the constituents before passing them on to a mass spectrometer for ~ppt level analysis. Graphite furnace atomic absorption (GFAA) also has been used during the development and adoption of ICPMS.

2.4.C. TMAH Process

The bevel etching process in this work begins with high grade (3 PPB) 25% TMAH from Moses Lake Industries Inc., a U.S. subsidiary of Tama Chemicals in Japan. Typical contamination limits (from supplier) from GFAA and ICPMS were less than 1.0 ppb for the vast majority of contaminants. Ion chromatography verified chloride levels less than 10 ppb, and a laser particle counter verified less than 15 particles/ml greater than 300 nm. Titration verified CO_2 levels at less than 30 ppm. At the time of writing, this product cost was $40/gal. While these details may seem superfluous, the historical trouble with etching repeatability may have

been due to differences in chemical preparation and purity. Many historical problems in silicon processing (e.g. microplasmas) were all but abolished with the advent of more pure and consistent materials. It is important only to make exact comparisons between processes with identical parameters.

The physical etching setup involves a heated reflux system. One liter of 25% TMAH is poured into a 2 liter beaker and then placed on a hotplate. The temperature is raised to 60-85 °C, and a glass bulb condenser is placed atop the beaker. Cold water flowing through the condenser causes evaporating TMAH to flow back into the beaker, thus keeping the 25% concentration at equilibrium throughout the etching process. This reflux system was developed for multi-hour etches at 90 °C where the amount of evaporation is much greater. This reflux method is retained herein for stability of the etchant and repeatability during many hours of short etches. Care is taken to remove impurities from the etchant, the samples, and any contacting labware by cleaning with electronic-grade solutions, 18 MΩ-cm deionized water (DI), and TexWipe II brand cleanroom wipes.

Masking is performed by depositing a 1500 Å LPCVD oxide and opening areas to be TMAH etched. Because of the inferior quality of deposited oxides as compared to thermally grown oxides, a high temperature densification step is performed. In this case, the wafers are pre-furnace cleaned and loaded into a furnace at its static temperature of 600 °C for 30 min. Higher temperatures would cause greater degrees of densification, but the thermal budget of most processes herein require a lower temperature to avoid appreciable diffusion of the donor or acceptor impurities.

After unloading, the wafers are immediately given a 5% HF dip for 30 s to remove any remaining native oxide. They are then placed face-down in a horizontal Teflon cassette which sits in the TMAH beaker. A two inch magnetic stirrer bar rotates at 100-400 rpm and agitates the TMAH to improve the uniformity of the etch over the entire 10 cm wafer. After etching, wafers are immediately removed into a DI water bath and rinsed for 2 min. Optical inspection (Figure 2.27) is used to determine the degree of undercutting, etch uniformity, gross roughness, and mask survival.

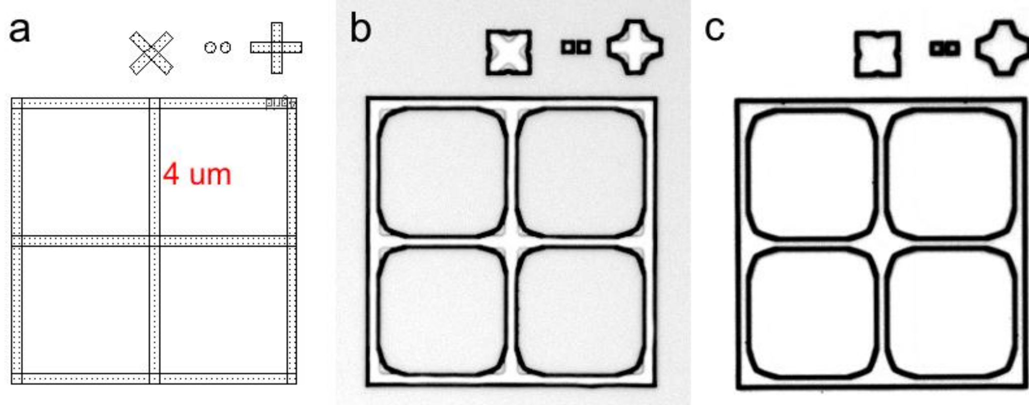


Figure 2.27. TMAH (25%) silicon etch at 60 °C for 10 min (2.3 µm deep). Original test feature mask (a) with 4 µm wide features in a 1500 Å *thermal* oxide. Microphotographs after etching (b) and after mask removal (c). Undercutting at corners is evident.

Scanning electron microscopy can also be used to confirm the smoothness and uniformity of the etch as depicted in Figure 2.28 and Figure 2.29.

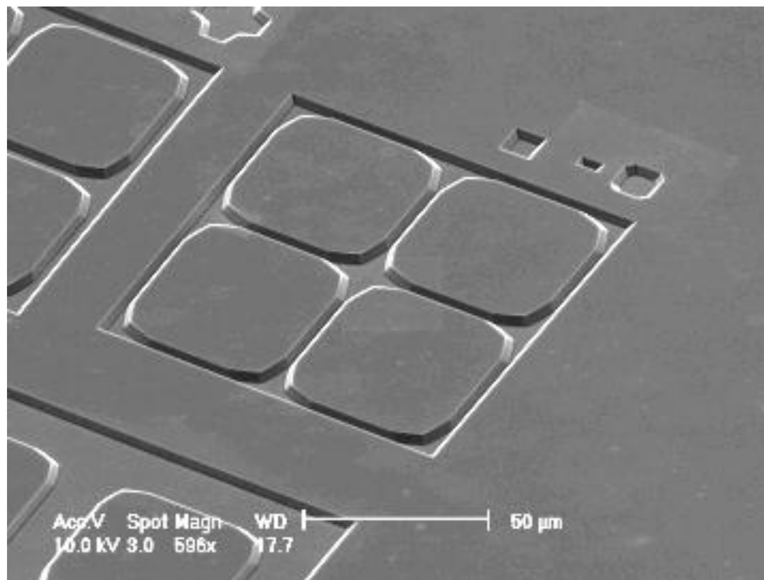


Figure 2.28. Scanning electron micrograph of 2.4 µm TMAH etch in a thermal oxide mask (removed). Scale marker is 50 µm. Note the presence of undercutting, i.e., non-square corners.

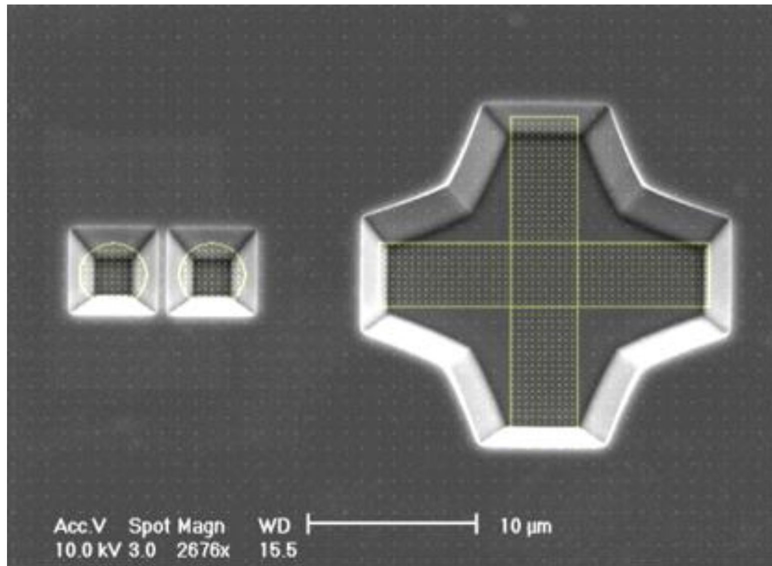


Figure 2.29. Scanning electron micrograph illustrating 25% TMAH wet etch and original mask design of area to be etched. Undercutting at convex corners is evident. Scale marker is 10 um.

A variety of temperatures and times were tested in order to find a controllable etch rate at low depths (1-5 µm) and reasonable times (<1 hr). Vertical etch rates (Figure 2.30) were obtained with a Dektak stylus profilometer after mask removal in HF. Shorter etches suffered from process variation in the time required to load and unload samples.

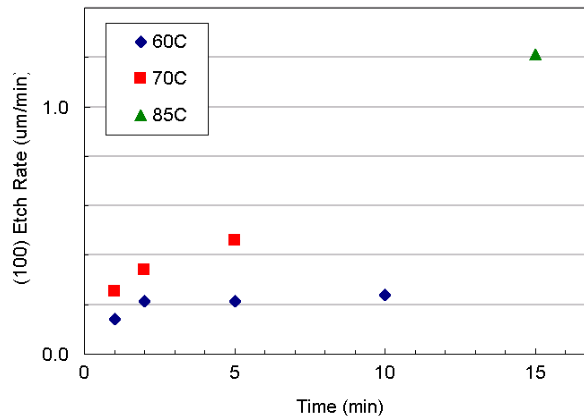


Figure 2.30. Silicon etch rates (µm/min) for 25% TMAH at several temperatures as determined by a Dektak surface profiler on 16 µm wide features.

A 60 °C TMAH temperature was selected to provide an appropriate etch rate of ~210 nm/min. The bevel angle (53.7°) and rms surface roughness ((100): 1.3 nm, (111): 6.1 nm) were determined through atomic force microscopy (AFM) using a 10 nm tip in tapping mode (Figure 2.31 and Figure 2.32). The slight deviation of

this bevel angle from the theoretical angle between lattice planes is due to a less than infinite etch selectivity between crystallographic axes.

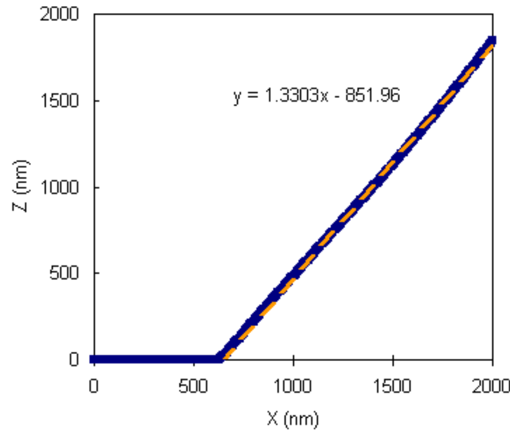


Figure 2.31. AFM trace of flat bevel trench transitioning to bevel edge, indicating 53.7° angle and bevel smoothness.

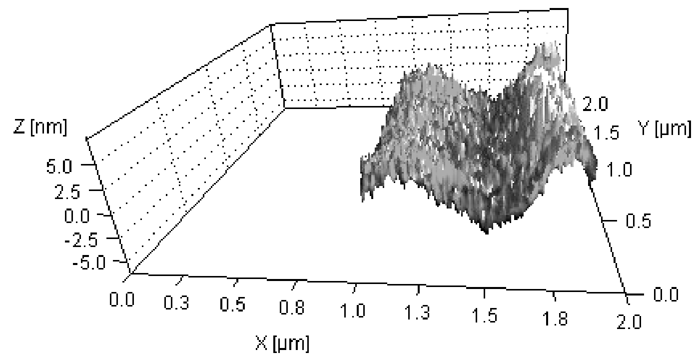


Figure 2.32. AFM image (256 x 256 points) of bevel edge (software leveled) illustrating 3 nm rms bevel smoothness. Note nm-scale in Z axis and µm-scale in X, Y axes.

The 3 nm rms smoothness of the 25% TMAH-etched bevel edge is of the same order as polished silicon and therefore no spikes in electric fields are expected. The degree of passivation expected to be achieved is thus comparable to that of the top surface of the polished wafer. The exposed bevel surface (111) will experience a slightly different mismatch during passivation and it will also have a slightly different oxidation rate. Having developed a successful process for smooth, shallow, bevel etching of silicon, the next task is to form diodes with this same procedure.

The degree of undercutting can be reduced by adding a surfactant which effectively reduces the surface tension [Gos09]. Although the mechanism is not

fully understand, the surfactant has the effect of dynamically altering the etch rate of certain planes. See [Gos09] for a full review of proposed theories. The surfactant chosen was Triton X-100 (CAS 9002-93-1), which is a non-ionic detergent and is available in highly purified grades. It is also CMOS-clean, consisting of only carbon, oxygen, and hydrogen as $C_{14}H_{22}O(C_2H_4O)_9$. The amount used for silicon etching is approximately 0.1% per volume of TMAH solution, or about 20 drops per liter. Being a high quality detergent, removal of residue may be an issue so cleanup requires multiple sequences of methanol and deionized water rinses for all labware and samples.

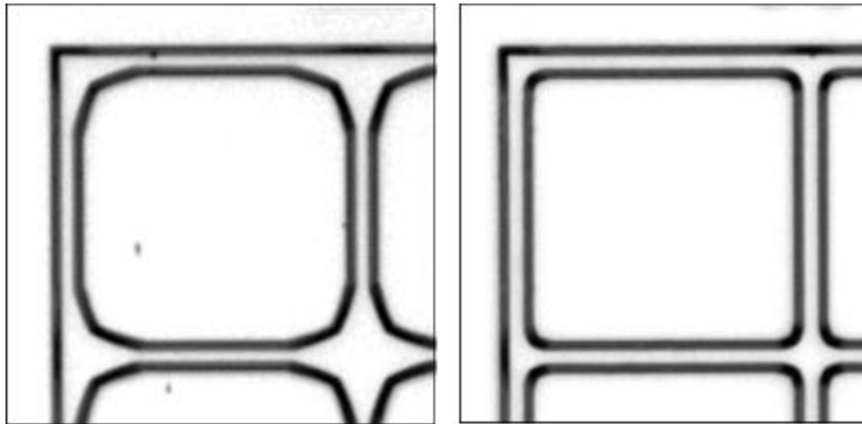


Figure 2.33. TMAH etch (2.4 µm deep) in silicon without surfactant (left), and with 0.1% v/v Triton X-100 surfactant (right).

The effects of the additional surfactant can clearly be seen in Figure 2.33 at relatively shallow etch depths. An undercutting ratio can be defined as the lateral distance removed from a convex corner divided by the etch depth. An improvement in undercutting ratio of 4.2 to 1.5 has been observed. The rounded corners are more advantages as they do not provide sharp geometries which can lead to spikes in the electric field.

An additional restriction on etching of diodes is the effect of boron as an etch stop at concentrations above $\sim 1\times10^{19}$ cm^{-3} [Sei90]. This fact is exploited in MEMS device processing in order to perform controlled through-wafer etching. Phosphorus however does not function this way and is readily etched by TMAH.

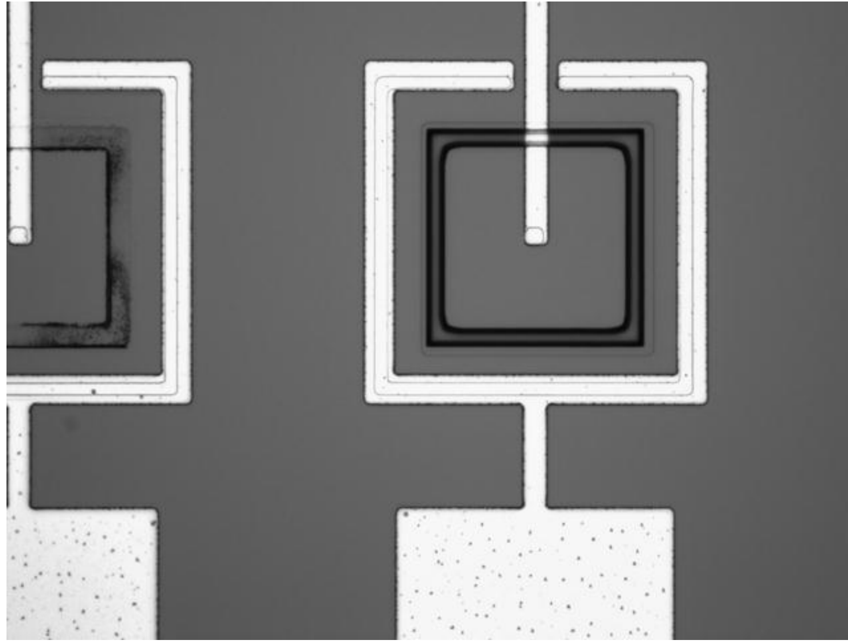


Figure 2.34. Optical micrographs illustrating poor etch rate of boron-doped silicon (left) and successful etching of highly phosphorus-doped silicon (right).

One can clearly see the difference in etch quality between the left (boron-loaded) and right (phosphorus-loaded) diodes of Figure 2.34. The etched trench is located just inside the edge of the highly doped region of the diode.

2.4.D. Bevel Etched Diode Results

Having established a successful process for the smooth and conformal bevel etching of silicon, we focus on several realized devices which illustrate some key considerations. A cross section of the bevel-etched diode structure initially considered appears in Figure 2.35. The fabricated diode appears in Figure 2.36. No passivation layer is applied, and the entire wafer is produced with only two masks: one for the bevel, and another for the metal layer.

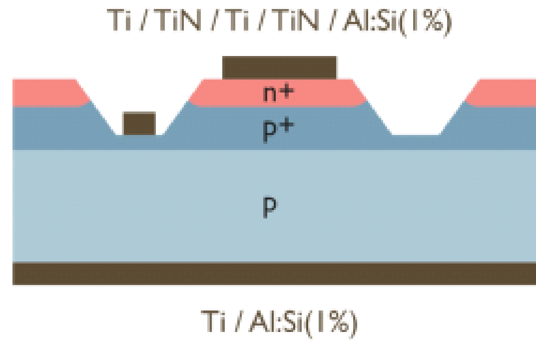


Figure 2.35. Structure of bevel diode with both top and bottom substrate contacts. The p+ layer is below the $1{\times}10^{19}$ cm^{-3} boron etch stop threshold.

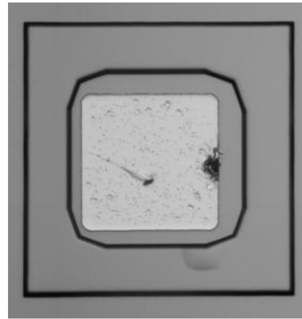


Figure 2.36. Photomicrograph of the 80 µm wide fabricated bevel diode with top contact to raised island.

The entire wafer receives a deep p+ diffusion followed by an unpatterned n+ diffusion. Diode islands are created via TMAH etching to a depth greater than the diode junction. Top contacts are made to the n+ diffusion as well as the recess of the beveled trench. The trench contact need actually only be made to one trench on the wafer, which is why it does not appear in Figure 2.36. A backside contact is also applied.

The concept behind the metal stack at the top is to provide a low resistance metal contact which does not alter the junction below. All deposited metal contacts must be heated to destroy the ever-present native oxide and also to create a smooth alloy transition between silicon and metal. Silicon dissolves readily in aluminum at elevated temperatures and the aluminum will form spikes into the silicon as the silicon diffuses into the aluminum bulk [McC71]. These spikes can be microns long, in which case they can pierce through shallow junctions, effectively shorting their rectifying characteristic behavior. This is not a problem for backside contacts or for very deep junctions. One method developed to

mitigate this junction spiking was to pre-dope the aluminum with its solid solubility of silicon (~2%). A refractory metal like titanium or tungsten can also be used as a barrier to discourage diffusion of silicon into the aluminum [Gha78]. Further, titanium nitride has a columnar structure which when filled with added titanium becomes a very good diffusion barrier. This is the motivation behind the multi-layer metal stack indicated in Figure 2.35, which is guided by previous work [Wu05].

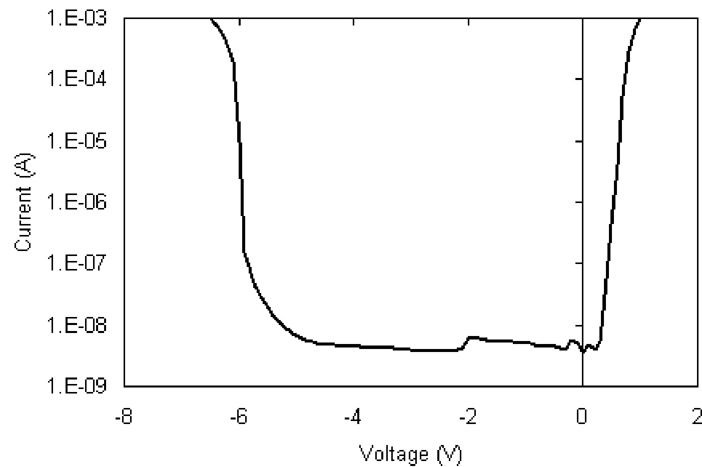


Figure 2.37. Current-voltage characteristic for diode in Figure 2.36.

Upon examination of the current voltage characteristic in Figure 2.37, we see a rather typical forward bias diode behavior and a rectifying reverse bias behavior until the breakdown voltage of -6 V. The noise in the curve at -2 V is due to system mischaracterization and should not affect interpretation. There is a reasonably stable reverse bias current of several nA, which incidentally was confirmed to be independent of diode size. This may indicate that surface leakage is playing a more dominant role. A forward biased diode (plus contact) resistance of approximately 400 ohms was obtained for this 80 μm wide diode. Just prior to breakdown, at about -5 V, there is an exponential increase in reverse bias current. This can either be due to enhanced tunneling, from impact ionization, from punch through at the bevel, or some combination of the three.

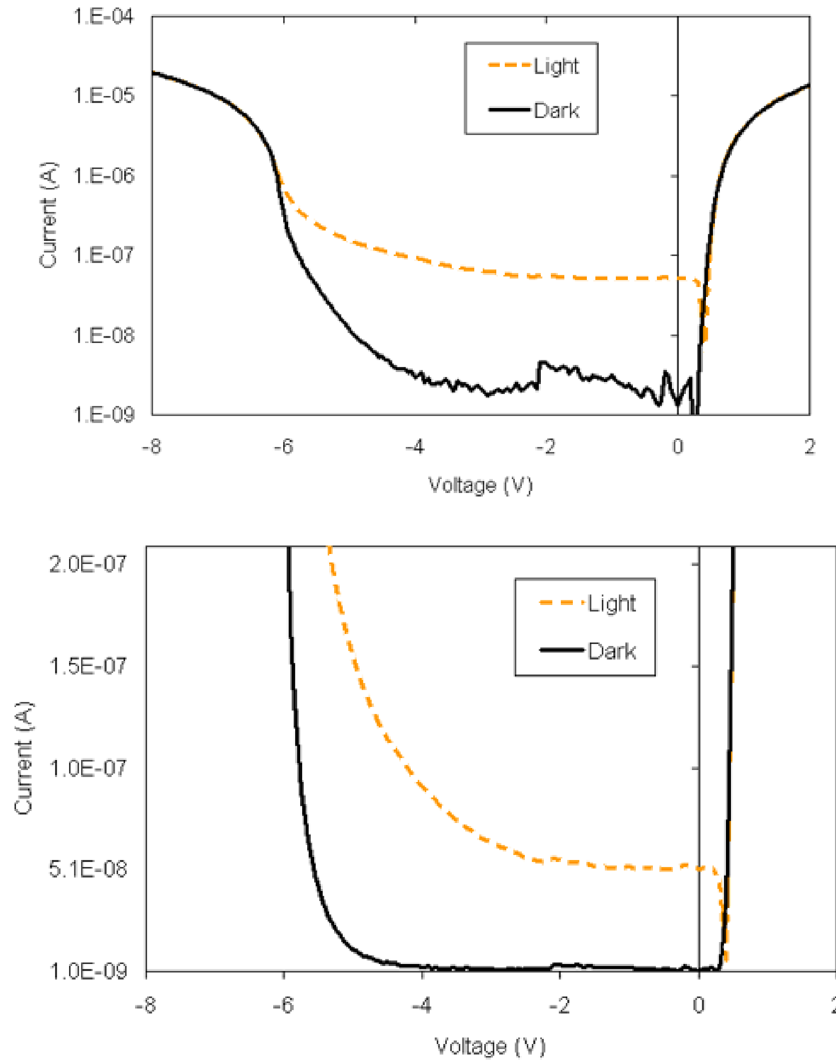


Figure 2.38. Current-voltage characteristic for diode in Figure 2.36, in series with 100 kΩ resistor, both with and without external photon stimulus. Log scale (top) and linear scale (bottom).

The theory for tunneling was originally developed for solid dielectrics [Zen34], but the energy-band theory turned out to be appropriate for semiconductors as well. In order to test whether or not we have an avalanche diode or a Zener (tunneling) diode, we can inject some fixed photon flux into the reverse bias junction and observe the diode response. If the diode were not sensitive to photons, then no increase in reverse bias current would be observed. However, we see a constant increase in zero-bias response to 50 nA. If the diode were to break down strictly by tunneling, then any additional photocurrent would simply add on top of the dark current. What we observe is in fact that the diode

responds with an amplified current as low as 2 V and steadily increases its amplification as the bias approaches the breakdown voltage.

Another confirmation that this diode is at least in part an avalanche diode is through the observation of the temperature coefficient of the breakdown voltage [Tya68]. That is, as the temperature is raised, to what degree does the breakdown voltage shift? When the temperature is raised, the increased atomic vibrations lead to an increased inter-atomic spacing, which in turn lowers the bandgap energy. Thus, less voltage is required for tunneling to occur. This corresponds to a negative temperature coefficient. On the other hand, when temperature increases, the added random thermal motion makes it more difficult for charge carriers to be accelerated. Therefore more voltage is required to achieve breakdown, and thus the temperature coefficient for impact ionization is positive. This polarity can be observed in any commercial avalanche or Zener diode. See the Transys Zener diode datasheet in Figure 2.39 for an illustration of the breakdown voltage temperature coefficient crossover at ~5.6 V. The actual transition between tunneling and impact ionization is graded and some authors [Sze07] suggest that this transition region is confined to within 4-6 times the bandgap. This would imply that avalanche diodes in silicon should have a room-temperature breakdown voltage greater than $6E_g = 6(1.1\text{ eV}) = 6.6\text{ V}$.

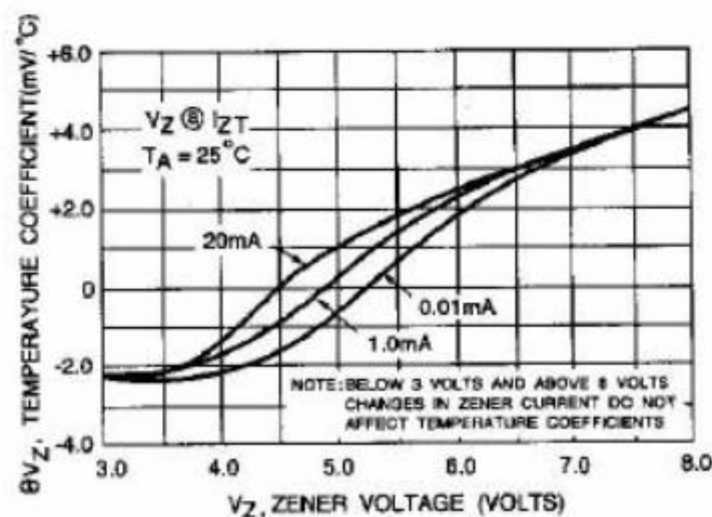


Figure 2.39. Temperature coefficient (in mV/°C) for Transys 1N4741A Zener diodes in silicon.

It is interesting to note that *avalanche* diodes (those with V_{BD} > 6 V) are actually marketed as *Zener* diodes, because of their near-identical performance in circuitry. Following this additional test for impact ionization or tunneling, the bevel diode above was subjected to a temperature controlled probe station, and the breakdown voltage was extracted from the current-voltage characteristic at each temperature (Figure 2.40).

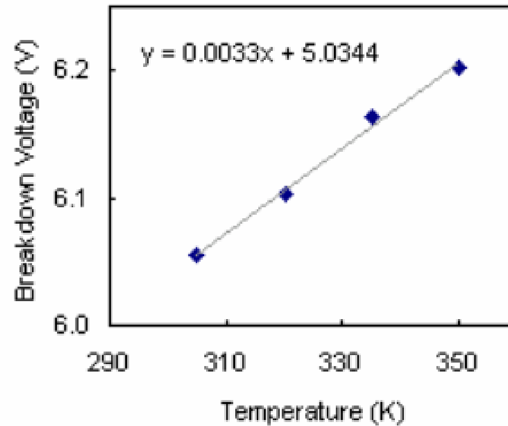


Figure 2.40. Temperature dependence (3.3 mV/°C) of the breakdown voltage of an 80 μm bevel diode.

The diode was observed to have a temperature coefficient of 3.3 mV/°C which indicates by its positivity that it is indeed at least partially an avalanche diode. The only other way of determining the balance of tunneling current to avalanche current is to somehow inject a known charge into the diode and measure the output current. This technique however is plagued by myriad assumptions and technical details which have made comparison and evaluation of data somewhat suspect. One particular line of commercial silicon photomultipliers (Hamamatsu MPPC) operating at 70 V have a temperature coefficient of ~60 mV/°C [Din10]. This is one motivation for producing avalanche diodes with a breakdown as low as possible, while remaining high enough to avoid any negative effects from band-to-band tunneling.

It would be instructive to somehow visualize the electric field before testing the Geiger mode behavior of this diode. The location of very high electric field regions can be probed by observing the steady-state reverse bias electroluminescence (EL) patterns from the hot-carrier emissions arising from the

recombination of energetic (e.g. avalanching) charge carriers. For the bevel diode above, a sensitive CMOS camera is fitted to a probe station microscope in order to record the two-dimensional location of those hot-carrier emissions (Figure 2.41).

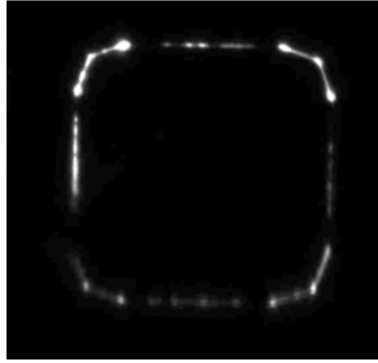


Figure 2.41. Electroluminescence pattern from vertically contacted 80 μm bevel diode operated at 10 V reverse bias and limited to 10 mA.

Unfortunately in this case, the electric breakdown appears to be wholly located at the periphery of the diode. Because this EL image was taken at 4 volts above breakdown, there is a possibility that the bulk breakdown is being overshadowed by the breakdown from the greater density of electron generation sites at the bevel edges. Even if this diode illustrated a uniform and confined electric field, the large contact atop the diode does not make for a sensible photodiode structure, and there is no passivation layer. Therefore, alternatives to this bevel diode structure will be presented in the following section. However, additional work should be done to optimize the doping concentrations and passivation layers before abandoning this extremely simple design.

2.4.E. Bevel Diodes Incorporating JTE Structure

Instead of creating bevel diodes with blanket or flood diffusions, a junction termination extension diode is first created, around which a trench is etched in order to combine the benefits of both processes. Because TMAH does not etch p+ layers ($>1\times10^{19}$ cm^{-3}) , only n+/p-well diodes will be studied, primarily from wafers #9-16.

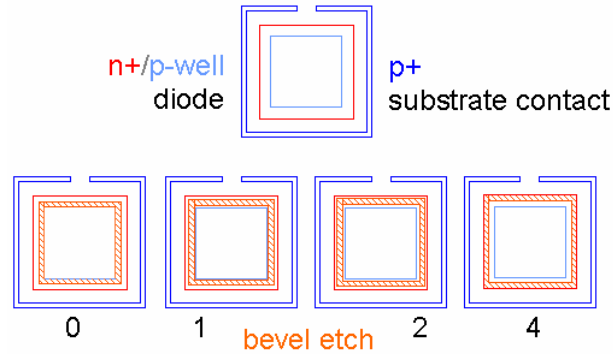


Figure 2.42. Revised bevel diode structure (top) and series (bottom) with bevel moat etch located at increasing distances from p-well. A p+ substrate contact surrounds the majority of the diode.

Like the JTE diodes, two contact schemes are available: vertical and lateral. Lateral contact is made between a center point contact and a ring surrounding the diode contacting the bulk silicon. A bevel trench some 4 µm wide encircles the active area (Figure 2.42). This bevel is located at increasing distances from the p-well diffusion defining the active area. The bevel should touch the edge of the p-well when the distance is zero.

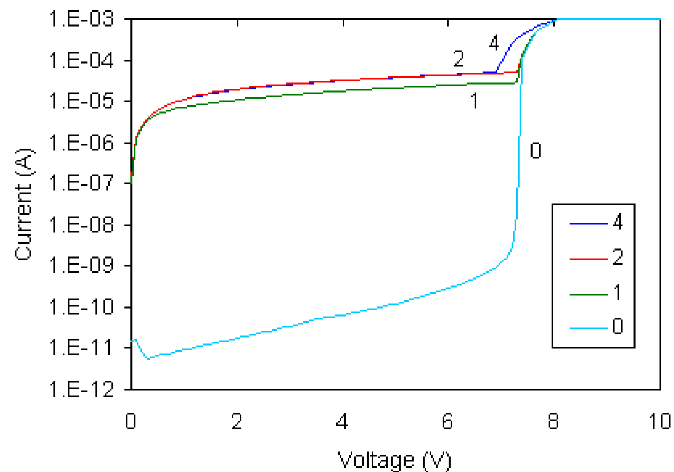


Figure 2.43. Square bevel diodes from wafer #11, contacted laterally. Distance in µm between bevel and p-well is indicated in legend.

Observing the reverse bias current-voltage behavior of a laterally-contacted square diode (Figure 2.34) from wafer #11 (Figure 2.43), a 7 V breakdown is evident for diodes of all bevel extension. However, only the 0 µm bevel distance diode appears to have a low-leakage blocking capacity. This is due to a low-resistance path that is opened up between the two lateral contacts. A dramatic

multi-decade rise in current is very well confined to just above the breakdown voltage, indicating low additional series resistance. A similar set of responses (Figure 2.45) is obtained from six-sided bevel diodes (Figure 2.44).

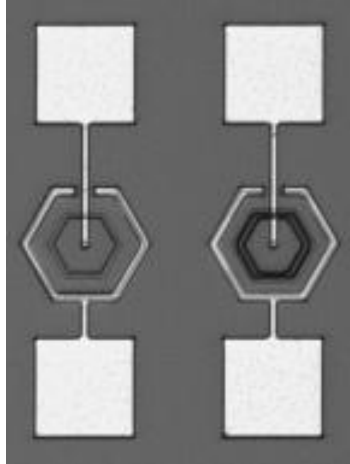


Figure 2.44. Hexagonal bevel etched diodes: p+/n-well (left) and n+/p-well (right). Not the slight asymmetry in the etch as the diode active area is longer in the vertical axis.

The fact that the 1 μm bevel distance diode is also low leakage is most likely a cause of the increased etch rates for the hexagonal planes. That is, the active p-well area is thought to be tightly bounded by the slightly over-etched bevel.

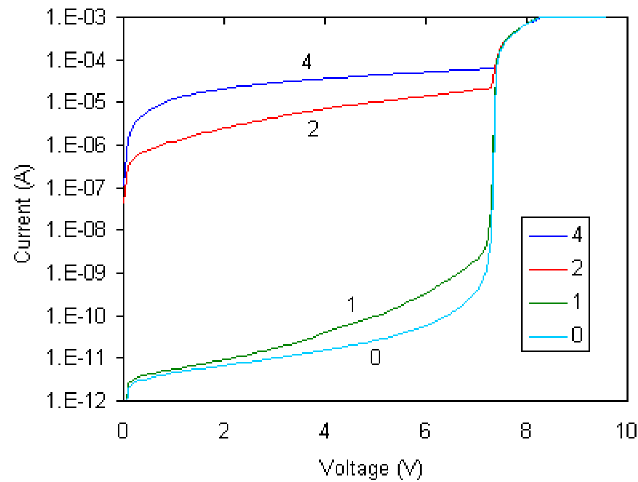


Figure 2.45. Hexagonal bevel diodes from wafer #11, contacted laterally. Distance in μm between bevel and p-well is indicated.

The low-leakage and high-leakage absolute current values agree well between the square and hexagonal shape. Not depicted are the forward-bias characteristics, which quickly reach the measurement saturation current. Although the high-

leakage device currents are significantly more leaky than those of low-leakage devices, they are still several orders of magnitude below the forward-bias currents.

Although the p+/n-well bevel diodes were not successfully beveled, one can still examine the n+/p-well diodes on wafers #1-8. This is because the extremely low-doping (~1×10^{-11} cm^{-3}) does not significantly contribute to the diode structure formed by the relatively highly doped n+ or p-well. Figure 2.46 presents examples of two such diodes from wafers #2 and #3.

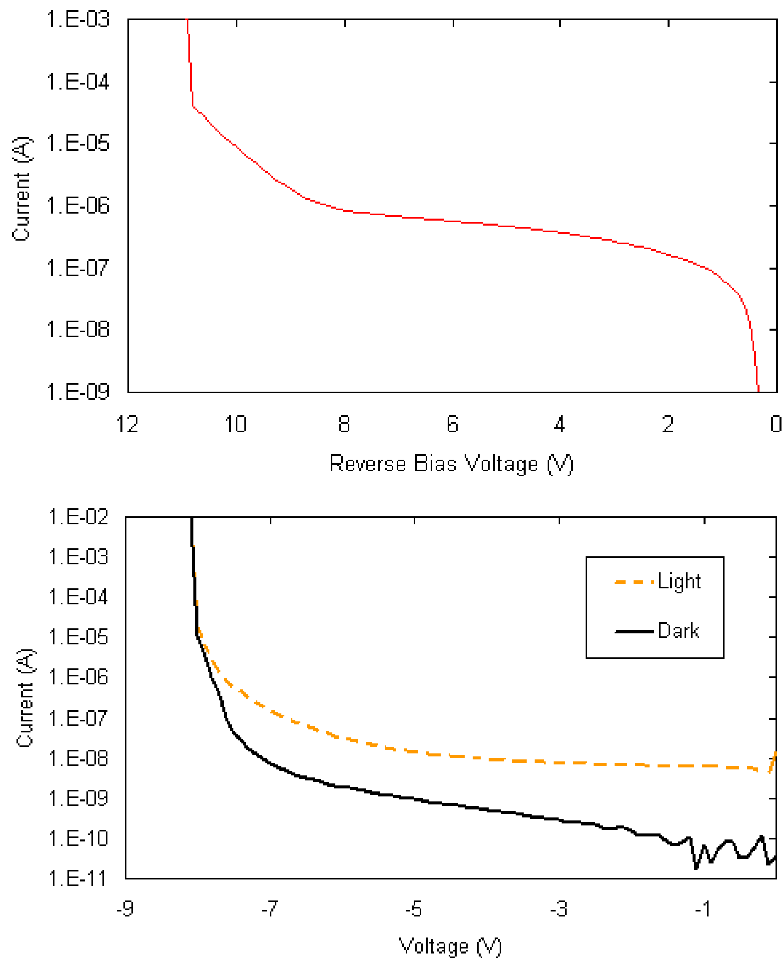


Figure 2.46. Steady state characteristic curves for *laterally* contacted n+/p-well bevel diode from wafer #3 (top) and #2 (bottom) with 0 distance between bevel and p-well. Response to light (top) indicates amplifying behavior prior to breakdown.

The deeper p-well diffusion of wafer #3 (top) results in a larger breakdown voltage of 11 V, as opposed to 8 V for wafer #2 (bottom). It is clear from the light response of the wafer #3 diode that some photoamplification is occurring beyond 5 V. The straightness of the #2 diode at 9-11 V would also seem to indicate an

exponential multiplication of the dark current before a much stronger electric field takes over at 11 V.

If no multiplication occurred (prior to breakdown), then the current curve would remain relatively flat until it met the dark current curve. Considering that the depletion region expands as reverse bias is increased, one would expect a greater fraction of photoelectrons to be swept to the contacts. However, photons are absorbed preferentially at the surface and are distributed exponentially as a function of depth. Because of the dual diffused doping profile, the depletion region will primarily expand into the lightly-doped region. That is, as the reverse bias is increased, an insignificant number of photoelectrons are included in the expanded depletion region (see for example Figure 2.10).

If contacted vertically, the hot-carrier emission patterns (Figure 2.47) from the bevel diodes appear reasonably well confined within the bevel, especially in the 0 μm bevel distance case. Some edge glow is apparent for the 1 μm and 2 μm distance diodes, while a rather unique centralization of hot-carrier emission is seen in the 4 μm diode. This increasing nonuniformity could be the result of higher leakage current. One possible explanation is that electrons closer to the edge of the diode experience a greater probability of being swept around the high field region through the opening presented by the 4 μm bevel distance. This would mean that those electrons which start out in the center of the diode would have a relatively increased probability of traveling through the high field region, and thus generating hot-carrier emissions preferentially at the center. It is proposed that what is being observed is the loss of carrier multiplication at the edges above breakdown which coincides with the increased leakage current below breakdown.

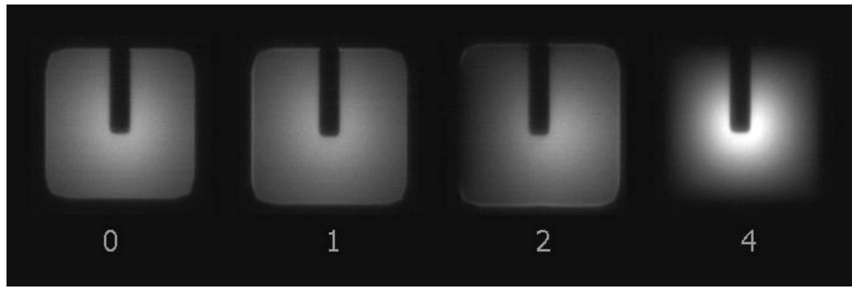


Figure 2.47. Electroluminescent images of *vertically*-contacted bevel diodes from wafer #3 at 10 V reverse bias. Numbers indicate bevel distance from p-well in µm.

One support for this conclusion can be seen in the electric field simulation of Figure 2.48, where current flow lines are shown to pass evenly throughout the entire area of the laterally-contacted bevel diode.

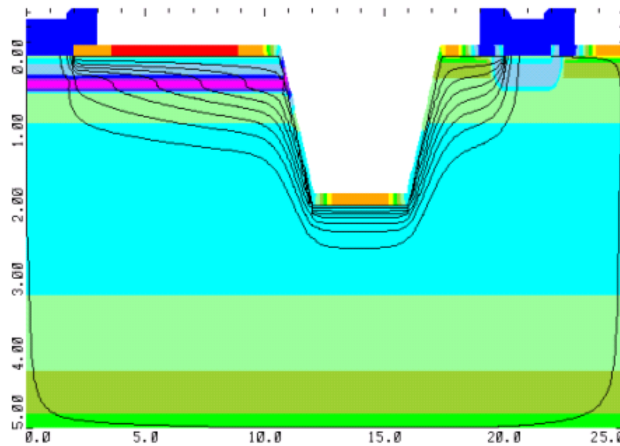


Figure 2.48. Simulation of electric field and current flow lines for a laterally contacted bevel etch diode. Note the even distribution of current flow lines throughout the active area on the left.

To further illustrate the features of various contact schemes and substrates, a series of electroluminescence images is presented in Figure 2.49.

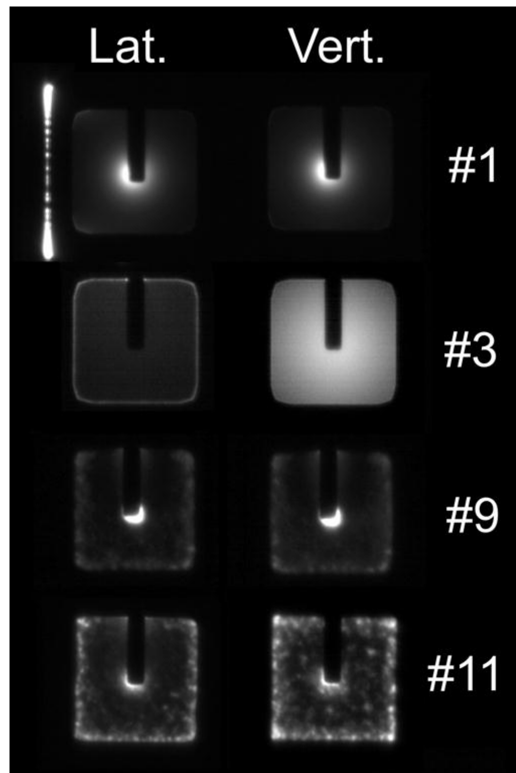


Figure 2.49. Electroluminescence images for laterally and vertically contacted n+/p-well bevel diodes from wafers #1,3,9,11, with 0 µm bevel spacing.

Recall that wafers #1 and #3 are high-resistivity n-type wafers, whereas wafers #9 and #11 are of low-resistivity p-type. An n-type substrate may be inappropriate for this doping structure because of the limitation in studying only n+/p-well diodes. This is evidenced in the difference in EL images from vertical to lateral contact schemes. Note that this is not the case for the p-type wafers; that both lateral and vertical contacts yield similar hot-carrier emission images. One final note is that many diodes exhibit significant breakdown at the point of contact. This most likely is due to overly-aggressive design rules which have led to misaligned contacts and potential junction spiking by the overlying aluminum.

2.4.F. Summary

A bevel etching process was developed to produce shallow etches with extremely smooth bevel edges. This process was performed with TMAH so that no ionic contamination would introduce electron generation sites or locations of premature breakdown. Simple bevel diodes proved to have low leakage current,

but appeared to break down preferentially at the bevel edges. A revised bevel diode, incorporating the JTE structure also was observed to produce low leakage currents (for some designs) and provided a well-confined hot-carrier emission pattern. However, the lateral contact scheme did not appear to produce a uniform electrical breakdown over the surface of the diode. The lack of a low-resistivity, high-purity lateral conduction path makes realization of this diode difficult with the present technology.

2.5. Guard Ring Diodes

A significant component of the leakage current observed from a simple double-diffusion avalanche diode is generated by electrons or holes which drift into the high electric field region from the substrate. This is particularly true for vertically contacted diodes, where the electric field can taper off gradually towards the bottom surface. Commercial transistor processes employ epitaxial layers, among other reasons, in order to isolate individual components from the substrate. Various doping techniques are employed to isolate devices by creating potential barriers, i.e., diodes. Some more involved processes create a patterned doped region (n buried layer) on the substrate before deposition of the epitaxial layer. Because this work relies on high temperature diffusion, epitaxial substrates are not feasible, and thus alternate methods must be used to isolate the junction from the substrate. It is hypothesized that a low-doped well can operate as an isolating tub or tank. So long as the well and the substrate do not break down at too low a voltage, charges in the active area junction should be limited to just those within the well.

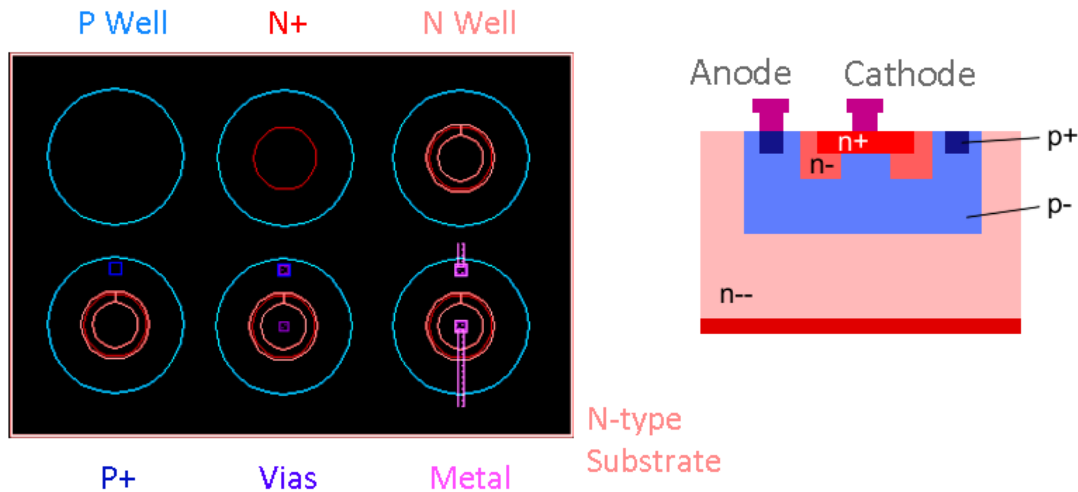


Figure 2.50. Guard ring design utilizing four diffusions: p-well (p-), n-well (n-), p+, n+. Top view (left) and cross section (right).

Given a high-resistivity n-type substrate (Figure 2.50), a low-doped p-well will effectively change the local substrate type to p-type. If a shallow n+ is then added inside the well, an n+/p-well diode is formed in addition to the deeper p-well/n- junction already established. Now that the diode is isolated, it needs to be protected against edge breakdown. From previous arguments, we know that the periphery of the n+ region will break down first, so we mitigate this unwanted effect by introducing a reduced concentration n-well ring around the periphery. This *guard ring* reduces the charge density gradient and thus a new peripheral diode is created which will break down at a much higher voltage. While preventing edge breakdown, this structure has eaten into the active area as well as the surrounding silicon real estate. The fourth and final diffusion is a p+ region where contact to the p-well is made. This high doping concentration will provide the necessary tunneling behavior to promote a good ohmic contact to the overlying metal. The choice of where to place the vias is not trivial, as a balance must be struck between large area, low-resistance contacts and maximum open area for photodetection.

2.5.A. Guard Ring Results

The guard ring structures realized are each 20 microns wide in active area with a 16 micron diameter opening in the cathode metallization (Figure 2.51).

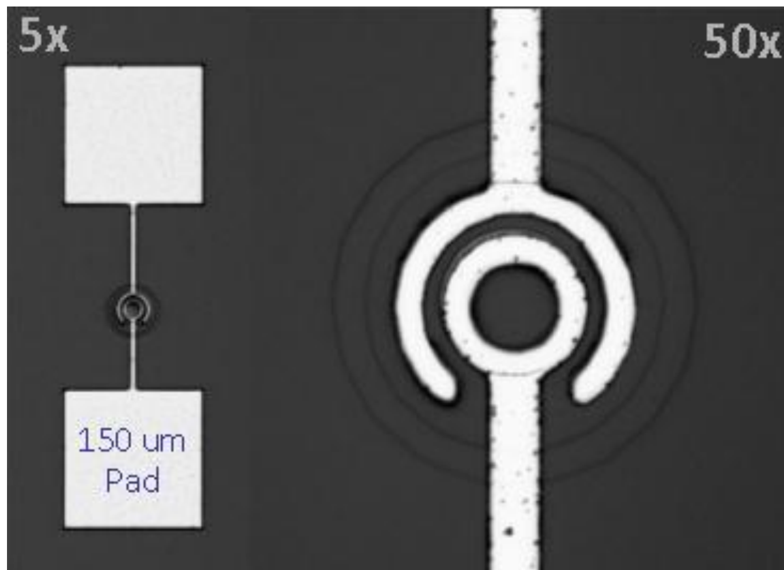


Figure 2.51. Micrographs of 20 µm guard ring diode and top side anode and cathode contacts and probe pads. The outer non-contacted ring is an n+ diffusion for local gettering.

The majority of guard ring structures across the 16 wafers previously mentioned exhibit especially low reverse bias leakage currents (Figure 2.52) congruent with those from the single-diffusion diodes used to illustrate gettering efficiency, discussed in Section 2.2.C.

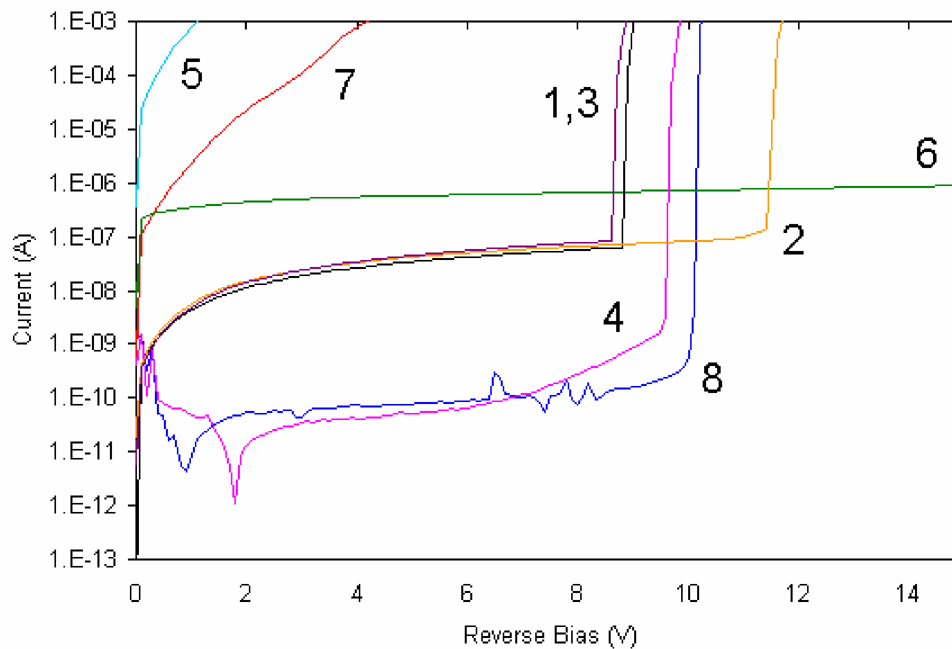


Figure 2.52. Characteristic curves for 20 µm laterally-contacted guard ring diodes on 8 different n-type wafers.

As in the section on gettering, wafers #4 and #8 exhibit the lowest reverse bias current, while wafers #1, #2, and #3 each exhibit the next highest leakage current. Wafers #5 and #6 are each missing n-well diffusions which contributes to their higher leakage currents. The deeper p-well diffusion of wafer #2 clearly increases the breakdown voltage. The remaining features are perhaps more difficult to explain without a complete understanding of where exactly the diodes are breaking down.

Incidentally, one concern with furnace doping is the potential for nonuniform doping across the wafer surface. The four-diffusion process detailed above should represent a fairly robust trial for across-wafer doping uniformity. The vertical (Figure 2.53) and horizontal (Figure 2.54) IV characteristic variations are detailed below.

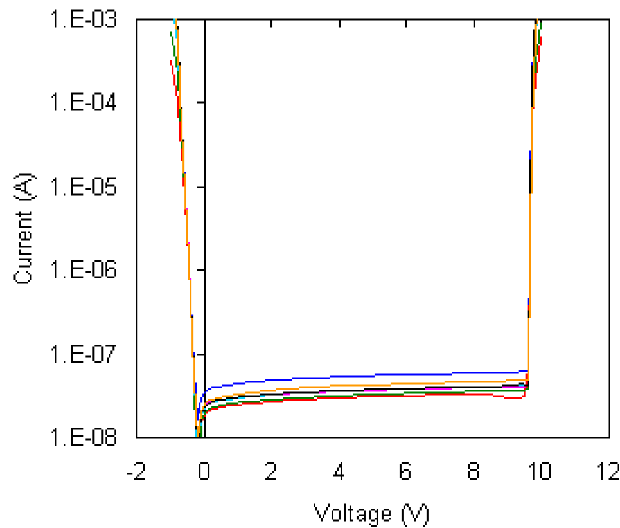


Figure 2.53. Characteristic curves for 20 µm laterally-contacted guard ring diodes from wafer #4 and 7 die distributed *vertically* across the wafer.

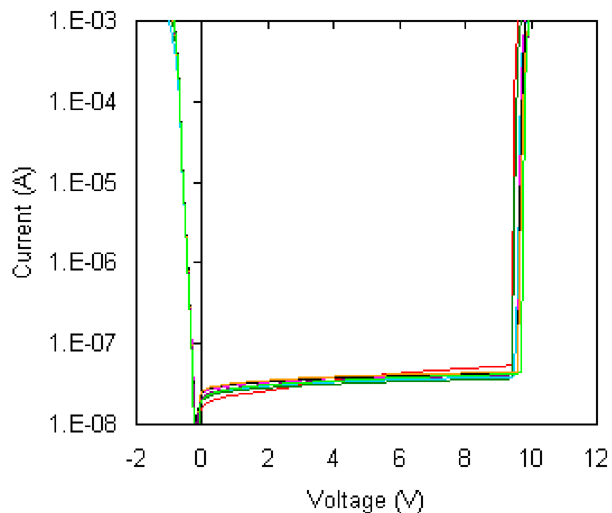


Figure 2.54. Characteristic curves for 20 µm laterally-contacted guard ring diodes from wafer #4 and 7 die distributed *horizontally* across the wafer.

The 20 µm diodes from wafer #4 for instance exhibited extremely stable behavior from die to die as evidenced in Figure 2.53 and Figure 2.54. The reverse bias leakage current varies by a factor of two or three, while the breakdown voltages are within a range of 0.5 V.

2.5.B. Breakdown Mapping

When contacted laterally, the two top side anode and cathode contacts are used, and the bottom substrate contact is left floating. Because the n-well in this series of diodes was diffused as the very first step, during the backside gettering process, it must necessarily be the deepest well. That is, any additional furnace diffusion steps will only drive the n-well deeper, and no other diffusions can hope to overtake its depth. Considering the structure above (Figure 2.50), this means that the n-well guard ring punches through the p-well tub/tank. The current from the p-well contact must then flow across the p-well/n- diode. While this is problematic, and correctible with proper choice of diffusion order, a more serious problem exists.

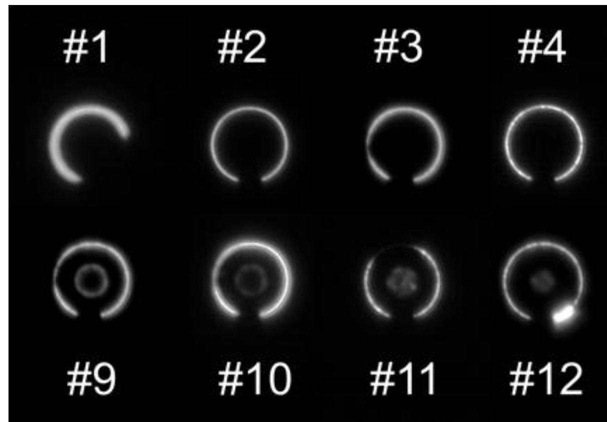


Figure 2.55. Electroluminescence images from *laterally* contacted 20 μm guard ring diodes. Wafer numbers indicated.

The EL images in Figure 2.55 indicate primary breakdown in a ring between the guard ring and the p-well tub contact. This is a result of the unreasonably small distance between these two features. During layout, the diffusions were planned to be less deep, and the diode was space-optimized. With the later choice of deeper doping, optimizing other structures, the guard ring and the p+ contacts came in very close contact and became the primary location for breakdown. This should also serve to illustrate the many 3D "diode" locations possible when attempting to create a single diode.

Several other features serve to illustrate specific fabrication concerns. The incomplete ring from the wafer #1 diode is confirmed to be the result of poor alignment, and so the diode breaks down preferentially towards the upper-left. If we focus instead on the lower-right corner which does not break down at this voltage, then we might conclude that an alteration of the contact spacing could prevent this ring breakdown altogether.

Another feature of interest lies in the difference between diodes from high-resistivity n-type (#1-4) and low-resistivity p-type (#9-12) wafers. The central region inside of the peripheral glow is observed to break down in the case of a low-resistivity p-type substrate. This would suggest that the n-type wafer diodes are experiencing a much greater blocking effect in their central diode region. The more highly doped p-type wafers decrease the central region breakdown voltage and are thus observed simultaneously with the peripheral glow. The inner diode

glow from wafers #9-12 also appears to be less uniform which might indicate the inability of the gettering layer to completely remove impurities that locally alter the electric field. This is the motivation driving the use of high-resistivity substrates. Looking at the inner glow from the wafer #9-12 diodes, a much more uniform glow is apparent for those diodes from wafers #11 and #12; their unique feature being a deeper n-well or n+ diffusion. Clearly, the lateral contact scheme has not been optimized for these doping conditions, and vice versa.

Figure 2.56. Electroluminescence images from *vertically* contacted 20 µm guard ring diodes. Wafer numbers indicated. Magnification is slightly larger than Figure 2.55.

However, when these diodes are contacted vertically, as in Figure 2.56, they are seen to break down in their central regions alone, as desired. The spatial variation in electroluminescence light is extremely smooth, and tapers off quickly and smoothly at the edges. If we were to judge an avalanche diode based on this criteria alone, there could be little improvement to the current design. This appropriate break down comes at a price however, since the current must now flow past the tank-substrate diode and through the entire wafer. This adds both resistance and leakage current to the vertically contacted diodes as depicted in Figure 2.57.

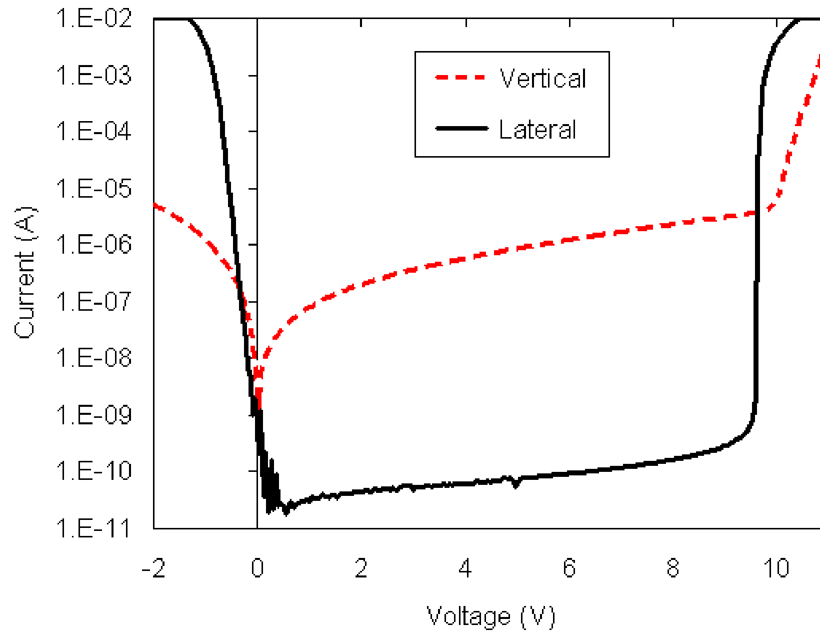


Figure 2.57. Characteristic curve for vertically and laterally contacted 20 µm guard ring diode from wafer #4.

The current passing through the forward biased diode (negative voltages) is extremely impeded by the added resistance. The straight line beyond 10 V indicates an exponentially increasing current above breakdown and is indicative of avalanche multiplication. However, Geiger mode behavior was not observed, presumably due to the relatively large density of free carriers causing the high leakage current.

For further understanding, we turn our attention back to a more realistic simulation, using the doping values derived from wafer #4 test structures. The simulated doping from an edge segment of the diode is depicted in Figure 2.58.

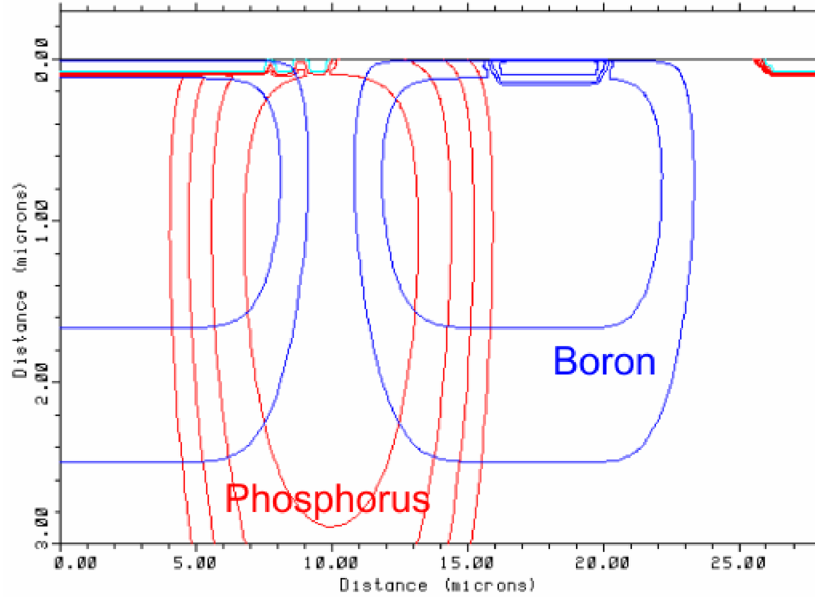


Figure 2.58. Doping contours at edge of guard ring diode. Blue is boron (p-type), and red is phosphorus (n-type).

Note the splitting of the p-well (at 10 μm) and the close proximity of the p+ contact to the n-well (at 15 μm). The active n+/p-well region extends from 0 to ~5 μm. A lone n+ gettering site is found beyond 25 μm. The simulated electric field at 10 V applied bias for both lateral and vertical contacts is illustrated in Figure 2.59.

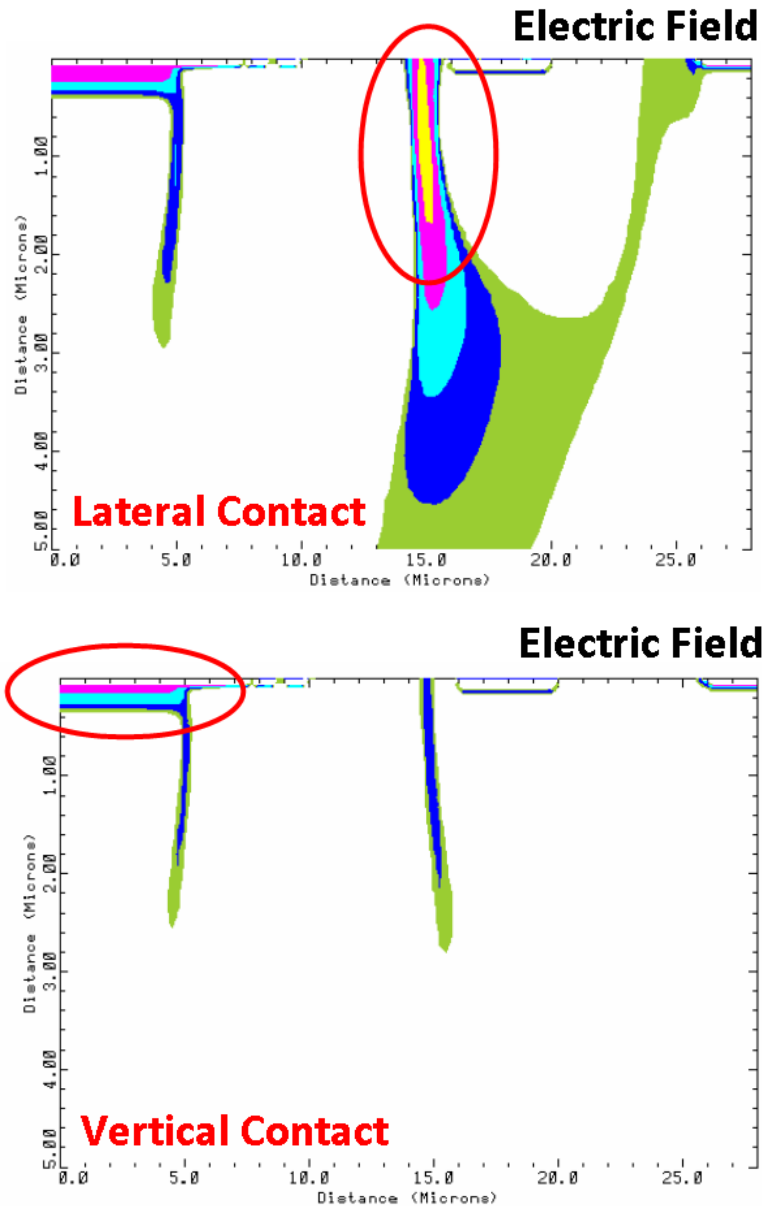


Figure 2.59. Simulated electric field at 10 V reverse bias for both lateral and vertical contacts. Circled are the regions of greatest electric field.

Observe that the electric field is greatest at the periphery of the guard ring in the laterally contacted case. When a vertical contact is employed, the region of highest electric field is confined to within the intended active area at the center of the diode. This accords well with the EL images from both vertically and laterally contacted diodes. If we are to trust the simulation, then it would appear that the laterally contacted diode would break down at a lower bias since its electric field

is higher at 10 V bias than in the vertical case. Referring to the measured IV curves in Figure 2.57, a slight difference in breakdown voltage is indeed observed.

2.5.C. Light Response

A preliminary response to incoming photons can be obtained by applying an arbitrarily increasing photon flux injected from a probe station microscope light onto a guard ring diode. The effect of the additional photoelectrons can be witnessed in Figure 2.60.

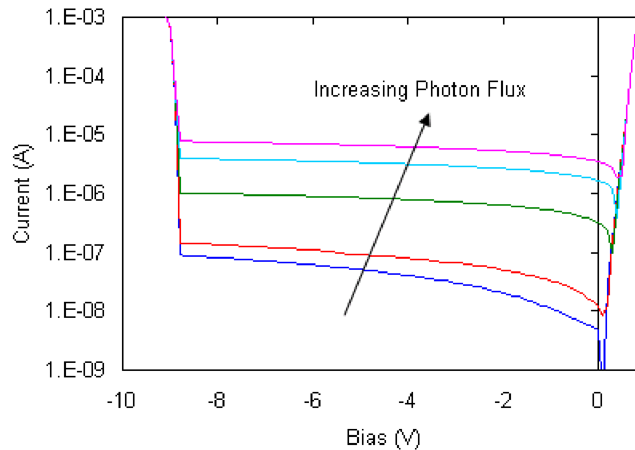


Figure 2.60. Current-voltage characteristic curve for guard ring diode from wafer #3. An increasing photon flux is observed to result in an increased reverse bias leakage current.

Starting with an absence of visible-wavelength photons applied to the diode, the reverse bias current is seen to increase by several orders of magnitude. This indicates that the diode is functioning as a standard PIN-like photodiode under reverse bias. One particular feature worthy of note is the continuation of the reverse bias current from negative to positive voltages. One might normally expect the current to always be zero when the applied bias is zero.

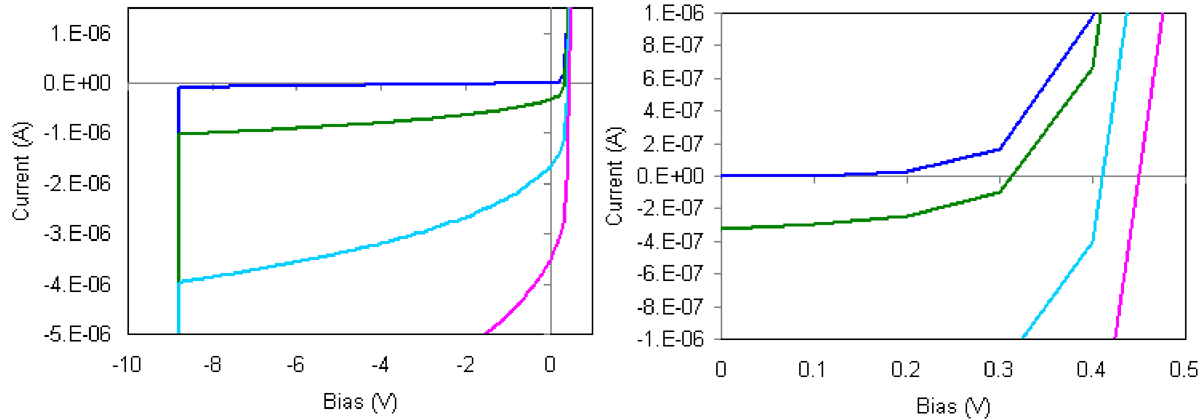


Figure 2.61. Current-voltage characteristic from Figure 2.60, only on a linear non-absolute-value scale (left) and magnified (right).

Taking a closer look at this phenomenon in Figure 2.61, we see again that the zero-voltage current increases negatively with increasing optical stimulation. The cause for this added electron motion is the built-in electric field which sweeps the additional photoelectrons in the depletion region. There is a simultaneous *photovoltaic* effect as witnessed by the shift in the zero-current voltage as the optical stimulus is increased. Because it was previously established that the high field region of this particular diode was confined to a small ring, the response to a very low photon flux cannot be expected to be as large as that from a properly designed diode. That is, the photon-sensitive region is limited to a smaller area. The absolute photon responsivity (in measured current for an incident photon power) is thus expected to be very poor for this diode.

2.5.D. Geiger Mode Behavior

Even if the probability of detection is rendered nonuniform by the nonuniform electric field, the question remains whether these diodes might still operate as Geiger-mode APDs. By adding an external quench resistance of 20 kΩ and observing the transient voltage (Figure 2.62), we can begin to answer that question.

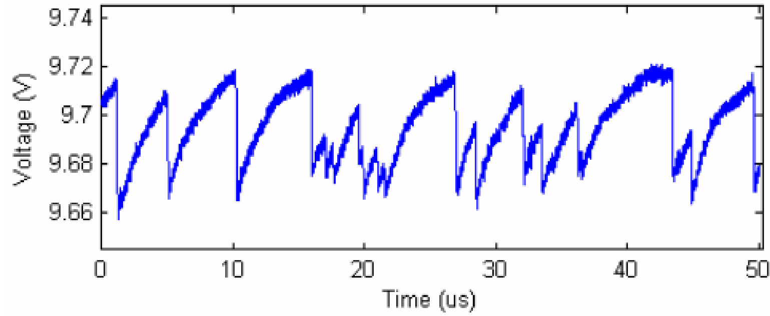


Figure 2.62. Transient voltage from a 20 µm (laterally contacted) guard ring diode (wafer #4) with an external 20 kΩ quench resistor in series, as measured by an oscilloscope.

Although effectively a ring diode, Geiger mode is observed with individual and quenched avalanche pulses. However at just 0.2 V above breakdown, a significant dark count rate of ~300 kcps is present. The external cabling capacitance dominates the recharge time constant, so increasing the excess bias leads to an improper estimate of the dark count rate trend. If the stray capacitance was reduced, then the recharge time would be shorter, and a greater number of dark events would be measurable.

A similar diode (8.6 µm diameter), implemented in a 130 nm process with a breakdown voltage of 9.4 V, experienced a dark count rate of 120 kcps at 1 V excess bias [Ger09]. Significant dark count rate improvement (to 30 *cps*) was obtained by reducing the implantation dose and achieving a higher breakdown voltage (12.8 V). This improvement was a direct result of the reduction of band-to-band tunneling which exists at lower breakdown voltages. Another diode (10 µm diameter), also implemented in a commercial 130 nm process, exhibited a room temperature dark count rate of 12 kcps at 1 V above a breakdown voltage of 10 V [Nic07]. Considering these two published results, there is reason to believe that improvement of the doping conditions to obtain uniform diodes might also yield diodes with similar dark count rates.

2.5.E. Quench Resistor Selection

Even with a diode which breaks down on the periphery of the guard ring, some avalanche signal features can still be analyzed. One remaining design question is the choice of a quench resistance by observing the dependence of the

Geiger mode signal on the value of the quench resistor. Take for example a guard ring diode (wafer #2) with the current-voltage characteristic depicted in Figure 2.63. This n+/p-well diode has a deeper and more lowly-doped p-well than that from wafer #4 (see Section 2.2.A), and therefore breaks down at a slightly higher voltage, 11.25 V. Note the reduction in both forward-bias and above-breakdown current conduction with the addition of an external 20 kΩ series quench resistor from a shielded resistor decade box (Extech 380400).

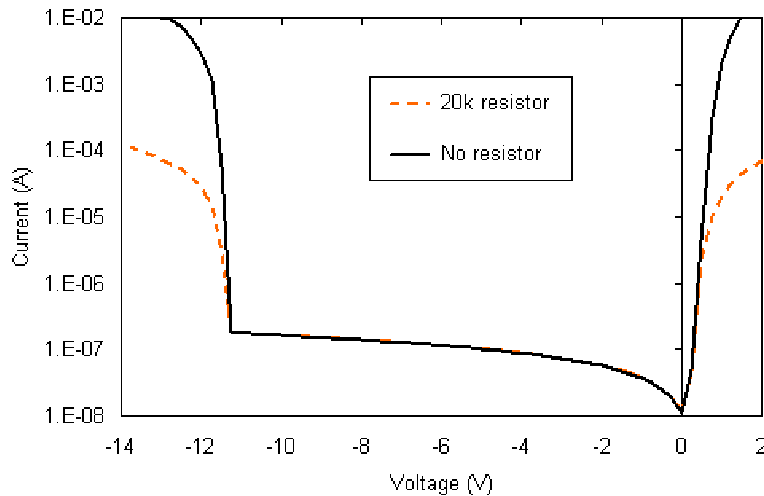


Figure 2.63. Current voltage characteristic for guard ring diode from wafer #2, with and without a 20 kΩ external series quench resistor.

By operating this diode at 11.45 V, 200 mV above breakdown, the avalanche signal can be observed (Figure 2.64) for varying values of externally added series resistance. The voltage signal is extracted by measuring the potential across the quench resistor with a 10 MΩ oscilloscope probe. A not insignificant "electronic" noise component is unavoidable because of the coaxial cabling involved and the unshielded probe station. This can be seen in the non-stationary baseline and the high-frequency or ringing component.

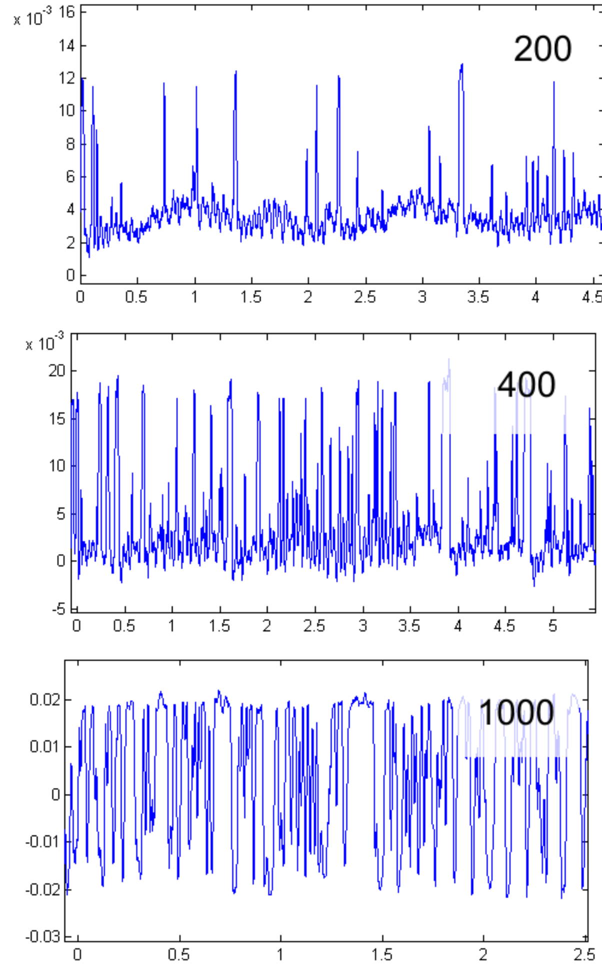


Figure 2.64. Progression of Geiger mode voltage signal from constant avalanche (200 Ω, top) towards a random telegraph signal (1000 Ω, bottom) for varying values of externally added resistance. Excess bias V_e = 0.2 V. Time scale is 1×10^{-4} s (third plot is time-expanded).

At first glance, it might appear that individual avalanche pulses are clearly observed in the 200 Ω quench resistor case. However, the opposite in fact is true. There is a sustained avalanche current with zero quench resistance, as in the majority of current-voltage characteristics presented in this work. The random probability that the avalanche will spontaneously quench begins to increase as the quench resistance is raised. There begins to be a slight probability that the avalanche current will momentarily quench even with a meager 200 Ω, only to be instantly reignited. The signal experiences near equivalent turn-on and turn-off probabilities at 1000 Ω and appears as a random telegraph signal. As the resistance is increased, there is an increasing probability for a zero-current state. When integrated by an IV curve tracer or semiconductor characterization system,

these increasingly frequent regions of no current yield a lower reverse bias current above breakdown, as in Figure 2.63. In fact, if the curve could be traced quickly enough (on the order of the dark count inter-arrival time), the breakdown rush of current might appear at any voltage above the breakdown.

It should be mentioned that the absolute values of these resistances are not necessarily directly transferable to other diodes, since the signal formed is largely dependent on the actual capacitance presented by the reverse biased diode *and* any external parasitic cabling capacitance. In fact, these resistances are also not even directly transferable to the values one might select for integrated thin-film quench resistors for this very diode without extraction of any parasitic effects.

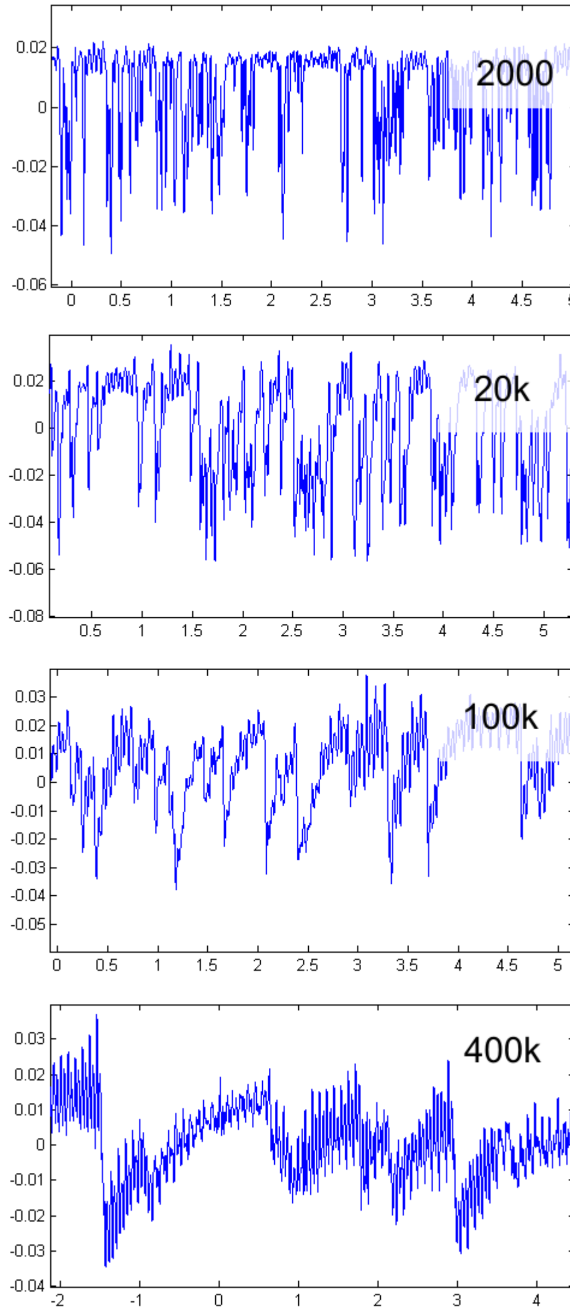


Figure 2.65. Progression of Geiger-mode avalanche signals with increasing quench resistance (indicated in Ohms) at V_e = 0.2 V. Time scale is $1{\times}10^{-4}$ s.

As the quench resistance is increased (Figure 2.65), the random telegraph signal begins to transition into a regime of increasing quiescence. The increasing resistance sets an increasing current threshold above which any observable avalanche signal must rise. The effect of the resistance on the diode recharge time constant is evident. In terms of design goals, a small quench resistance is

desirable so that the dead time is decreased. Reduction of diode size and all capacitances involved also will help to achieve a faster and less noisy diode. However, the quench resistance must not be so small that the diode does not have a stable quiescent state. Quench resistors in commercial devices can range from kΩ to MΩ, because of the wide array of diode designs and fabrication techniques. One method for selecting a quench resistor might be to measure the true capacitance of the diode alone and then select a resistance from a set of known Geiger mode behaviors.

2.5.F. Multiple Breakdown

The three-dimensional interplay of dopants often intrinsically creates multiple junctions in addition to the desired diode. Premature breakdown may occur if these additional diodes are not properly taken into account. Poorly aligned contacts can also cause premature breakdown if aluminum layers are allowed to directly touch silicon during the final anneal. Figure 2.66 presents a guard ring diode from wafer #12 with multiple breakdown sites. The first breakdown occurs at 6 V and can be seen as the brightest spot in the lower right-hand corner. Its most likely causes are poor contact alignment and aggressive design tolerances.

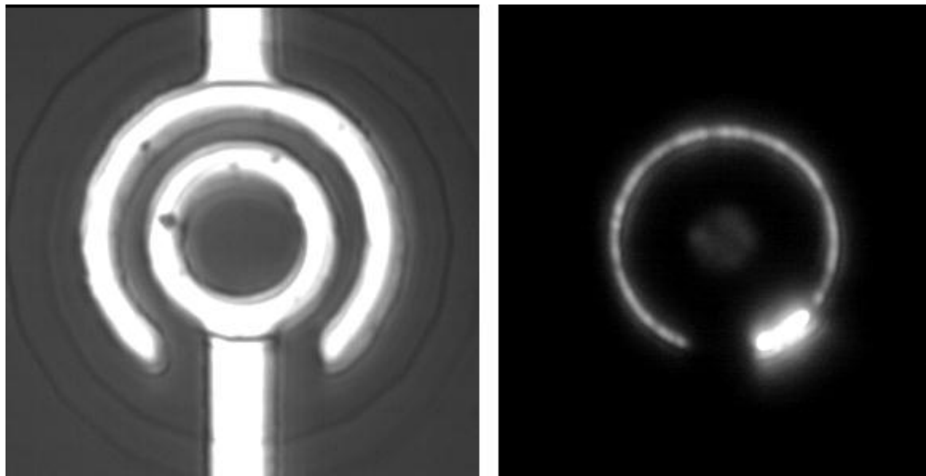


Figure 2.66. Micrograph (left) of a 20 µm guard ring diode (wafer #12) and steady-state electroluminescence of laterally contacted diode at -15 V and 10 mA. Multiple breakdown sites are evident: at the center, in a ring, and at the tip of that ring.

An additional breakdown occurs at 12 V, corresponding to the luminescent ring characteristic of other ~12 V diodes. The final breakdown occurs just above

this voltage and is not visible in the IV curve (Figure 2.67) but manifests itself as the central faint luminescent circle in the desired active area of the diode.

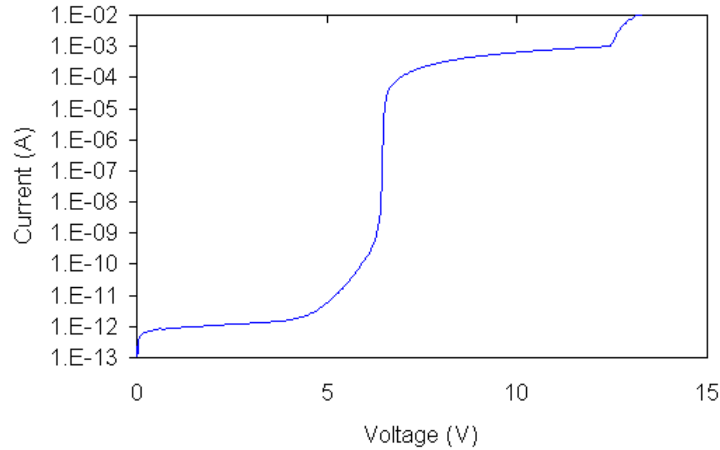


Figure 2.67. Characteristic curve for diode depicted in Figure 2.66. Note the multiple breakdown voltages at 6 V and 12 V.

Although this diode does eventually break down in its central area, the previous break downs will completely drown out any single photoelectrons arriving in that central region. If the premature breakdown sites could be completely eliminated, then this diode would breakdown only at its center somewhere between 12 V and 15 V.

Now consider a doubly-diffused diode where one diffusion is poorly aligned to the other. Unless adequate design rules are created and followed, asymmetrical alignment can lead to stronger electric fields on one side of the diode (Figure 2.68). In this case, the edge breakdown occurs at such a low voltage that any other breakdown sites are completely obscured.

Figure 2.68. Electroluminescence image of poorly aligned diode indicating premature edge breakdown.

Multiple breakdown sites are not strictly undesirable. The principle requirement is simply for the first breakdown to be that of the designated active area. Because single-photon avalanche photodiodes are typically operated at several volts above breakdown, any additional intrinsic diodes must be designed to break down above this operating point.

2.5.G. Summary

The guard ring diodes presented in this section were designed with a tank diffusion to eliminate leakage from the substrate in much the same way as an epitaxial layer does for CMOS devices. A central diffusion of opposite type defines the outer edge of the diode, while a ring of lesser doping concentration serves to reduce the electric field at the edges. Because of the complex temperature budget and the four diffusions required for this type of diode, not all diffusion parameters were optimized for these designs. This was evidenced by the poor edge breakdown characteristics indicated by the electroluminescence images of laterally contacted diodes. However, vertically contacted diodes displayed extremely smooth and centrally contained breakdown. The benefits gained from this contact scheme unfortunately were overshadowed by the added resistance and leakage current from current passing through the entire wafer.

Even with a peripherally active diode, a photodiode response was observed at reverse biases below the breakdown voltage. Geiger mode behavior was also observed for the peripheral diode, and the illustration of avalanche quiescence

was presented as a function of the quench resistance. An absolute determination of the requisite quench resistance was complicated by the external capacitance from the probe station, cables, and resistor selection box.

2.6. Transparent Quench Resistors

Besides the avalanche diode itself, one of the only additional integrated components in a solid-state photomultiplier is the quench resistor. There are two fundamental methods for creating resistors in commercial integrated circuits. Resistors can either be formed in bulk silicon by controlled doping, or they can be formed from thin-films and placed above the passivation layer during back end of the line processing. Diffused resistor values fall in the range of 5 to 5000 Ω/square, however they consume potentially photoactive silicon real estate and are thus not considered for this application. It is worth noting that one group is achieving success with a buried quench resistance structure below the diode, although this approach currently requires cooling due to bulk injection of thermally generated carriers [Nin10]. Patterned polysilicon is the primary material used for interlayer thin-film resistors in commercial CMOS processes. High resistivity polysilicon values are limited to approximately 2000 Ω/square and can have a temperature coefficient of resistance (TCR) of approximately -1000 ppm/°C. A common polysilicon thickness is 200 nm, and the optical properties of polysilicon are reasonably close to those of monocrystalline silicon [Lag97].

The thin-film quench resistor may be required to cover some portion of the diode active area for very closely packed SPADs in a solid-state photomultiplier. This is especially true if longer resistors are needed due to upper limits on thin-film resistor sheet resistivity. Traditional polysilicon resistors will decrease the probability of detection, especially for photons of shorter wavelength. For this reason, we wish to implement a transparent conductor as a replacement for the polysilicon quench resistor. Several candidate materials are available including Poly(3,4-ethylenedioxythiophene) (PEDOT), indium tin oxide (ITO), Al-doped Zn oxide (AZO) and with significant development, carbon nanotubes. ITO was initially selected as a material of interest because of its process maturity [Bas98].

The resistivity of stoichiometric ITO is ~1×10^{-4} Ω-cm depending on deposition and annealing conditions. Its TCR is ~200 ppm/°C which would represent an improvement over polysilicon resistors. In this work, a In_2O_3:SnO_2(10%) RF sputtering target is used to deposit ITO at room temperature. After hotplate annealing in a nitrogen ambient at 500 °C for 10 minutes, a film of 4×10^{-4} Ω-cm is achieved with refractive index of 2.0 throughout the visible spectrum with a negligible extinction coefficient down to 400 nm. This provides a sheet resistivity of 20 Ω/□ for standard film thicknesses of ~200 nm. By decreasing the thickness to the minimum controllable limit of ~10 nm, a higher sheet resistivity of 400 Ω/□ can be obtained. When considering a ~50 kΩ quench resistor, 125 squares of thin ITO would be required. Provided that the minimum realistic patterned line *width* of ITO is 3 μm, a 375 μm long resistor would be necessary. This length is clearly unacceptable for the tens-of-microns scale of most SPADs. Therefore, an alternative material must be sought.

We first consider that the conductivity of ITO cannot compare with other metals because of the oxygen present. One may then naïvely infer that an even lower conductivity might result from an increase in the oxygen content either during or after deposition. Because a plasma sputtering source is being used to deposit the ITO, wherein the In_2O_3 and SnO_2 species are inherently dissociated, there exists a convenient environment in which to introduce additional oxygen. Sputtering is typically performed in the inert gas argon at pressures of 7 mTorr with base vacuum pressures of <10 μTorr. In the process herein, oxygen (8-10%) was introduced in order to *increase* the resistivity of the deposited films. Similar work exists which attempts to minimize ITO resistivity by converging on an ideal O_2 gas ratio [Wu 1996].

Patterning of thicker ITO films (>200 nm) is accomplished relatively easily by lift off patterning (Figure 2.69) using a 3 μm thick Shipley SPR 220-3.0 resist. Immediately after the spin and softbake, the photoresist receives a 20 s spray developer treatment to harden the top layer. This process provides slightly overhanging edges which facilitates lift off. Sputter deposition is followed by extended soaking (hours to days) in a photoresist solvent like acetone, PRS-2000.

Photoresist strippers like PRS-2000 or 1116A are preferred because of their additional surfactants which reduce readhesion of lifted films. Heated solutions allow for faster lift-off times, but can etch certain metals at elevated temperatures. It should be noted that while many sputter tools provide conformal coatings, covering photoresist sidewalls, the specific tool used in this work provides a more unidirectional source.

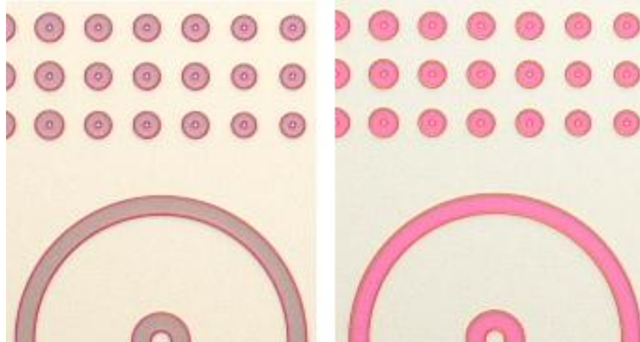


Figure 2.69. Sputtered ITO (2000 Å) on Si, patterned by lift off (left) and 1:7 HCl etch (right). Smallest openings are 5 µm in diameter.

Patterning of thinner ITO films (~10 nm) by lift off is far less successful due to the high degree of readhesion in this rather gross and uncontrolled process. A wet etchant offers a more controlled process solution.

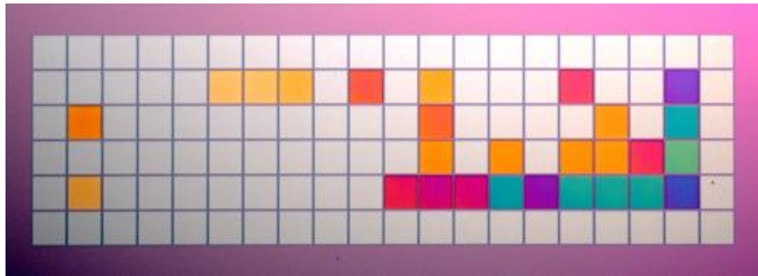


Figure 2.70. Micrograph illustrating *incomplete* HCl etch of ITO film due to high surface tension. Colors indicate ITO thickness, similar to traditional colors for silicon dioxide thickness.

A dilute HCl acid etch (~1:10 H_2O:HCl) has been observed to provide extremely reliable results. A 30 s, 80 Watt, O_2 plasma descum has been found to be beneficial in removing any photoresist monolayers which would increase surface tension and impede the etch, as in Figure 2.70.

2.6.A. ITO Resistor Results

In order to successfully integrate thin oxygen-rich *resistive* ITO (RITO) quench resistors and thicker *conductive* ITO (GITO) readout lines, ohmic contact must be established between the diode, the resistor, and the readout pad. Early investigation indicated that the oxygen in either ITO film reacts at higher temperatures with both silicon and aluminum to form SiO_2 and Al_2O_3. These insulators effectively block all current flow. Given the need for post-deposition ITO annealing, some intermediary material must be sought which acts as an effective diffusion barrier at higher temperatures and provides ohmic contact across the entire stack of thin-films.

An experimental design was created in which the interaction and conductivity between five materials could be simultaneously tested. The principle test structure utilized is the transfer length method structure (Figure 2.71) [Ous05], sometimes combined or confused with the "transmission line model" [Ber72].

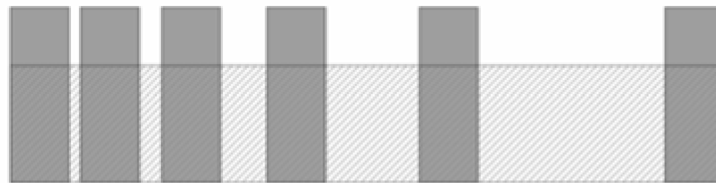


Figure 2.71. Transfer length method structure with 6 metal pads contacting a single strip of resistive material. The 5 resistors are each contacted by equivalent area contacts.

The concept is simply to create a series of resistors with increasing lengths and equivalent contact areas. The zero-length resistance is extrapolated and is equivalent to twice the contact resistance (and any probe resistance). The *specific contact resistance*, in Ohms per unit area, can then be extracted by considering the metal-film contact area. As an additional benefit, the sheet resistance, in Ohms per square, can be extracted from the slope of the resistance vs. length plot. This structure is repeated for five materials in a combinatorial experimental design (Figure 2.72) utilizing one mask set for many experimental process variations.

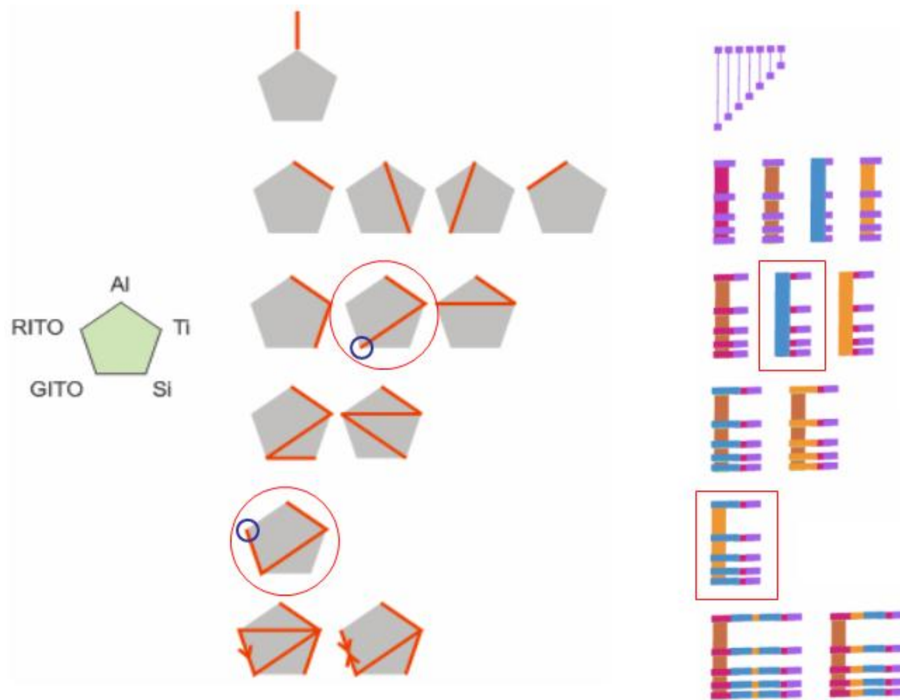


Figure 2.72. Combinatorial experimental design schematic (left) for the extraction of contact and sheet resistances for multiple materials (Al, Ti, Si, RITO, and GITO). Actual mask design shown at right.

The first row at the right of Figure 2.72 is a series of Al resistors of increasing length meant to test the sheet resistance of the aluminum as well as the contact resistance of the tungsten probe tips to the aluminum pads found in the remaining structures. The second row allows extraction of the sheet resistance of the remaining four materials, provided they make ohmic contact with aluminum. If ohmic contact is made, then the specific contact resistance between aluminum and that material can also be extracted and used in later tests. The second feature in the very last row illustrates silicon resistors contacted by Ti:GITO:RITO:GITO:Ti:Al:(probe). In this way, annealing variations and material substitutions can be made while pinpointing the exact location of any non-ohmic behavior. Many iterations were attempted with various material deposition parameters and annealing conditions and sequences. The majority of tests resulted in very high contact resistance or severely non-ohmic behavior and are therefore not reported here. The final sheet resistances of 200 nm GITO and 10 nm RITO were 21 Ω/□ and 12.4 kΩ/□. The absolute resistivities of the films, independent of thickness, were 4 μΩ-m for GITO and 100 μΩ-m for RITO,

illustrating a factor of 25 increase in resistivity by sputtering in an Ar:O_2(8%) RF plasma. Resistors of reasonable complexity (Figure 2.73) have been successfully fabricated.

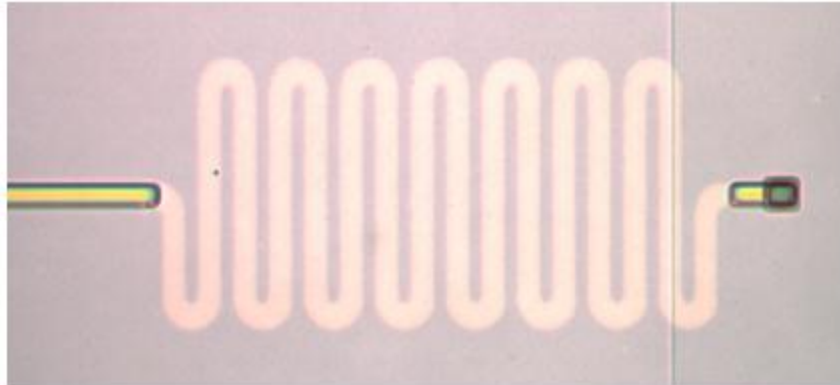


Figure 2.73. Serpentine RITO resistor some ~145 squares long and 5 µm in width. Measured resistivity is 1.9 MΩ.

Sputtered titanium and titanium-doped tungsten each performed well as refractory barrier layers for creation of ohmic contacts to both silicon and aluminum. The very thin RITO layer was deposited before the GITO layer, due to morphological concerns and the need for chemical etching. That is, the 100 Å RITO was not expected to conform to a 2000 Å step without experiencing discontinuities. An interdielectric layer would facilitate the chemical etching of both ITO layers but would increase process complexity. The final metal stack and micrograph appear in Figure 2.74.

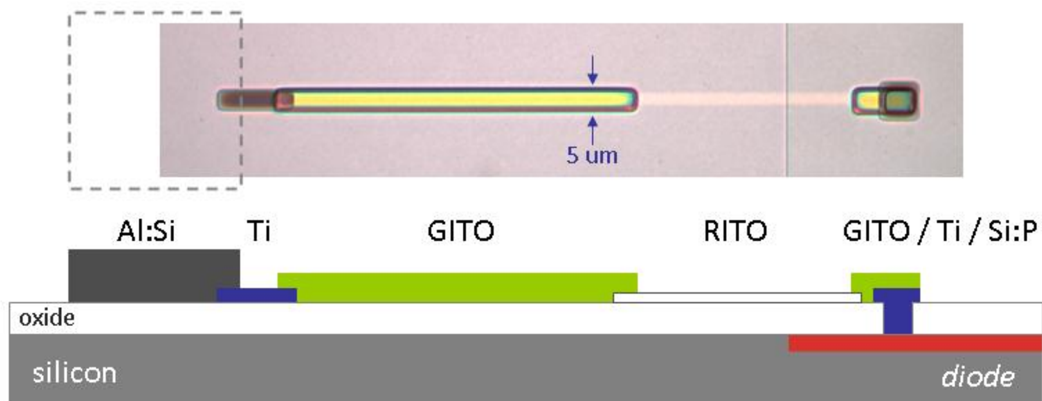


Figure 2.74. Micrograph and cross section illustrating the method for ohmically contacting ITO and oxygen-rich ITO to both silicon and aluminum.

GITO was the only material identified which provided ohmic contact to RITO. It is expected that during annealing, any surplus oxygen present in the RITO will safely diffuse across the GITO boundary providing a gradual material transition. The transparency of both RITO and GITO was measured with a white light reflectometer, to be 80-95% transmissive down to 350 nm.

One additional concern when using very thin-films is their current handling capacity. Because of the avalanches of charge flowing through the resistor, any heating can potentially be detrimental. However, the post-deposition annealing temperature of the ITO films is 500 °C, so any self-heating effects would have to heat the ITO well above this fabrication temperature.

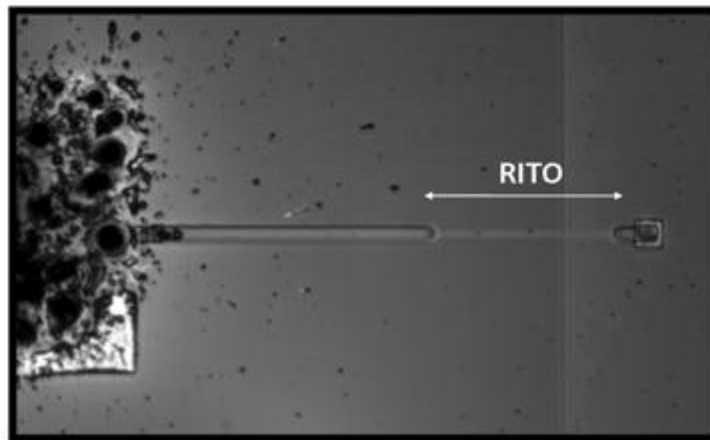


Figure 2.75. Current carrying capacity of RITO film stack. Evaporation occurs at the Ti:Al interface at the far left while the GITO, RITO, and Ti-via to the PN junction at the right remain intact.

Testing the current capacity of the GITO/RITO film stack indicated that the weakest point is in fact the Ti:Al boundary (Figure 2.75).

In summary, a process for the deposition, patterning, and ohmic contact of highly resistive oxygen-rich ITO films has been successfully developed for the purpose of integrating high resistivity transparent thin-film quench resistors onto Geiger-mode silicon avalanche photodiodes.

Chapter 3

Stochastic Model of the SiPM

The motivation for the development of a stochastic model of the SiPM is the improvement in the estimation of the energy, position, and timing of incoming radiation. Detectors convert energy from physical phenomena into data. With an accurate knowledge of that conversion process, the data can be processed into meaningful information. The conversion process and the method of information extraction can each degrade the utility of the final result. With an appropriate physical model of the detection process, collected data can be mapped onto an estimate of the underlying stimulus. This estimation process requires some sort of inversion of the physical model. In simple cases, this inversion will provide a uniquely defined one-to-one relationship between datum and stimulus. In many real detection schemes however the detector will provide identical data for multiple stimuli, representing an ill-defined inversion and a questionable estimate of extracted information.

As with many technical innovations, successful application often precedes the development of a complete physical model. Silicon photomultipliers can for instance be applied to gamma-ray spectroscopy [Pav05] and suitable results may be obtained by simply fitting Gaussian distributions to full-energy photopeaks. The trouble arises when energy, timing, or spatial resolutions are greater than might be expected or desired from traditional technologies (e.g. the PMT). Then an accurate physical model is absolutely necessary to determine the ultimate limits of the device in-hand or of the technology in general. In addition to providing appropriate technology selection criteria, physical models allow for correction of known nonlinearities and systematic errors [Joh09, Bur07]. The

statistical derivations developed in this chapter augment current models by offering a framework with which to derive expected energy resolutions.

First, we derive from first principles the probability that a single pixel will fire. Then the mean and variance in a collection of pixels is determined. It is recognized that the discretized nature of the SiPM leads to a nonlinear response which can be treated as an estimation problem using the methods of information theory. Having obtained an estimate of the incident photon flux from the number of pixels which fire, several methods are proposed to predict the variance in that estimate which leads to the intrinsic energy resolution.

Several other additional phenomena are compounded with this inherently nonlinear response. The reality that each avalanche contains a slightly different amount of charge must be considered to fully account for potential spectral broadening. Thermally generated events, afterpulsing, and optical crosstalk are each treated separately with the acknowledgment that their effects are often interdependent. The finite recovery time of each pixel, if fast enough, may negate the inherent SiPM nonlinearity. Finally, an effort is made to point toward the components required for the development of a complete model for a given device.

3.1. Scintillation Detection

The resolution of any scintillation detection system arises from a variety of physical mechanisms. Assuming a monoenergetic source of radiation, there will be a variable amount of energy lost in the scintillator and a variable amount of light produced. As higher energy radiation is transferred to electrons, X-rays are produced whose energy is transferred to additional electrons. This cascade process, among other factors, leads a scintillator to respond nonlinearly in its light output. Higher energy radiation may Compton scatter and not be entirely collected within the limited volume of the crystal. This can lead to position-dependent energy depositions.

Scintillation photons are isotropically emitted and may scatter or be absorbed by the bulk of the scintillator, depending on their wavelength. Photons reaching non-detecting boundaries will either be absorbed or scattered, with a directional

and wavelength dependence. Photons reaching a detecting boundary must first cross a barrier of differing refractive indices. Several optical interface layers may also impede this transfer efficiency. Once a photon arrives at the surface of a photodetector, many additional loss mechanisms and noise processes may contribute to the degradation of the energy resolution in the final electrical signal. The purpose of this chapter is to illustrate the complexity of those mechanisms for the silicon photomultiplier.

3.2. Single Pixel Firing Probability

In contrast with the spatially continuous photocathode of the PMT, a silicon photomultiplier is fundamentally an array of independent discrete devices. Each pixel, or single photon avalanche photodiode (SPAD), operates to produce a large avalanche of charge in response to a single electron. This *electron detector* is applied in the context of radiation detectors to detect single photoelectrons from incident scintillation photons. In order to evaluate the consequences of this photosensor discretization on energy resolution, we begin with the fundamental stochastic processes involved. First, we let some random number N photons be incident upon an M-pixel device as in Figure 3.1.

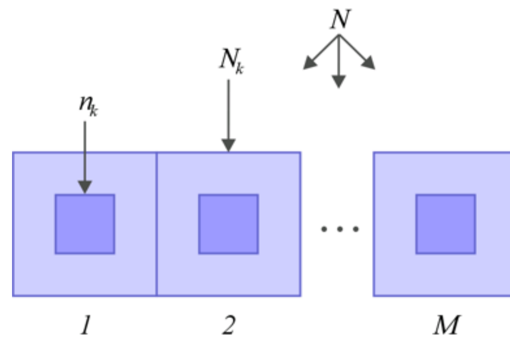


Figure 3.1. Diagram of random variables describing photons incident on a discretized detector.

Let the random variables

$$
\begin{aligned}
N &\equiv \text{number of photons incident on entire device} \\
N_k &\equiv \text{number of photons incident on pixel } k \\
n_k &\equiv \text{number of photons incident on sensitive area of pixel } k \\
\alpha_k &\equiv \text{number of photoelectrons capable of causing an avalanche in pixel } k
\end{aligned}
$$

be defined by the following (conditional) probability distributions

$$\begin{aligned} P(N) &= \text{Poisson}(\lambda) \\ P(N_k \mid N) &= \text{Binomial}(\text{N},1/\text{M}) \\ P(n_k \mid N_k) &= \text{Binomial}(\text{N}_\text{k},\varepsilon) \\ P(\alpha_k \mid n_k) &= \text{Binomial}(n_k, p) \end{aligned},$$

where

$$\begin{aligned} \lambda &\equiv \text{average number of incident photons} \\ M &\equiv \text{number of pixels} \\ \varepsilon &\equiv \text{geometric efficiency (fill factor)} \\ p &\equiv \text{avalanche initiation probability} \end{aligned}.$$

Initially, a random number of photons N are simultaneously incident on a SiPM with M pixels. The conditional probability $P(N_k \mid N)$, the probability of N_k photons given N photons, represents a uniformly distributed incident photon flux. That is, the expected number of photons at each pixel is the Binomial mean N/M. The number of photons landing in the active area of a single pixel is reduced by the probability ε, the geometric fill factor. The assumption is made (in this analysis) that any avalanche extinction happens very early on, and thus the beginnings of multiple avalanches (within the same pixel) remain physically separated. This allows those n_k avalanche precursors to be treated as independent events and therefore with binomial probability $P(\alpha_k \mid n_k)$. That is, after n_k trials, there is a $P(\alpha_k \mid n_k)$ probability of α_k photoelectrons causing an avalanche.

In order to derive the probability of a single pixel firing, we begin by combining the distributions $P(N)$ and $P(N_k \mid N)$ to obtain the joint probability $P(N_k)$.

$$
\begin{aligned}
P\left(N_k = x\right) &= \sum_{y=0}^{\infty} P\left(N = y\right) P\left(N_k = x \mid N = y\right) \\
&= \sum_{y=0}^{\infty} \frac{e^{-\langle N \rangle} \lambda^{y}}{y!} \binom{y}{x} \left(\frac{1}{M}\right)^{x} \left(1 - \frac{1}{M}\right)^{y-x} \\
&= \left(\lambda / M\right)^{-x} \frac{e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{1}{\left(y-x\right)!} \left(\lambda \left(1 - \frac{1}{M}\right)\right)^{y-x} \\
&= \left(\lambda / M\right)^{-x} \frac{e^{-\lambda}}{x!} e^{\lambda\left(1 - 1/M\right)} \\
&= \left(\lambda / M\right)^{-x} \frac{e^{-\lambda / M}}{x!} \\
&= \text{Poisson}\left(\frac{\lambda}{M}\right)
\end{aligned}
\tag{3.1}
$$

which is the probability of x photons striking anywhere on pixel k, given an average total photon flux of λ. Note that the combination of a Poisson random variable with a Binomial random variable results in a new Poisson random variable, a result which can be directly reapplied (with a sufficient proof). Repeating for the remaining random variables, we first obtain the probability of finding n_k photons in the active region of pixel k,

$$
P\left(n_k\right) = \text{Poisson}\left(\frac{\lambda \varepsilon}{M}\right) \tag{3.2}
$$

and the probability of those photons (now photoelectrons) in the active area of pixel k initiating an avalanche,

$$
P\left(\alpha_k\right) = \text{Poisson}\left(\frac{\lambda \varepsilon p}{M}\right). \tag{3.3}
$$

The parameter $\lambda \varepsilon p / M$ is thus the average number of photons that land in the active area of pixel k that *could possibly* cause one avalanche.

Next, we treat the unique binary nature of each pixel. Because only one avalanche can be created at a time, the probability that pixel k will fire is simply the complement to the probability that no photoelectrons initiate an avalanche. That is, the probability that one or more photoelectrons *do* initiate an avalanche will be

$$
\begin{aligned}
p_{\text{fire}} &\equiv P(\text{pixel } k \text{ fires}) \\
&= 1 - P(\alpha_k = 0) \quad . \\
&= 1 - e^{-\lambda\varepsilon p/M}
\end{aligned}
\tag{3.4}
$$

This probability is the root of the potentially nonlinear response of a silicon photomultiplier.

3.2.A. Non-Uniform Spatial Distribution of Incoming Photons

When incident photons are equally distributed over the surface of the detector, each pixel will see an average of $1/M$ of the N incident photons. This can occur for very thick (or high aspect ratio) crystals or when there is a sufficiently thick optical interface (e.g. light pipe) between the scintillator and the photodetector. However, if incident photons are concentrated in a particular area of the device, as with thin crystals, then the pixel firing probability p_{fire} will depend on the pixel k. A more locally dense flux will fire more pixels, but there will be more nonlinearity from photons lost to simultaneous detection by a single pixel. If the spatial density is both constant and known, then it can be easily factored into the proceeding analysis.

As an example, let photons be linearly distributed across the face of a one-dimensional detector such that the lowest probability of incidence occurs at pixel $k = 1$. The probability of N_k photons at pixel k, given N initial photons becomes

$$
P(N_k \mid N, k) = \text{Binomial}\left(N, 2k/M^2\right) \tag{3.5}
$$

which leads to a pixel-dependent firing probability

$$
p_{\text{fire}}(k) = 1 - e^{-2\lambda\varepsilon pk/M^2} \tag{3.6}
$$

which is now dependent on the specific pixel of incidence k. More realistic spatial distributions should be obtained from Monte Carlo optical simulations for a given system description.

3.3. Multiple Pixel Mean and Variance

Having determined the probability that any given pixel will fire, the next step is to determine the statistics for m, the number of pixels which fire. By taking M

trials (M pixels total) each with success probability p_{fire}, the number of pixels fired is determined to be binomially distributed.

$$
\begin{gathered}
P(m \text{ pixels fire}) = \text{Binomial}(M, p_{\text{fire}}) \\
P(m) = \binom{M}{m} p_{\text{fire}}{}^{m} (1 - p_{\text{fire}})^{M-m}
\end{gathered}
\tag{3.7}
$$

Given this binomial nature, expressions can be immediately obtained for the mean number of pixels fired

$$
\langle m \rangle = M p_{\text{fire}} = M\left(1 - e^{-\lambda \varepsilon p/M}\right) \tag{3.8}
$$

as well as the variance in the number of pixels fired

$$
\sigma_m^2 = M p_{\text{fire}} \left(1 - p_{fire}\right) \tag{3.9}
$$

These statistics describe some of the basic features observed in a silicon photomultiplier spectrum for a number of pixels M, geometric efficiency ε, and mean number of incident photons λ. There are however other factors which must eventually be incorporated to accurately represent the spectra from physical devices. Some of these *intrinsic* device factors include:

- quantum efficiency
- avalanche excess noise
- thermal dark counts
- afterpulsing
- optical crosstalk
- recovery time

Extrinsic factors such as:

- temperature
- excess bias
- photon distribution (in number, space, and time)

will also affect the relative contributions of a number of the intrinsic factors above. A full model of a silicon photomultiplier therefore must be inclusive of the inter-relationships between the intrinsic and extrinsic parameters above.

3.3.A. Non-Uniform Spatial Distribution

Returning to the spatially linear photon flux above, we find that the total number of pixels that fire also depends on the spatially-dependent pixel firing probability.

$$\langle m \rangle = \sum_{k=1}^{M} p_{\text{fire}}(k) \tag{3.10}$$

For instance, let $M = 400$, $\varepsilon = 0.5$, $p = 0.8$, and $\lambda = 4000$. When the 4000 photons are distributed uniformly, an average of 393 pixels fire, whereas only 351 pixels fire for the linearly distributed case. This 11% difference would be a significant contributor to the energy resolution if both spatial distributions were to occur over the course of one measurement. For instance, lower energy radiation may be preferentially absorbed near the surface of a crystal, while higher energy photons would penetrate more deeply. The origin of the scintillation photons within the crystal may impact the exiting spatial photon flux seen by the photodetector. Therefore, the spatial distribution of incoming photons, as well as the *range* of possible distributions, should be well understood.

3.4. Incident Photon Flux Estimation

While the number of pixels that fire may be useful in a limited number of relative measurements, an accurate estimate of the incident photon flux is usually the information of interest. Taking a maximum-likelihood approach, we maximize the log-likelihood with respect to p_{fire} since it is directly related to λ. In the case of scintillation detection, λ will vary from event to event; not only because of stochastic loss mechanisms, but also due to the distribution of incident and absorbed energies. Therefore, without significant prior knowledge of the system physics and radiation field we are attempting to measure, we cannot make use of more than one scintillation event at a time to estimate the incident number of photons. Each recorded scintillation event must be individually transformed from a number of pixels fired into an estimated number of incident photons.

3.4.A. Maximum Likelihood Estimator

The principle of maximum likelihood provides a method for obtaining the parameter values that make observed data most likely. Of primary interest is an estimate for the number of incident photons λ, which appears in the pixel firing probability p_{fire}. We therefore first seek a maximum likelihood estimator (MLE) for the non-random parameter p_{fire} using what we measure: the number of pixels fired m.

Let $m_1, m_2, \ldots, m_n$ be independent and identically distributed (i.i.d.) realizations / outcomes / samples from a random variable $m \sim \text{Binomial}(M, p_{\text{fire}})$. The probability mass function (PMF) for m is

$$f_m(m_i) = \binom{M}{m_i} p_{\text{fire}}^{\;m_i} (1 - p_{\text{fire}})^{M - m_i} \tag{3.11}$$

which gives the likelihood function

$$\begin{aligned} L(p_{\text{fire}}) &= \prod_{k=1}^{n} f_{p_{\text{fire}}}(m_k) \\ &= \prod_{k=1}^{n} \binom{M}{m_k} \cdot \prod_{k=1}^{n} p_{\text{fire}}^{\;m_k} \cdot \prod_{k=1}^{n} (1 - p_{\text{fire}})^{M - m_k} . \\ &= C \cdot p_{\text{fire}}^{\sum_{k=1}^{n} m_k} \cdot (1 - p_{\text{fire}})^{nM - \sum_{k=1}^{n} m_k} \end{aligned} \tag{3.12}$$

The log-likelihood function is

$$\begin{aligned} l(p_{\text{fire}}) &= \log L(p_{\text{fire}}) \\ &= \log C + \log(p_{\text{fire}}) \sum_{k=1}^{n} m_k + \log(1 - p_{\text{fire}}) \left(nM - \sum_{k=1}^{n} m_k \right) \end{aligned} \tag{3.13}$$

which has the derivative

$$\begin{aligned} \frac{d}{dp_{\text{fire}}} l(p_{\text{fire}}) &= \frac{1}{p_{\text{fire}}} \sum_{k=1}^{n} m_k - \frac{1}{1 - p_{\text{fire}}} \left(nM - \sum_{k=1}^{n} m_k \right) \\ &= \frac{1}{p_{\text{fire}}(1 - p_{\text{fire}})} \sum_{k=1}^{n} m_k - \frac{nM}{1 - p_{\text{fire}}} \end{aligned} , \tag{3.14}$$

and the equations

$$
\begin{aligned}
\frac{d}{dp_{\text{fire}}} l\left(p_{\text{fire}}\right) &= 0 \\
\frac{1}{p_{\text{fire}}\left(1-p_{\text{fire}}\right)} \sum_{k=1}^{n} m_k &= \frac{nM}{1-p_{\text{fire}}} \\
\frac{1}{p_{\text{fire}}} \sum_{k=1}^{n} m_k &= nM
\end{aligned}
\tag{3.15}
$$

have the solution

$$
\hat{p}_{\text{fire}} = \frac{\bar{m}}{M}. \tag{3.16}
$$

where $\bar{m}$ is the sample mean number of pixels fired. It can be shown that MLE's are invariant to one-to-one functional transformation [Kay93]. For instance, the MLE parameters of the log-normal distribution are equivalent to those of the normal distribution fitted to the log of the data. Because from equation (3.4), λ can be stated as a one-to-one function of p_{fire}, the MLE $\hat{\lambda}$ for λ is simply

$$
\begin{aligned}
\hat{\lambda} &= \frac{-M}{\varepsilon p} \log\left[1-\hat{p}_{\text{fire}}\right] \\
&= \frac{-M}{\varepsilon p} \log\left[1-\frac{\bar{m}}{M}\right]
\end{aligned}. \tag{3.17}
$$

This accords well with a perhaps naïve intuition of obtaining an estimate of λ by inverting equation (3.4). It would however be an inappropriate assumption to simply invert equation (3.4) without sufficient mathematical reason. Again, because we assume we can only make estimates based on one scintillation event at a time, the average number of pixels fired $\bar{m}$ in the equation (3.17) above will simply be a scalar value *m*.

This is admittedly a rather weak application of the maximum likelihood principle since our number of observations $n = 1$. However, it provides a reliable framework from which to posit and compare the mean and variance in the number of pixels and the number of photons that relate to the radiation energy we are attempting to measure. Any nonlinearities in a pixels-fired spectrum can be linearized into a photons-estimated spectrum via the MLE estimator from equation (3.17).

3.5. Variance in the Estimate of Incident Photon Flux Estimation

Having developed an estimator for the number of incident photons, we are next interested in deriving the approximate variance of that estimator. This variance will allow the prediction of the width of spectral peaks and ultimately the energy resolution from a given silicon photomultiplier. First, we extend the information theory approach to compute a lower bound on the variance. Following that, several approximations will be introduced and shown to be equivalent to that lower bound.

3.5.A. Cramér-Rao Lower Bound on Estimator Variance

We can compute the lower bound of the variance in the estimator from equation (3.17) through the Fisher information and the Cramér-Rao lower bound (CLRB) [Kay93]. The first derivative of the log-likelihood function, with respect to λ is

$$\begin{aligned}\frac{d}{d\lambda}l(\lambda)&=\frac{d}{d\lambda}\left(\log C+\log\left(1-e^{-\lambda\varepsilon p/M}\right)\sum_{k=1}^{n}m_k+\log\left(e^{-\lambda\varepsilon/M}\right)\left(nM-\sum_{k=1}^{n}m_k\right)\right)\\&=\frac{\varepsilon p}{M}\frac{1}{e^{\lambda\varepsilon p/M}-1}\sum_{k=1}^{n}m_k-\frac{\varepsilon p}{M}\left(nM=\sum_{k=1}^{n}m_k\right)\end{aligned}, \quad (3.18)$$

whose second derivative is

$$\begin{aligned}\frac{d^2}{d\lambda^2}l(\lambda)&=\frac{d}{d\lambda}\left(\frac{\varepsilon p}{M}\frac{1}{e^{\lambda ep/M}-1}\sum_{k=1}^{n}m_k-\frac{\varepsilon}{M}\left(nM-\sum_{k=1}^{n}m_k\right)\right)\\&=-\frac{\varepsilon^2p^2}{M^2}\frac{e^{\lambda ep/M}}{\left(e^{\lambda ep/M}-1\right)^2}\sum_{k=1}^{n}m_k\end{aligned}, \quad (3.19)$$

and the expectation of this value becomes the Fisher information

$$\begin{aligned}-\left\langle\frac{d^2}{d\lambda^2}l(\lambda)\right\rangle&=\frac{\varepsilon^2p^2}{M^2}\frac{e^{\lambda\varepsilon p/M}}{\left(e^{\lambda\varepsilon p/M}-1\right)^2}\sum_{k=1}^{n}\langle m_k\rangle\\&=\frac{\varepsilon^2p^2}{M^2}\frac{e^{\lambda\varepsilon p/M}}{\left(e^{\lambda\varepsilon p/M}-1\right)^2}M\left(1-e^{-\lambda\varepsilon p/M}\right)n\,,\\&=\frac{\varepsilon^2p^2n}{M}\frac{1}{e^{\lambda\varepsilon p/M}-1}\end{aligned} \quad (3.20)$$

from which the Cramér-Rao lower bound follows as

$$\sigma_{\hat{\lambda}}^2 \geq \frac{1}{-\left\langle \frac{d^2}{d\lambda^2} l(\lambda) \right\rangle}$$
$$= \frac{-M}{\varepsilon^2 p^2 n}\left(1 - e^{\lambda \varepsilon p/M}\right). \tag{3.21}$$
$$= \frac{M}{\varepsilon^2 p^2 n} \frac{p_{\text{fire}}}{1 - p_{\text{fire}}}$$

Note the dependence on n, the number of observations. Again, because we can only make use of one scintillation event at a time (i.e., $n = 1$), this is perhaps a weak application of the CRLB.

In terms of the number of pixels fired m, an approximate CRLB can be stated as

$$\sigma_{\hat{\lambda}}^2 \geq \frac{m}{\varepsilon^2 p^2 d} \tag{3.22}$$

where $d = (1 - m/M)$ is the fraction of pixels that do *not* fire and the assumption is made that $p_{\text{fire}} \approx \hat{p}_{\text{fire}} = m/M$. Intuitively, this means that increasing geometric efficiency ε and avalanche initiation probability p can have a dramatic impact on the energy resolution, as expected. The m in the numerator should accord well with an understanding of the underlying Poisson process of the incident photon flux where mean is equivalent to variance. The d in the denominator illustrates an improvement in resolving power when a small fraction of pixels fire. This would tend to be an argument for a greater number of pixels. That argument of course must be balanced by the additional dark noise that a greater number of pixels can bring if total active area is increased. If SiPM active device area can be held constant, then the ideal geometry would be an infinite number of infinitely small pixels (like a continuous photocathode). Electric field and edge breakdown concerns of course limit the minimum pixel size.

We can also specify a kind of signal to noise ratio based on the estimator in equation (3.17) and the square root of the CRLB in equation (3.21).

$$\text{SNR} = \frac{\hat{\lambda}}{\sigma_{\hat{\lambda}}} = \frac{1}{\log(1 - p_{\text{fire}})} \sqrt{M \frac{1 - p_{\text{fire}}}{p_{\text{fire}}}}. \tag{3.23}$$

Likewise, if the photon spectrum peaks are reasonably Gaussian, then the best estimated photon spectrum peak resolution we can expect would be

$$\text{Resolution} \approx 2.35 \frac{\sigma_{\hat{\lambda}}}{\hat{\lambda}} = 2.35 \log\left(1 - p_{\text{fire}}\right) \sqrt{\frac{p_{\text{fire}}}{M\left(1 - p_{\text{fire}}\right)}} . \tag{3.24}$$

This complex behavior can be much different from the traditional Poisson relationship and will be illustrated in later sections.

3.5.B. Functional Expansion Method

The variance of the function of any random variable can be approximated by expanding the function about its mean and dropping several higher order terms [Pap91].

$$\begin{aligned} \sigma_{\hat{\lambda}}^2 &\cong \left|\hat{\lambda}'(m)\right| \sigma_m^2 - \left(\left|\hat{\lambda}''(m)\right|^2 \frac{\sigma_m^2}{2}\right)^2 \\ &\cong \frac{M p_{\text{fire}}\left(1 - p_{\text{fire}}\right)}{\varepsilon^2 p^2 \left(1 - m/M\right)^2} - \left(\frac{p_{\text{fire}}\left(1 - p_{\text{fire}}\right)}{2\varepsilon p\left(1 - m/M\right)^2}\right)^2 , \end{aligned} \tag{3.25}$$

or truncating to the first order

$$\begin{aligned} \sigma_{\hat{\lambda}}^2 &\cong \left|\hat{\lambda}'(m)\right|^2 \sigma_m^2 \\ &= \frac{M}{\varepsilon^2 p^2} \frac{p_{\text{fire}}\left(1 - p_{\text{fire}}\right)}{\left(1 - m/M\right)^2} . \end{aligned} \tag{3.26}$$

Taking m to be $\langle m \rangle$, the approximate variance of our estimator becomes

$$\begin{aligned} \sigma_{\hat{\lambda}}^2 &\cong \frac{M}{\varepsilon^2 p^2} \frac{p_{\text{fire}}\left(1 - p_{\text{fire}}\right)}{\left(1 - M p_{\text{fire}}/M\right)^2} \\ &= \frac{M}{\varepsilon^2 p^2} \frac{p_{\text{fire}}}{1 - p_{\text{fire}}} \\ &= \frac{M}{\varepsilon^2 p^2}\left(e^{\lambda \varepsilon p/M} - 1\right) \end{aligned} \tag{3.27}$$

which we note achieves the Cramér-Rao lower bound in equation (3.21), when the number of observations n = 1. Therefore, our estimator may be considered "efficient" to the first order.

3.5.C. Linearity of Logarithm Method

If the fraction of pixels which fire is small, the argument within the logarithm of the MLE will be close to 1, and we may expect $E\left[\log(X)\right] \approx \log\left(E[x]\right)$. We can use this proposition to demonstrate the approximate unbiasedness of our estimator.

$$\begin{aligned} \langle\hat{\lambda}\rangle &= \left\langle \frac{-M}{\varepsilon p} \log\left[1 - \frac{m}{M}\right]\right\rangle \\ &\approx \frac{-M}{\varepsilon p} \log\left[1 - \frac{\langle m \rangle}{M}\right] \\ &= \frac{-M}{\varepsilon p} \log\left[1 - \frac{M\left(1 - e^{-\lambda \varepsilon p/M}\right)}{M}\right] \\ &= \lambda \end{aligned} \tag{3.28}$$

The MLE estimator is thus an efficient and unbiased maximum likelihood estimator *to the first order*, when the fraction of pixels fired is small.

Using the same argument, the variance in the logarithm of a random variable X might then be approximated as

$$\begin{aligned} \sigma^2_{\log(X)} &= E\left[\log(X)^2\right] - E\left[\log(X)\right]^2 \\ &\approx \log\left(E\left[X^2\right]\right) - \log\left(E[X]^2\right) . \\ &= \log\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \end{aligned} \tag{3.29}$$

The application of this approximation along with several expectation identities produces an approximation to the variance in the estimated incident photon flux.

$$\begin{aligned} \sigma^2_{\hat{\lambda}} &= \operatorname{var}\left[\frac{-M}{\varepsilon p} \log\left(1 - \frac{m}{M}\right)\right] \\ &\approx \frac{M^2}{\varepsilon^2 p^2} \log\left(1 + \frac{\operatorname{var}\left[1 - m/M\right]}{E\left[1 - m/M\right]^2}\right) \\ &= \frac{M^2}{\varepsilon^2 p^2} \log\left(1 + \frac{1}{M^2} \frac{\sigma_m^2}{\left(1 - \mu_m/M\right)^2}\right) \\ &= \frac{M^2}{\varepsilon^2 p^2} \log\left(1 + \frac{1}{M} \frac{p_{\text{fire}}}{1 - p_{\text{fire}}}\right) \end{aligned} \tag{3.30}$$

Note the equivalence to the CRLB when $\log(1+x)$ is approximated to the first order by *x*.

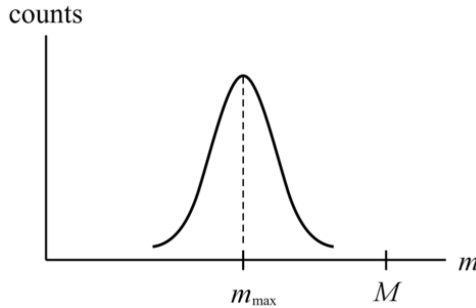
$$\log(1+x) \sim x$$
$$\Downarrow$$
$$\frac{M^2}{\varepsilon^2 p^2}\log\left(1+\frac{1}{M}\frac{p_{\text{fire}}}{1-p_{\text{fire}}}\right) \sim \frac{M}{\varepsilon^2 p^2}\frac{p_{\text{fire}}}{1-p_{\text{fire}}} \tag{3.31}$$
$$= \frac{M}{\varepsilon^2 p^2}\left(e^{\lambda\varepsilon p/M}-1\right)$$

The equivalent estimator variance relationships presented in this section serve as the basis for the estimation of the energy resolution intrinsic to the silicon photomultiplier and will be illustrated in following sections.

3.6. Systematic Features

3.6.A. Dynamic Range

As the incident photon flux increases, there exists an increasing probability that the device may saturate, defined as all pixels firing simultaneously. When $M-1$ or fewer pixels fire, accurate estimation of the incident photon flux is possible. However, should all *M* pixels fire, the estimated incident photon flux may be anything greater than that predicted by $\hat{\lambda}(m=M-1)$.



We would like to specify the maximum *mean* number of pixels that may be allowed to fire so that the probability of total device saturation is low. We do this with a confidence interval defined by

$$\langle m\rangle + s\sqrt{\sigma_m^2} = M \tag{3.32}$$

where s is the number of standard deviations between the maximum allowable *mean* number of fired pixels $\langle m \rangle$ and the total number of available pixels M. We can then solve for both the maximum allowable number of pixels fired

$$m_{\max} = \frac{M}{1+s^2/M}\ , \tag{3.33}$$

and the maximum allowable *mean* incident photon flux

$$\begin{aligned}\lambda_{\max} &= \hat{\lambda}(m_{\max}) \\ &= \frac{-M}{\varepsilon p}\log\left(\frac{1}{1+M/s^2}\right)\end{aligned}. \tag{3.34}$$

For example, at six standard deviations, a 1600-pixel device would be nearly saturation-free below 1564 pixels (97.8% of total pixels) or ~24000 incident photons. This expression for dynamic range will require modification as additional terms are added to the definition of pixel variance and the measured distribution is widened.

3.6.B. Quantization Error

The information obtained directly from a silicon photomultiplier is primarily a discrete pixel spectrum. If the pixel spectrum is quantized before being transformed into a photon intensity spectrum, then the constant pixel-to-pixel quantization error of 1 (pixel) will be nonlinearly transformed into a photon-dependent error. We can examine this error by taking the first finite difference of our estimator. The forward, backward, and central pixel differences yield similar results. Here we state the estimation quantization error for a central pixel difference.

$$\begin{aligned}\hat{\lambda}(\Delta m) &= \hat{\lambda}(m+1) - \hat{\lambda}(m-1) \\ &= \frac{-M}{\varepsilon p}\log\left(\frac{M-m-1}{M-m+1}\right)\end{aligned} \tag{3.35}$$

Relative to the estimated photon flux, this truncation error can be stated as the fraction

$$\frac{\hat{\lambda}(\Delta m)}{\hat{\lambda}(m)} = \frac{\log\left(\frac{Md-1}{Md+1}\right)}{\log(d)} \tag{3.36}$$

where d is the fraction of pixels which do not fire. For a sufficiently large number of pixels, this error can be small enough to be neglected, as indicated in Figure 3.2.

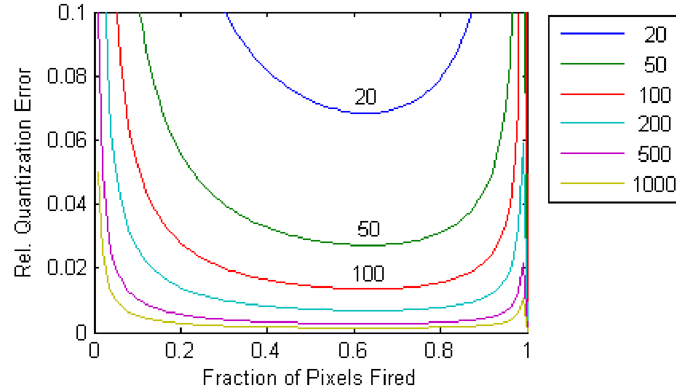


Figure 3.2. The figure above illustrates the relative quantization error for devices with M = 20 to 1000 pixels.

When relatively few pixels fire, the quantization error is relatively low compared to the $1/E$ uncertainty from counting statistics. On the right side of Figure 3.2, the quantization error increases because of the expansion of an increasingly nonlinear response. Still, this error will likely only be seen in simulation or in cases of extreme saturation. This effect is included in the simulation of Section 3.7 and provides for a greater match between theory and simulation.

3.6.C. Avalanche Excess Noise Error

The ideal silicon photomultiplier would produce a single integer representing the number of pixels which fire. In fact, low fill-factor devices with integrated digital electronics have already realized this type of digital SiPM [Fra09]. However, the amount of charge generated and collected per pixel is *not* a constant in an analog SiPM with passive quenching. The variance in the charge collected from a single pixel adds in quadrature as many pixels fire. Eventually, this will lead to a variance in the estimate of the number of pixels which fire. It is this variance that prevents visual observation of individual peaks in a passively quenched pixels-fired spectrum beyond 10-20 pixels. This variation is due to

variation in doping, quench resistance, and avalanche evolution. The *avalanche* excess noise factor F is defined by

$$F = \frac{E\left[m^2\right]}{E\left[m\right]^2} = \frac{\sigma_m^2 + \langle m \rangle^2}{\langle m \rangle^2} = \sigma_1^2 + 1 \tag{3.37}$$

where σ_1^2 is the variance in the number of pixels fired when *one* pixel fires $\left(\langle m \rangle = 1\right)$. Note that the excess noise factor tends to 1 as the single pixel peak narrows to a delta function. Thus, the additional variance in the number of pixels which fire is given by

$$\begin{aligned} \sigma_{m,ENF}^2 &= m\sigma_1^2 \\ &= m\left(F-1\right) \end{aligned} \tag{3.38}$$

which must be added to the original pixel variance definition and will contribute to the estimated photon variance.

As an illustration, assume that the peaks in a pixels-fired spectrum are half-blurred at 10 pixels. This means that the full width at half maximum (FWHM) is one pixel wide at 10 pixels-fired. Another way of saying this is that the width of the sum of 10 Gaussians is equivalent to the distance (in charge, voltage) between individual pixels. Simply

$$\begin{aligned} \mathrm{FWHM}_{10} &= 2.35\sigma_T \\ &= 1\ \mathrm{px} \end{aligned} \tag{3.39}$$

so the total variance in the tenth pixel from avalanche excess variation will be

$$\begin{aligned} \sigma_T^2 &= \frac{1\ \mathrm{px}^2}{2.35^2} \\ &= 0.18\ \mathrm{px}^2 \end{aligned}, \tag{3.40}$$

and the single pixel avalanche excess variation is

$$\begin{aligned} \sigma_1^2 &= \frac{1}{10}\sigma_T^2 \\ &= 0.018\ \mathrm{px}^2 \end{aligned}, \tag{3.41}$$

making the avalanche excess noise factor

$$F = \sigma_1^2 + 1 \\ = 1.018 \quad (3.42)$$

Now assume that an average of 500 such pixels fire. What is the effect on the pixel peak width from this avalanche excess error? From Poisson counting considerations, the peak will have a *minimum* FWHM (in pixels) of

$$\begin{aligned} \text{FWHM}_{\text{counting}} &= 2.35\sigma \\ &= 2.35\sqrt{500} \\ &= 52.5 \end{aligned} \quad (3.43)$$

The single-pixel avalanche excess variance will add in quadrature, 500 times, so that the contribution to the peak width will be

$$\begin{aligned} \text{FWHM}_{\text{excess}} &= 2.35\sqrt{500\sigma_1^2} \\ &= 2.35\sqrt{500 \cdot 0.018} \\ &= 7.05 \end{aligned} \quad (3.44)$$

We can see from this example that even though the pixel peaks begin to be obscured relatively quickly in the spectrum, the effect on the full-energy peaks themselves is relatively low.

3.7. Simulation Example

In order to illustrate the spectral features described in this chapter, a Monte Carlo simulation is performed with a 1600 pixel device with $\varepsilon p = 0.25$. Photons are limited to an average of 24436 (or 1565 pixels fired) in order to avoid total device saturation. The MLE and CRLB are used to estimate the incident photon flux and its variance. The quantization error is also taken into account to improve the variance estimate at high photon flux.

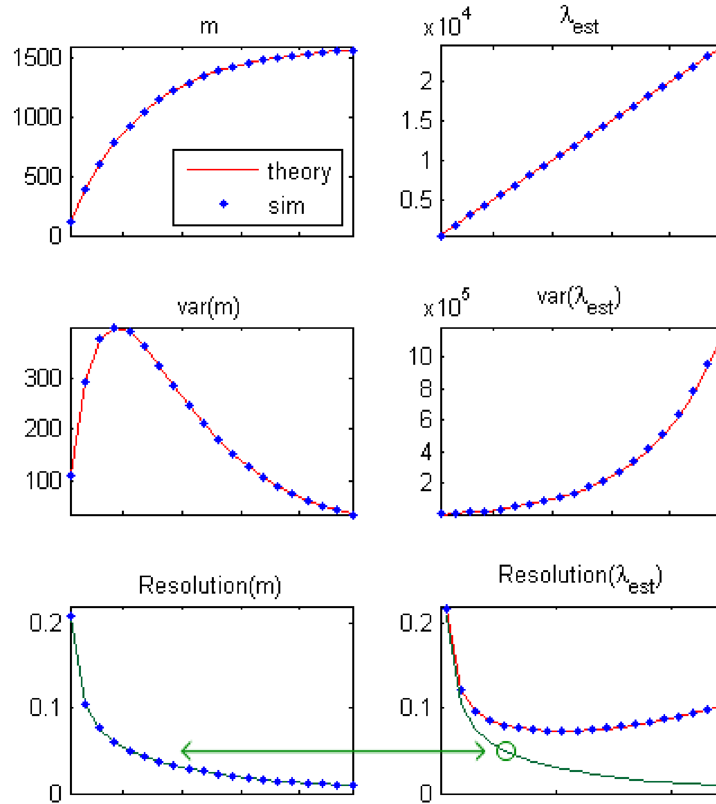


Figure 3.3. The mean, variance, and relative FWHM (resolution) are displayed for both number of pixels fired *m* (left-hand column) and estimated number of incident photons λ_{est} (right-hand column), as a function of known incident flux (from 1 to 24436).

Notice first in Figure 3.3 the nonlinear behavior of *m* with the number of incident photons (horizontal axis). The MLE is observed to *linearize* the response of the detector. The pixel spectrum resolution in the bottom left panel is a result of the binomial nature of the saturated pixel response. In the bottom right panel however, we note that the resolutions for both *m* (green line) and λ_{est} (red line) diverge significantly as the number of incident photons increases beyond the linear regime. The minimum in the estimated-photon resolution curve represents a limit on the spectral peak resolving power for a measured energy spectrum. This intrinsic photodetector resolution may in fact degrade the resolution obtained from otherwise highly-resolving bright scintillators and should be considered thoroughly before selection of scintillator or silicon photomultiplier parameters.

3.8. Recovery Time

In the preceding discussion, it was initially assumed that all photons impinge upon the photodetector simultaneously, when in fact scintillators can vary significantly in their exponential decay time constants. Additionally, stochastic optical photon transport tends to broaden the scintillation photons into a bi- or multi-exponential pulse. In this section, we relax the instantaneous arrival assumption and consider the consequences of relatively slowly recharging pixels.

As each avalanche is quenched, the lowered bias slowly returns to the original excess bias with a time constant governed by the quench resistor and the pixel diode capacitance.

$$V(t) = V_q \exp\left(-t/\left(R_q C_d\right)\right) \tag{3.45}$$

If the time constant of the scintillator is *larger* than this recovery time, then it becomes possible for one pixel to produce *more than one* avalanche per scintillation event, thus reducing the nonlinearity in the discussion above. As an example, the decay constant τ_{scint} for NaI is 230 ns. Letting C_d = 500 fF and R_q = 100 kΩ (similar to real devices), the pixel recovery time constant will be 50 ns. Defining, somewhat arbitrarily, the recovery time as 95% of the original bias, the recovery time will be

$$t_{\text{recovery}} \equiv \text{-100 k}\Omega \cdot \text{500 fF} \cdot \log(1 - 0.95) = \text{150 ns}. \tag{3.46}$$

This is one motivation for decreasing the pixel size, and thus its capacitance and recovery time. In reality, the definition of "recovery" is difficult to accurately define, because avalanches of increasingly large total charge will be produced during recovery as the bias increases above the breakdown voltage. This range of pulse magnitudes will alter the measured energy from event to event.

The scintillation process (and optical transport) can be modeled as a non-homogeneous Poisson process (NHPP). That is, the number of scintillation photons arriving at the photodetector in an infinitesimal time interval will follow a Poisson distribution with a rate parameter $\lambda(t)$ that is a function of time. In this case $\lambda(t)$ is the exponential function of the scintillation decay process. There are

many congruent problems in the theory of stochastic processes, the foremost of which belong to the discipline of queuing theory [Gro85]. One can envision the present detection process as a single server who takes a defined amount of time to serve a customer. Any customer seeing a busy server, leaves the establishment and is considered lost. The original theory for this type of loss model was developed by Erlang [Erl09] to address the distribution of waiting times and congestion in telephony applications. Fundamental to queuing theory is the theory of Markov chains whose state transition probabilities also find application in birth-death processes. Of significance in applying stochastic theories to this problem is that the vast majority of aforementioned theories have been developed only for stationary processes, whereas the scintillation event is definitively a time-bounded and time-varying stochastic process. Before applying any theory, we examine the stochastic features of potential interest.

3.8.A. Monte Carlo Illustration

A Monte Carlo approach is used to illustrate the effects of recovery time for a single pixel. For simplicity, let $\varepsilon p = 1$ and the average number of photons striking the pixel $\langle n_k \rangle = \{5,1\}$. In the non-transient or static case, these parameters lead to pixel firing probabilities of

$$p_{\text{fire}} = 1 - e^{-\{5,1\}} = \{0.993, 0.632\} \, . \tag{3.47}$$

From a frequency viewpoint, these two probabilities can be interpreted as the fraction of times a single pixel will fire from an ensemble of many scintillation events. In the transient case, this single pixel may now have the opportunity to fire multiple times for each scintillation event.

The simulation begins by generating uniform Poisson points and transforming them into NHPP points representing incident photons. The time intervals between successive photon arrivals are sequentially compared with the recovery time to identify photons which will be discarded. The first interval that is less than the recovery time indicates a photon arriving during recovery and is removed from the set. The entire set of photon arrival intervals is repeatedly tested until all

photons forming overlapping recovery intervals are removed simulating a paralyzable process. This process is repeated a statistically significant number of times, representing many scintillation events. The distribution (Figure 3.4) and mean (Figure 3.5) number of *detected* photons per pixel are illustrated below. In accord with intuition, the largest number of avalanches are observed with the shortest recovery time (or *slowest* scintillators).

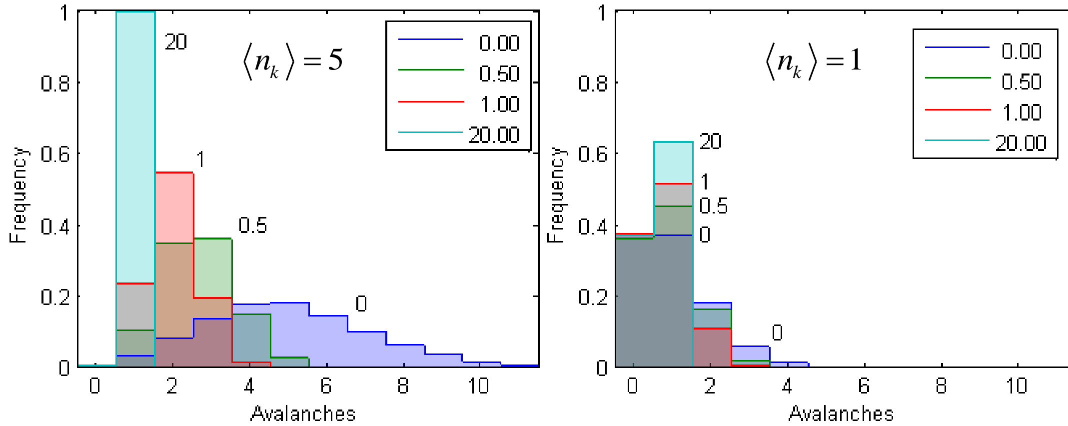


Figure 3.4. The probability distributions for the number of avalanches (from a single pixel) per scintillation event are shown for several different relative recovery times $t_{\text{recovery}}/\tau_{\text{scint}}$, as indicated by the legend entries. Mean incident photon fluxes of $\langle n_k \rangle = 5$ (left) and $\langle n_k \rangle = 1$ (right) are illustrated.

For the $\langle n_k \rangle = 5$ case (left of Figure 3.4), with $t_{\text{recovery}}/\tau_{\text{scint}} = 20$ (relatively slow recovery or fast scintillator), we note that there is almost zero probability of the pixel *not* firing. In fact, this probability is $1 - p_{\text{fire}} = 1 - 0.993 = 0.007$. There is also zero probability that the pixel will fire more than once per scintillation, which is consistent with the instantaneous detection assumption of the static case. As the recovery time decreases (or scintillator slows), both the mean and the variance increase. This variance will play an important role in determining the ultimate photon detection resolution for saturated devices with various scintillators.

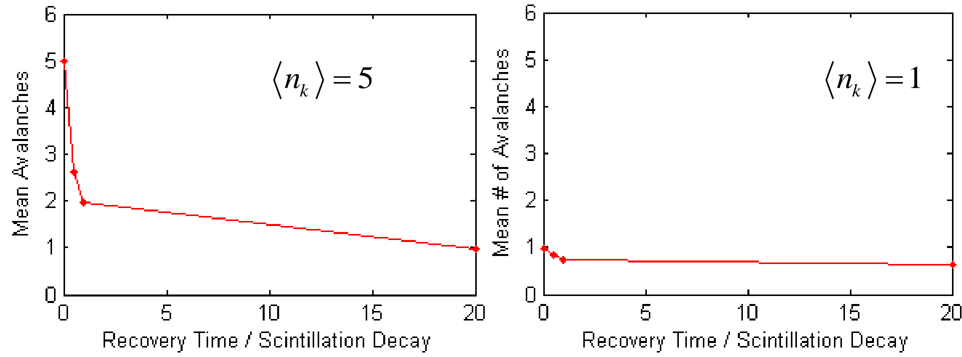


Figure 3.5. The mean number of avalanches from the distributions in Figure 3.4.

For the $\langle n_k \rangle = 1$ case (right of Figure 3.4), the slow-recovery pixel will still fire only once at most, but there is an increased probability that the pixel will not fire (0.368). Even with such a low fluence, there is still a slight gain in number of avalanches for fast-recovery pixels. Considering the *high* and *low* fluence cases together, it is concluded that high photon fluence (or spatial density) may be permissible for fast-recovery pixels or for slower scintillators. Therefore, a slow, dim scintillator may in fact outperform its faster, brighter counterpart if pixels are saturated. However, more device dark counts will also be included when using slower scintillators or longer integration times. This additional noise may outweigh any benefits obtained by single pixels firing multiple times per scintillation.

3.9. Noise Mechanisms

3.9.A. Thermally Generated Dark Counts

Any electron-hole generation site in or near an active region of the device will contribute dark counts at a rate that is dependent on the local trap states.

The energy spectrum noise contributed by thermally generated dark counts will largely be a function of the scintillation decay constant and thus the integration time or filter window width. A NaI crystal for example has a decay constant of 230 ns, therefore an integration time of 500 ns might be used to collect a majority of the photons $\left(1-e^{-500/250}=86\%\right)$. Now imagine that this crystal is coupled to a silicon photomultiplier with 1000 pixels, each with a dark count rate

of 10^4 cps. Assuming the dark counts from each pixel can be represented by a Poisson process, the total signal will be the sum of many such processes, which is itself a Poisson process. During the integration window, there will be a Poisson average of

$$1000 * (10^4\ s^{-1}) * (500\times10^{-9}\ s) = 5 \text{ counts.} \tag{3.48}$$

This dark count distribution with an average of 5 counts per event will be added on top of any scintillation events. For instance, the measured signal from a scintillation event firing say 132 pixels will be increased by an average of 5 pixels, resulting in a *relative* increase in energy resolution of $\sqrt{137/132} = 1.9\%$. It should be apparent that thermally generated dark counts are increasingly problematic for lower energy events. For reasonably bright events, a very large dark count rate may be tolerable, providing for much larger device areas. Clear application guidelines and specifications must be decided prior to selection of a SiPM based on noise considerations. Many applications may be able to tolerate very high dark noise devices which may significantly reduce costs. Triggering half of the 1000 pixels and tolerating a 10% *relative* increase in energy resolution at 500 ns shaping times would allow the single pixel dark count rate to be

$$(1.10^2 * 500 - 500) / 1000 / (500\times10^{-9}\ s) = 210 \text{ kcps.} \tag{3.49}$$

This level of tolerance is also the fundamental limiting factor to large area arrays of silicon photomultipliers. Scaling a square SiPM by a factor of two in length will increase the dark count rate by a factor of four. The cost of SiPMs (single die) will also increase nonlinearly with area due to production yield factors. For this reason, bonded and packaged arrays of dice with high fill fraction likely will be the preferred solution to large area photodetection areas. One alternative solution to the size vs. noise constraint is to focus the incoming photons into sufficiently few detectors with micro or macro lens arrays. High resolution SiPM scintillation detection systems may benefit from a reduced pixel count, reduced per-pixel dark count rates, and faster scintillators and shaping times.

3.9.B. Afterpulsing

The phenomenon of afterpulsing in avalanching junctions is a result of deep energy-level traps being filled and emptied at some later time. These traps are filled during an avalanche, and the electron or hole is confined there with an exponential time constant related to the energy of the trap. Many traps are sufficiently shallow (close to the electron or conduction bands) that they de-excite with a time constant much less than the pixel recovery time. In this case, the liberated electron or hole simply rejoins the hoard of energetic carriers waiting to recombine while the electric field is returned to its original excess bias. Traps which are more deep can be released with time constants up to milliseconds, especially if the device has been cooled. These slower traps will cause a single pixel to retrigger itself at a later time, and will be correlated with the previous occurrence of an avalanche. This can be a significant component of the dark current, depending on the applied bias.

Due to the exponential nature of the afterpulse retriggering process, the majority of afterpulses will be located in close temporal proximity to the original avalanche. Once a pixel has been retriggered once, the process can be considered to start again with similar probabilities. The actual physics of which traps fire when and if traps are filled or released during an avalanche are difficult to assess experimentally. For this reason, the experimenter will typically determine the average number of extra avalanches per photo-generated avalanche. This number, like the optical crosstalk probability, is typically low (<5%).

A number of factors can increase the effective afterpulse "multiplier". A very long integration time will record a greater number of afterpulses. Contamination in processing can produce a greater density of traps. Larger pixels will include a greater volume of traps. The larger capacitance from larger pixels will result in more charge per avalanche which may fill more traps. Bigger pixels may also require longer recharge times which may obscure charged carriers released from traps. Colder temperatures increase the effective trap time constants, so more afterpulsing may occur in a given time window. However, if the integration time

is on the order of the recharge time, then afterpulsing may only contribute to the low energy background. Because this process is so interdependent with other detection and noise mechanisms, an appropriate simulation or model must necessarily be limited to specific devices.

3.9.C. Optical Crosstalk

The phenomenon of optical crosstalk occurs when an energetic electron recombines with a hole and excess energy is released in the form of one or more optical photons. These photons can have a broad range of wavelengths and thus potentially can travel some distance in silicon. A second avalanche can be triggered if one of these photons is absorbed in the active region of a neighboring pixel. Devices with very closely spaced pixels are more prone to optical crosstalk due to the increased solid angle seen by emitted "avalanche photons".

The assumption is often made that photons emitted from an avalanching pixel do so at the very beginning of the avalanche. The consequence of this assumption is that once a pixel fires, it cannot be retriggered by its own optical crosstalk chain. If the crosstalk probability is reasonably large, then one pixel may trigger its neighbor and so on until the last photon is either absorbed in a non-active region, or it is absorbed in a currently firing pixel. This leads to a rather interesting statistical phenomenon akin to the "random walk". However, a true random walk is not constrained by the disallowance of revisitation. Instead, optical crosstalk in a silicon photomultiplier is more like a self-avoiding walk (SAW) which terminates upon the first revisitation of a previously firing pixel. In order to determine the probability of a single pixel triggering some number of other pixels, one must determine the total number of possible SAW's and their lengths. Due to the inherent complexity in this branching process, there is no known analytical formula for determining the number of self-avoiding walks [Liś03]. Instead, exhaustive numerical simulations are performed to arrive at the number of possible walks and their lengths [Hay98]. The length describes the number of pixels recorded. In extreme cases, a high crosstalk probability may be treated with percolation theory, in much the same way as a wildfire enveloping an entire forest.

This type of behavior is unlikely to be seen in real devices, although SSPM's made from III-V materials may exhibit much higher rates of hot-carrier emission.

The direct simulation of this phenomenon in its entirety is complicated further by the probable presence of many scintillation photons arriving near-simultaneously at many pixels throughout the detector. This can potentially shorten the length of each optical crosstalk SAW. It is therefore proposed that the optical crosstalk chain length distribution is a function of the photon fluence, particularly at higher photon flux densities. However, the optical crosstalk probabilities for the majority of devices is low enough to only require accounting for crosstalk SAW's of length two or three, and this effect may be of minor significance.

A simplified alternative to accounting for the exact distribution of extra pixels fired for each incident photon is to simply weight the incoming number of pixels fired by some "multiplier" [Joh10] related to the mean number of extra pixels measured. For the majority of devices with low crosstalk probability (<5%), this excess "gain" will allow the experimenter to account for the few additional pixels and correctly calibrate the peaks in an energy spectrum.

3.10. Nuclear Counter Effects

The nuclear counter effect is a potential source of noise in scintillation detection systems [Gro84], wherein a photodetector directly responds to the radiation particles meant for scintillation detection. If operated in proportional mode, solid-state detectors will provide a signal proportional to the number of electron-hole pairs created along some ionized track within the semiconductor. This can add a significant continuum of background events to the desired scintillation spectrum and add uncertainty by broadening peaks of interest. Avalanche photodiodes and particularly PIN photodiodes are most susceptible to this effect. Given the shallow junction depth (~1 μm), small pixel size (tens of microns), and Geiger mode binary nature of SPADs in a SiPM, the largest signal one can expect from any ionizing particle is equivalent to that of a single optical photon. Further, a blue-sensitive SiPM will only be responsive to those ionizing

particles that are stopped in the first several microns of silicon. The background radiation spectrum measurable with a SiPM alone will appear comparable to that of the thermally generated dark counts, making the photodetector a poor radiation detecting counter.

3.11. A Complete Model

The purpose for developing a model of the silicon photomultiplier is really two-fold. Looking forward, the experimenter would like to either accurately interpret the results of some measurement or decide on a device based on the expected data. In the case of scintillation detection for ionizing radiation, this would mean obtaining some model which provides for the estimation of the energy, position, or timing of incoming particles based on measured data. Looking backward, a complete device model would allow the device designer to probe the limits of a given processing technology. Such a model would allow the determination of the ultimate efficiency achievable and any potential tradeoffs.

In terms of applying a model to measured data, what often is done is simply to ensure the detection scheme is linear in energy, position, or time and then calibrate the scale based on interactions of known energy, position, or time. Thus, no physical model is needed of the detection whatsoever except what is necessary to linearize the detector response. This linearization in energy for the silicon photomultiplier requires the estimation theory set forth in Sections 3.2-3.4. However, even this theory may prove unnecessary if the incident photon flux is below the threshold of nonlinear saturation effects. Complicating noise factors like afterpulsing and optical crosstalk will broaden the measured energy spectrum, in much the same way as would an analog gain. Ignoring second order peak broadening effects (due to dark counts, afterpulsing, crosstalk, etc.), this linearization process is most likely sufficient for a large number of applications.

The purpose for prediction of second-order effects in detection systems is in the determination of which system elements contribute the greatest to the uncertainty with which a measurement is translated into an estimated energy, position, or time. Let the dark counts from a large silicon photomultiplier be

contributing significant uncertainty to the peaks in an energy spectrum. A potential solution might be to select a brighter or faster scintillator or concentrate the existing photons down onto a smaller photodetector. It is important to understand the relative contribution to energy resolution from the finite number of detected photons as well as the noise mechanisms inherent in a larger photodetector.

Chapter 4

Simulation of Optical Processes

The interface between an avalanche photodiode and the photon source for many applications consists of a large expanse of air or glass. Scintillation detectors, on the other hand, present the photodetector with a closed optical system in which the photodetector itself must be considered as an integral optical element for the reflection and refraction of isotropically emitted scintillation photons. For this reason, an intelligent choice of photodetector optical coating must be made. Photons arriving at the photodetector interface may be reflected back into the crystal, only to be returned for detection a second time. This marks a fundamental difference in photodetector response for different applications. For instance, silicon photomultipliers for laser ranging would not be intimately coupled to a closed system of optical reflectors, and therefore may require a different antireflection coating. In addition, the range of wavelengths of interest can vary dramatically from bioluminescence studies to scintillation detection to infrared ranging. Thus, the optical interfaces must be carefully chosen for the application at hand in order to maximize the number of detected photons, and minimize the detection of unwanted photons (e.g. hot-carrier emissions).

4.1. Antireflection Coatings for Scintillation Photodetectors

The inevitable change in index of refraction at the photodetector window boundary represents an increased probability that incoming scintillation photons will be turned away back into the crystal and potentially lost or absorbed. Given a known spectrum of scintillation photon wavelengths, a suitable antireflection coating (ARC) may be applied to reduce the overall probability of total internal

reflection at the photodetector surface. Simple single-layer ARC may consist of a thickness of transparent optical medium which is one quarter the wavelength of the incoming light. This would provide destructive interference for reflection, and would increase the transmission through the interface. For scintillation emissions of a range of wavelengths, multi-layered AR coatings with multiple indices of refraction may be tuned to provide optimal transmission from scintillator to photodetector at a given angle.

4.1.A. Incident Angle Distribution

In order to determine the ideal ARC for a photodetector, the incident photon angular distribution must be known. This distribution may depend on the specific crystal and optical interface geometries as well as the indices of refraction of all materials involved. In order to predict the incident angular distribution, the optical Monte Carlo code DETECT2000 was modified to output the entire trace, providing times and locations for each interaction (Figure 4.1). It should be noted that the optical tracking module within GEANT4 should work equally well for this type of optical tracking simulation.

comp	fin	x	y	z	age	nsurf	fate
1	0	0.87	-0.18	0.98	0.000	00	0
1	1	0.84	-0.17	1.00	0.000	01	0
1	1	1.00	-0.22	0.84	0.002	02	0
1	1	-1.00	-0.87	-0.69	0.017	03	0
1	2	-0.56	-0.88	-1.00	0.020	04	0
1	1	1.00	-0.93	0.10	0.032	05	0
1	1	-0.69	1.00	0.13	0.047	06	0
1	1	1.00	-0.13	0.34	0.060	07	0
1	1	0.89	-0.48	1.00	0.064	08	0
1	1	1.00	-0.78	0.87	0.066	09	0
1	1	-1.00	0.65	-0.43	0.083	10	0
1	2	-0.08	0.97	-1.00	0.090	11	0
1	1	0.01	1.00	-0.94	0.090	12	0
1	1	1.00	0.47	-0.15	0.099	13	0
1	2	0.52	0.54	-1.00	0.105	14	0
2	3	0.37	0.57	-1.20	0.106	15	1
1	0	0.53	-0.72	-0.06	0.000	00	0
1	1	1.00	-0.54	-0.31	0.003	01	0
1	1	-0.09	-1.00	-0.21	0.011	02	0
1	1	-0.01	0.67	1.00	0.023	03	0
1	2	0.14	0.47	-1.00	0.035	04	0
2	3	0.16	0.44	-1.20	0.036	05	1

Figure 4.1. Trace of two detected photons from DETECT2000 indicating: component id, reflector finish, position (cm), age (ns), number of surfaces, and detection fate.

The two photons traced in Figure 4.1 each begin at different locations and experience a number of interactions with both the diffusely reflecting sidewalls

(finish 1) of the 8 cm^3 crystal as well as the optical interface (finish 2) located at z = -1 cm. Depending on the crystal geometry and surface conditions, there can be a significant difference in the number of interactions between photons and thus a spread in the times of detection.

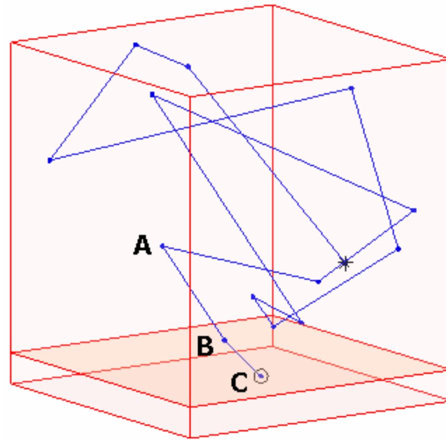


Figure 4.2. Geometry of cubic scintillator and bottom photodetector indicating one possible path of a single photon interacting at diffusely reflecting boundaries.

Figure 4.2 illustrates one possible path a scintillation photon may take in a diffusely reflecting cubic scintillation crystal. Ray AB is the angle incident upon the optical interface while ray BC is the ray incident upon the photodetector. If we take the index of refraction of the scintillator to be n_1 = 1.8, and the optical interface index to be n_2 = 1.45, then the critical angle for this boundary is $\theta_c = \sin^{-1}(1.45/1.8) = 53.7°$. A multi-photon simulation was performed wherein scintillation positions were uniformly distributed throughout the crystal.

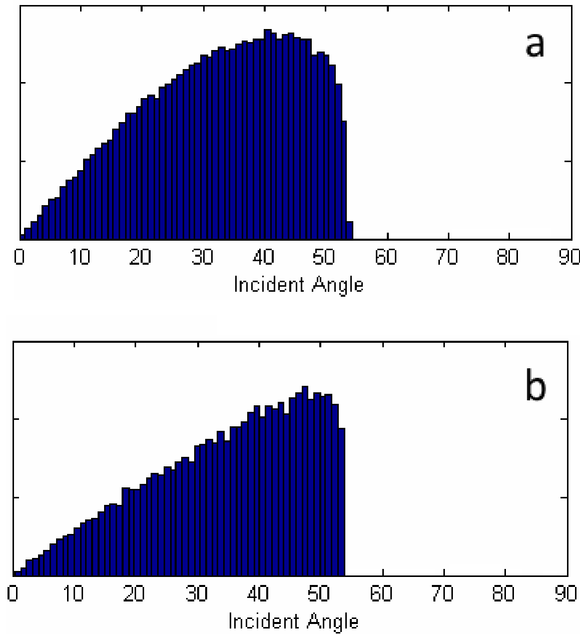


Figure 4.3. Angular distribution (ray AB from Figure 4.2) of photons refracted across the optical interface of a scintillation crystal covered by diffuse (a) and specular (b) reflectors.

Upon observing the angular distribution of photons that make it across the boundary (Figure 4.3), indeed we see a critical-angle-limited distribution, where the majority of photons exit the crystal at angles close to the critical angle. Although intuition is rarely useful in boundary-valued branching process problems, one might conclude that the more uniformly distributed set of angles in the diffuse case is a result of the wide distribution of possible angles at each reflection. As for the low probability of small incident angles, one should consider that the ensemble of angles are not confined to within a single plane and that the differential solid angle seen by the detector is $d\Omega = \sin(\theta)d\theta d\phi$. This is one explanation for the reasonably sinusoidal increase in probability for greater incident angles. Because the incident angles range from 0° to the critical angle, the refracted angles will be essentially stretched into a distribution from 0° to 90° (Figure 4.4).

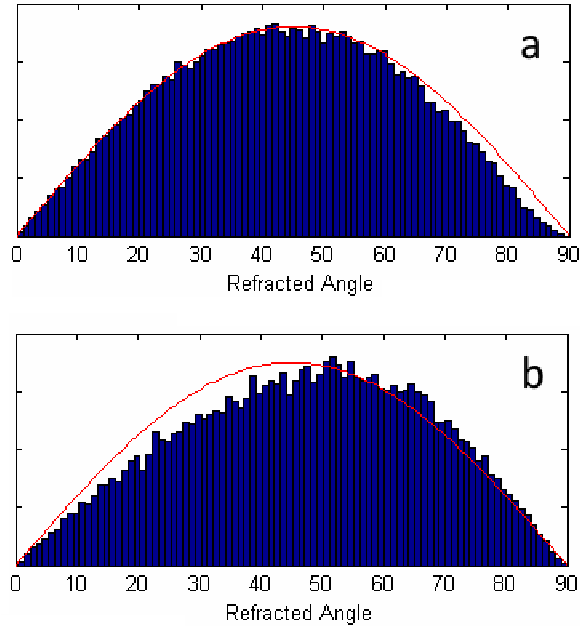


Figure 4.4. Angular distribution (ray BC from Figure 4.2) of photons refracted across the scintillator-optical interface boundary for diffuse (a) and specular (b) crystal reflectors.

It is these distributions which the photodetector sees and for which an ARC should be designed. The distributions from both specularly and diffusely wrapped cubic crystals appear to each be reasonably well approximated by a sin(2θ) distribution. One might expect these distributions to be slightly altered for non-cubic (e.g. cylindrical) or higher aspect ratio crystals. The distribution may also be altered by the location of scintillation within the crystal (e.g. low energy particles absorbed near a surface).

4.1.B. Thin-film Interference Filters

A destructive interference optical filter can be created from a single thin-film of appropriate optical thickness (index of refraction times film thickness). The three materials involved in the schematic of Figure 4.5 are the scintillator's optical coupling compound (n_1), the interference filter (n_2), and the silicon photodetector (n_3).

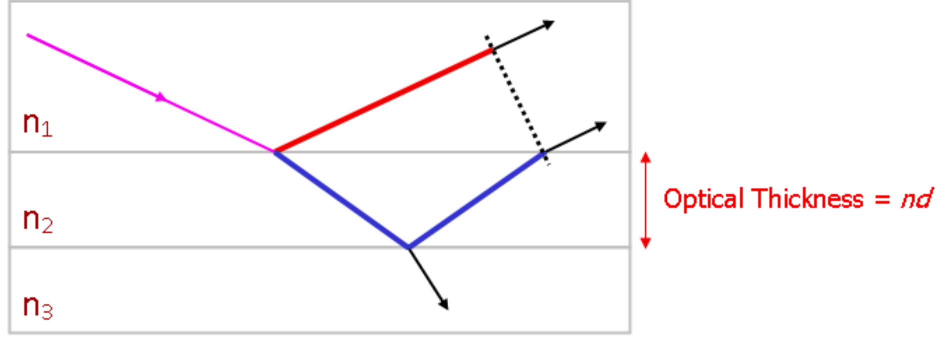


Figure 4.5. Destructive optical interference filter schematic where superposition occurs between refracted (blue) and reflected (red) rays.

The optical thickness (n_2d) is selected such that the angular phase difference (i.e., phase) between the reflected and refracted rays,

$$\beta = \frac{2\pi}{\lambda} n_2 d \cos\theta_2 \tag{3.50}$$

is equivalent to π radians or 180° phase difference at normal incidence ($\theta_2 = 0°$). In order for the reflected and refracted paths to destructively interfere, it is required that

$$n_2 d = \frac{\lambda}{4}. \tag{3.51}$$

This represents a quarter wave optical thickness. The single interface amplitude reflectances from interface *i* to *j* for both *s*- and *p*-polarizations are

$$\begin{aligned} r_{ijp} &= \frac{n_j \cos\theta_i - n_i \cos\theta_j}{n_j \cos\theta_i + n_i \cos\theta_j} \\ r_{ijs} &= \frac{n_i \cos\theta_i - n_j \cos\theta_j}{n_i \cos\theta_i + n_j \cos\theta_j} \end{aligned}, \tag{3.52}$$

while the reflectance accounting for both interfaces is

$$R_k = \frac{r_{12k}{}^2 + r_{23k}{}^2 + 2r_{12k}r_{23k}\cos(2\beta)}{1 + r_{12k}{}^2 r_{23k}{}^2 + 2r_{12k}r_{23k}\cos(2\beta)}, \tag{3.53}$$

and the total reflection coefficient is the average of *s*- and *p*-polarized photons

$$R = \frac{1}{2}\left(R_p + R_s\right). \tag{3.54}$$

Selecting some values for the indices of refraction in a typical system, we next examine the reflected energy from a single thin-film interference filter of varying optical thicknesses and at various incident angles (Figure 4.6).

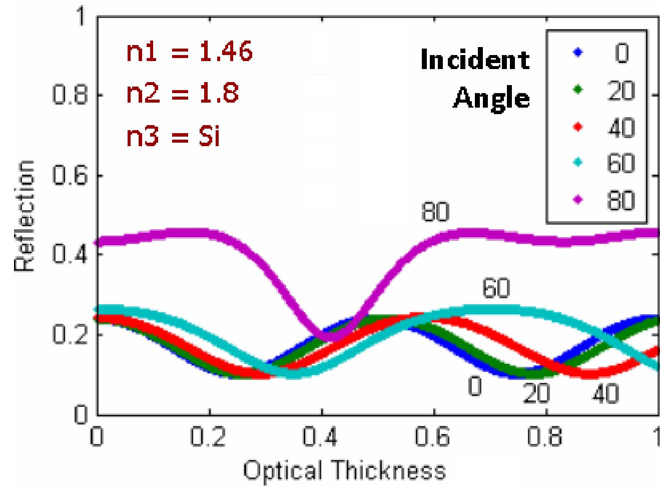


Figure 4.6. Single layer thin-film interference filter (n_2) as a function of *relative* optical thickness for varying incident angles (at 500 nm).

Clearly, the quarter-wave filter achieves the lowest reflectivity (highest transmission) at normal incidence. However, the optimal optical thickness shifts to 0.35 for angles of 60° and above 0.4 for angles of 80°. Note also the periodic behavior which indicates that optical thicknesses of 1/4 + *n*/2 wavelength (where *n* is a whole number), would function similarly.

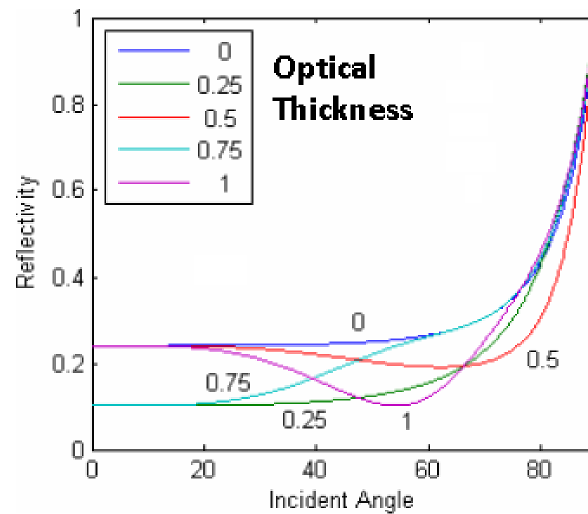


Figure 4.7. Single layer thin-film interference filter (n_2) as a function of incident angle for varying *relative* optical thicknesses (at 500 nm).

Taking a similar view (Figure 4.7), one can see the degree of improvement for various filters as a function of incident angle. It is interesting to note that at no angle does any filter do worse (absorb more) than a photodetector with no filter. There is however some variation in the minimum reflected distribution, as was shown in the previous figure.

By combining this reflectivity versus optical thickness and incident angle (Section 4.1.A), the scintillator-specific reflectivity can be illustrated (Figure 4.8).

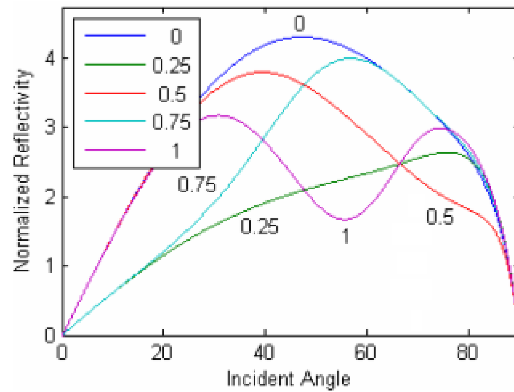


Figure 4.8. Reflectivity for 500 nm scintillation photons exiting a diffusely wrapped cubic crystal (Figure 4.4) and crossing an interference filter (n = 1.8) of specified *relative* optical thickness, as a function of incident angle.

The effects of the sinusoidally distributed incident angles are observed to pull down the reflectivities (Figure 4.8) at higher angles of incidence. By visually integrating the curves in Figure 4.8, one may ascertain that a quarter-wave optical filter may indeed be the most appropriate choice for this single-wavelength case. The numeric integrations presented in Figure 4.9 confirm this hypothesis.

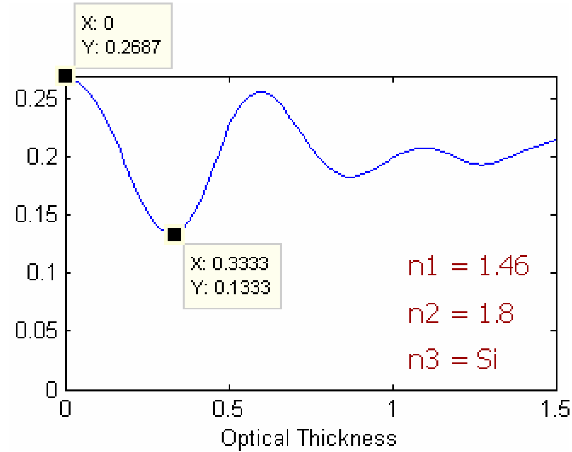


Figure 4.9. Integrated reflectivities from Figure 4.8 as a function of *relative* optical thickness with a minimum at 0.33.

We note first that without an ARC, only 73% of the 500 nm scintillation photons will make it into the silicon. However, with an optimal thickness filter of 0.33 * 500 nm = 165 nm, the transmission is improved to 87%. This represents a 19% improvement with an appropriate single thin-film interference filter. In terms of applied results, these additionally detected photons could lead to a $\sqrt{1.19} \rightarrow 9\%$ improvement in the measured energy resolution. It should be noted that a *traditional* quarter wavelength filter does perform similarly, for this wavelength, at these refractive indices, and with this incident angular distribution. It is also interesting to note that as the optical thickness increases beyond unity, that the reflectivity tends to damp towards a stable, still improved value of ~20% from 27%. The next step in determining the optimal filter thickness would be to include the distribution of wavelengths emitted from a given scintillator. It is assumed for the purposes of this work that the relative spectral width of the scintillator is sufficiently narrow such that the peak wavelength can be representative of the entire distribution.

Sensitive optical photodetectors are employed in many fields for a variety of applications, the optimization of optical transmission into these photodetectors often can be greatly improved with the application of a suitable antireflection coating. While many detectors operate from a normally-incident source of photons which travel through the air, scintillators for nuclear detection represent a closed optical system in which the light source is intimately coupled to the

photodetector. In addition the isotropic scintillation emission process provides a wide distribution of incident angles to the surface of the photodetector. Optical Monte Carlo simulations were performed to illustrate the expected angular distributions, and the design process for an ideal single layer thin-film interference filter was presented. A traditional quarter wavelength optical filter is predicted to perform adequately provided it is tuned to the peak scintillator emission wavelength. Moreover, the presence of any single layer thin-film optical filter is determined to transmit a greater number of photons than would the absence of such a filter.

4.2. External Optical Crosstalk

The traditional description of optical crosstalk involves a hot-carrier emission being ejected laterally from a pixel and being absorbed in the active region of a neighboring pixel. There are however other paths (Figure 4.10) these photons may take to end up in the active region of another pixel.

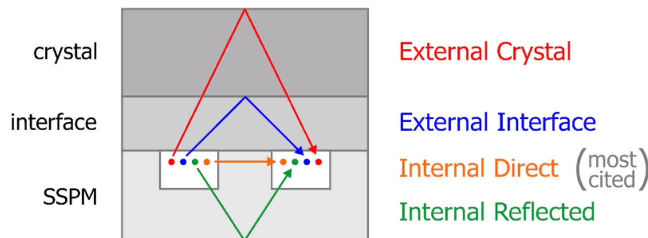


Figure 4.10. Optical crosstalk mechanisms in a closed optical detection system (e.g. scintillator).

The substrate of a SiPM is typically of standard wafer thickness, ~0.5 mm. Depending on the base material upon which it is mounted, there is some probability that higher wavelength hot-carrier emissions will be internally reflected off the bottom surface and be detected in another pixel. Unlike the internally direct method of optical crosstalk, this reflected crosstalk mechanism is not limited to regions immediately surrounding the originating pixel.

Two other forms of optical crosstalk are possible as a direct result of placing a scintillator atop the silicon photomultiplier. These mechanisms involve hot-carrier emissions being emitted away from the surface and interacting with either the optical interface or the highly reflective scintillator surfaces. These forms of

external optical crosstalk also provide a relatively distributed spatial probability in terms of the secondary pixel location. In fact, simulations show [Bar09] that for a cubic crystal with specular reflector, the location of hot-carrier photon absorption is distributed uniformly over the majority of the device. The degree of impact these mechanisms have will be a strong function of the excess bias, as more hot-carrier emissions are produced when the total avalanche charge is increased. The hot-carrier emission flux will also depend strongly on the fabrication process and has been known to vary by orders of magnitude from manufacturer to manufacturer [Bar09].

Chapter 5

Conclusions and Future Work

5.1. Silicon Avalanche Diode Fabrication

Methods have been described for the creation of high electric fields in silicon diodes which amplify photoelectrons into an avalanche of charge via impact ionization. Specifically for scintillation detection, a photodetector which is sensitive to more blue/UV photons necessitates photoelectron amplification within the first several microns of the surface. This high electric field region will be nonuniform without careful design and thus represents a nonuniform probability of detection over the majority of the avalanche diode.

Specific structures were explored to prevent premature edge breakdown. This thesis uniquely relied upon furnace doping and diffusion instead of implantation for the creation of diode structures. The reason for this decision was in the supposed damage and incomplete dopant activation that potentially accompanies ion implantation. Preliminary gettering methods were explored to reduce the reverse bias leakage current, and success was obtained with a low-stress polysilicon backside layer, highly doped with phosphorus. Various wafers and doping levels were used to illustrate the relative efficacy of the gettering process developed herein. Current-voltage characteristic curves were obtained from wafer-level probing, and images of the hot-carrier emission locations (i.e., reverse bias electroluminescence) were obtained with a sensitive CMOS camera mounted to a probe station microscope.

Junction termination extension diode structures were observed to reduce edge breakdown, given a minimum extension value. Several contact geometries and

schemes were explored to identify suitable designs for scalability to close-packed arrays of diodes. A balance of low reverse bias leakage current, suitably sharp and high breakdown voltage, and uniform hot-carrier emission pattern was unable to be identified from within the diodes fabricated. Unwanted excess charge was injected from peripheral areas due to lack of diode isolation. Unsuitably high doping levels led to lower breakdown voltages, and an increased probability of tunneling current. This particular structure may perhaps best be suited for epitaxial substrates and ion implantation.

A bevel diode etching process was specifically developed to yield an extremely smooth and shallow etch with good conformity to a low-temperature oxide mask. This was achieved by first densifying an LPCVD oxide mask at 600 °C, and subsequently etching in a 60 °C 25% TMAH solution including 0.1% per volume Triton-X 100 surfactant. Bevel diodes were fabricated with this process and yielded reasonably low reverse bias leakage currents of 10 pA when combined with the JTE structure.

Guard ring structures were designed to prevent leakage current from the substrate and produced the lowest leakage currents of <100 pA for 20 μm diodes. Greater breakdown voltages were also obtained which reduces the relative contribution of breakdown by band-to-band tunneling. Several issues with the doping order prevented laterally contacted diodes from breaking down uniformly, but Geiger-mode behavior was ultimately observed for various quench resistances. A dark count rate of 300 kcps was obtained for a guard ring diode at 0.2 V excess bias.

A process was developed to deposit and pattern transparent conducting thin-film quench resistors from sputtered ITO in an oxygen rich ambient. Films of 10 nm thickness achieved a sheet resistivity as large as 12.4 kΩ/square. This represents a suitable resistivity to achieve the tens to hundreds of kΩ typical for diodes of 10 to 100 μm in size. Thicker layers of stoichiometric ITO can also be used for low resistance readout lines with a sheet resistance of 8 Ω/square for a 500 nm film.

5.2. Silicon Photomultiplier Statistics

In addition to specific fabrication efforts, several key statistical models were developed which can be used to estimate the number of incident photons from the number of pixels fired. The probability that a given pixel will fire was derived from first principles in terms of physical quantities such as the geometric efficiency and the avalanche initiation probability. The probability that multiple pixels will fire was then derived and shown to be the source of silicon photomultiplier nonlinearity at higher photon fluxes.

Methods from information and estimation theory were lightly applied to obtain a measure of the variance in both the number of pixels that fire and the estimate of the number of incident photons for a Poissonian incident photon flux. Mention was made to spatially non-uniform photon fluxes which increase the local flux density and may require special treatment to estimate the true incident photon flux.

Other more systematic factors such as the dynamic range, avalanche excess noise, and quantization errors were each derived as unique attributes of the silicon photomultiplier arising from the nature of its discretized detection surface. The recovery time of a diode was predicted to play a complex role in the number of photons detected at high photon fluence. The choice of integration time constant and the dark count rate will each play a key role in the selection of an optimal recovery time (and thus pixel size, quench resistance, etc.). The physical mechanisms of thermally-generated dark counts, afterpulsing and optical crosstalk were each introduced, but a full second-order statistical treatment of their combination is best left to specific cases to prevent ill-drawn conclusions.

An improvement in photon detection efficiency (and thus counting statistics) was described through the development of an ideal antireflection coating specific to detection of scintillation photons. It was determined through simulation that the angular distribution of photons exiting a cubic scintillator followed a $\sin(2\theta)$ distribution centered at 45°. The suggestion put forth in this work is to design an

antireflection coating targeted for 45° incidence and for the centroid wavelength of the scintillation emission wavelength distribution.

5.3. Future Work

Many aspects of detector fabrication depend on well characterized and maintained tools and processes. Some of the more influential of these tests and processes are detailed below as suggestions for endeavors which would improve the reliability of processing, especially in a shared facility.

5.3.A. Commercial Foundry Run

While ion implantation may be damaging to the surface of an avalanche diode, there may be ways to combine the relative benefits of both ion implantation and furnace diffusion. For instance, a controlled dose of dopants may be implanted into a layer (e.g. oxide or polysilicon) above the silicon without doing damage to the underlying lattice [Sci03]. The dopants may then be driven into the silicon in such a way that near complete activation is achieved.

5.3.B. Passivation

It was unclear from the work performed to date exactly what effect various passivation techniques had on surface leakage. A hydrogenated, thermally grown silicon dioxide is known to perform very well [Sze07], however little is known as to the required thickness or composition. Many plastics, organics, and deposited materials (e.g. polyimide) have also been reported to perform as well as thermal oxides for some applications. A comprehensive test of basic structures would be prudent to elucidate viable passivation techniques.

5.3.C. Reliable Ohmic Contacts

The optimization of ohmic contacts to shallow silicon junctions requires significant trial and error testing. There is some evidence (as seen in the contact glow in this thesis) that the sharp vertical edges of metal contacts may provide the potential for breakdown. Following historical trends of the CMOS industry,

optimization of an appropriate silicidation (e.g. $TiSi_2$) process would perhaps be the next most likely step towards reliable contact integration.

5.3.D. Diffusion Masking

There is evidence that low pressure chemical vapor deposited (LPCVD) low temperature oxide (LTO) functions differently as a diffusion mask than thermally grown oxide masks. This is due to the more porous nature of the LTO. However, there is also evidence that LTO can be *densified* at higher temperatures. The degree of densification will certainly depend on the specific LTO deposition and growth characteristics as well as the densification ambient and temperature. Furnace diffusions require a slow ramp to diffusion temperatures, and thus a built in densification may occur. In the case of gaseous diffusion, the substrate can be held at diffusion temperatures before any gas is introduced. For solid proximity diffusion systems, the dopant is introduced constantly (and to varying degrees) at some lower threshold temperature. Therefore, furnace diffusions contain a built-in densification step. The question remains as to what times and temperatures are required to properly densify the LTO. Another question is what thickness of LTO is required to adequately mask a diffusion at a given time and temperature? That is, what is the diffusion rate in a given LTO oxide?

5.3.E. VMOS

Methods have been developed for V-groove MOS (VMOS) transistors employing bevel etching and thin gate oxides [Rog74]. Given the relative maturity of the bevel etching processes developed herein, the co-integration of VMOS transistor logic with existing bevel-edge terminated diodes may be feasible.

5.3.F. Variation of Lateral Doping

One additional method of creating a guard ring is to use a method called variation of lateral doping to continually reduce the doping at diffused edges

[Ste85]. An oxide implantation or diffusion mask is created with decreasing pinholes around the main diode in order to control the lateral doping.

5.3.G. LOCOS Isolation

Local oxidation of silicon (LOCOS) is one method of forming isolation regions between transistors in a CMOS process. It has been replaced by higher density methods such as deep and shallow trench isolation formed by plasma reactive ion etching. The local oxidation process creates a bevel which may be appropriate for controlling the electric field at the boundaries of an avalanche diode. However, this method may consume too much real estate to be a viable method for *very* closely packed pixels in a SiPM.

5.3.H. Active Quenching and Micro-Optics

Active quenching is no doubt a more elegant solution but requires significantly more development effort and expense as well as consuming more silicon real estate. One way around this loss of geometric efficiency is to integrate micro-focusing optics. The drawback with this approach is that scintillation photons are multidirectional by nature, and may therefore be more difficult to focus.

Appendices

Appendix A
Process and Device Simulation

The realization of optimal semiconductor structures is greatly aided by technological computer aided design (TCAD) simulation software. With proper input parameters, these simulation tools allow the designer to iteratively test processes and devices without the need for expensive refabrication at every step. The primary purpose of semiconductor *process* simulators is to estimate the charge density distribution by simulating the interaction of specific impurities with the semiconductor lattice. The role of device simulators is to take a known charge distribution and determine the resulting electrostatics and electrodynamics under varying applied biases. Many process effects may be included, such as stress, oxidation, ion implantation and the interaction of multiple dopant species with lattice interstitials and vacancies. Due to the inherent variability in process tool configurations, simulations must be calibrated with measurements from real test runs.

For instance, a four inch wafer diffusion furnace which operates in the 600 to 1200 °C range may act very differently depending on the differences in gas flow dynamics, resulting in inexplicable variation in doping from die to die, or from wafer to wafer, or from run to run, or from lab to lab. The thermocouples which provide constant control feedback may be tuned slightly differently which may create hot spots and locally enhanced diffusion. Environmental effects such as the relative humidity and temperature of the lab, or how long the tool has been powered, can also add variability to processing. Unless two *seemingly* identical tools, from lab to lab, are engineered to be identical, there is significant room for doubt as to the congruency of process outcomes. For this reason, process modeling is more of an art than a science, wherein the more calibration data that

is collected, the greater degree of trust the designer can place in the process simulations.

The process modeling software used throughout this thesis is SUPREM, the Stanford University Process Engineering Module, which evolved from a one-dimensional silicon diffusion modeling program [Ant78]. Since its inception three decades ago, SUPREM has undergone continual development and is now a standard tool for process simulation in silicon. The version of SUPREM used in this thesis is TSUPREM4 v2007.03 from Synopsis, Inc.

Appendix B
Process Flows

Process development requires a significant number of iterations of trial and analysis, a fact which is difficult to plan for unless "skilled in the art." Textbooks on fabrication can only scratch the surface as to the expected variance in the outcome of a single processing step. The number of variables simply is too large comprehensive factorial testing, so significant efforts are made to standardize and identify consistent and measurable behaviors and sub-processes which effectively reduce the mass of process parameters. The following process methods were developed and tested to provide that performance repeatability.

B.1. Experimental design

The following experimental design details the individual steps taken to produce the 16 wafers referenced within this work.

			1	1	2	1	Oxide		Gettering (poly)				Getter (n)		LOCOS	Boron (p)				Boron(p+)		Phos (n+)				Bevel		
	Poly		N	P	N-	P-	OX	TCA	**G0**	GB	GF	GFB	GN	GD	LOC	B1	B2	BS	BI	B+	BS+	P1	P2	PS	PI	V1	V2	VA
G1	x	**normal**		**x**				**x**				**x**	**x**			**x**				**x**		**x**				**x**		
G2	x	deep p		x				x				x	x				x			x		x				x		
G3	x	deeper n+		x				x				x	x			x				x			x			x		
G4	x	deep get		x				x				x		x		x				x		x				x		
G5	x	back get, deep p,n+		x				x		x			x				x			x			x			x		
G6		noget		x				x	x							x				x		x				x		
G7		front normal		x				x			x		x			x				x		x				x		
G8		front deep		x				x			x			x		x				x		x				x		
G9	x	**normal**			**x**			**x**				**x**	**x**			**x**				**x**		**x**				**x**		
G10	x	deep p			x			x				x	x				x			x		x				x		
G11	x	deeper n+			x			x				x	x			x				x			x			x		
G12	x	deep get			x			x				x		x		x				x		x				x		
G13	x	back get, deep p,n+			x			x		x			x				x			x			x			x		
G14		noget			x			x	x							x				x		x				x		
G15		front normal			x			x			x		x			x				x		x				x		
G16		front deep			x			x			x			x		x				x		x				x		
													900,10			875,10						850,10						
													900	3hr		1000,6hr				900,5			900,5					900,30
													1100	3hr			1100,6hr											

B.2. Lithography

When compared to a commercial CMOS process, the lithographic methods demands for single photon avalanche photodiodes are meager, requiring relatively few masks, feature sizes larger than 3 µm, and alignment tolerances less than 1 µm. These features are readily achieved with the 15+ year old technology found in the majority of university fabrication laboratories. Equipment trickles down from industry, and universities are currently moving from 4” to 6” capable tools due to the decline in serviceability of older tools. A greater degree of process cleanliness and control can be expected due to the inherently more stringent demands placed on newer tools.

B.2.A. Masks

Soda lime glass (5” × 5” × 0.090”) masks are purchased pre-coated with 1.0 µm of AZ1518 resist from Nanofilm in Westlake Village, CA. Mask designs are drawn in L-Edit or LayoutEditor and the resulting .GDS files are fractures into rectangles by either PDRACULA or ASM2600 GDS2PG software. An Interserve Electromask II mask maker takes the fractured .INT files and exposes ~700 ms 500 W H-line bursts of light through a 2 µm minimum aperture on a 0.1 µm grid. Typical flash rates are 3000 flashes per hour. Critical (i.e., small) features require an exposure test to first be performed. Identical rows of varying exposure can be used to simultaneously perform an exposure-develop matrix test. Masks are developed for 55 s in Microposit MF-319 developer, and inspected under a filtered-light microscope for opening of small features. A 90 s etch in Cyantek Cr-14 removes the chrome underlying the developed photoresist. After another inspection, the photoresist is stripped in a “metals-allowed” heated bath of J. T. Baker PRS-2000. Prior to use, masks are cleaned of any photoresist residue in a large bath of acetone (J.T. Baker CMOS™), while being gently scrubbed with a smooth swab. Without letting the acetone dry, the mask is quickly transferred to a primary bath of isopropyl alcohol or IPA (J.T. Baker CMOS™). An optional final rinse in 18 MΩ-cm deionized water removes any additional residue from the shared baths. Masks are removed and blown dry on all sides with clean nitrogen and placed into a mask box or carrier for future use.

Features smaller than 2 μm can be created by using the image repeat feature on the Electromask or by using the GCA Autostep 200. Each method employs a complex 5:1 reducing lens to either create a reduced mask or directly expose a reduced pattern onto the wafer. This provides the capability for 400 nm features, provided the selected photoresist is capable of sub-micron features. Special alignment marks are needed for each system in order to align the wafer to the system.

B.2.B. Photoresist Application

Manual application of photoresist requires a clean, dry substrate. The first step in photoresist application is a 7 min cycle in a spin-rinse-dry unit which removes any gross surface contamination. Even after wafers appear dry, microscopic water molecules may persist. The wafers are then dehydrated with a longer oven bake at 200 °C for 20 min or a shorter hotplate bake of 90 s at 115 °C. wafers are transferred to a vacuum chuck which holds the wafer and spins at speeds up to 5000 rpm. The adhesion of many photoresists is significantly improved with the application of hexamethyldisilazane (HMDS). Enough drops are applied from a plastic pipette to cover the wafer when spun. A slower 500 rpm cycle runs for 5 s in order to spread the HMDS or photoresist, and is followed by a 30 s spin at a faster speed (up to 5000 rpm). Another method for HMDS application is in a heated vacuum chamber, where the HMDS is introduced as a vapor. This largely removes the need for the dehydration bake and yields very repeatable results. Enough photoresist is applied to the wafer in order to just cover the surface at higher spin speeds. Many photoresists are defined by a 4 digit label such as 1827 or 1518, where the last two numbers indicate the photoresist thickness at 5000 rpm; e.g. 2.7 μm of 1827 at 5000 rpm. The primary resist used in this work is Megaposit SPR-220 3.0. Very thin photoresists would be desirable for very high resolution processes. Once the photoresist has been spun, it is placed on a 115 °C hotplate to remove the solvents in what is called a *softbake*. Instead of handling wafers directly, a wafer track system (Suss ACS 200) employs a robotic arm which can vapor prime, spin, bake, and develop an entire cassette of wafers

without manual intervention. This system provides extremely good uniformity and repeatability.

B.2.C. Alignment and Exposure

The mask pattern is transferred onto the photoresist in an aligner. A Suss G-line contact aligner places the mask in near contact with the wafer, and alignment is performed by microscopically looking through the mask on the wafer and moving the wafer into correct alignment. Exposure times for SPR-220 at 3 μm is 6 s. An I-line projection aligner (GCS AS200) automatically aligns the wafer to the aligner system. The wafer is then *manually* aligned to the aligner system, and the mask reticle is reduced by 5x in order to create each die on the wafer at exposure times of 0.35 s. Vernier alignment marks on adjacent layers allow sub-micron alignment to be microscopically verified, with 0.5 μm being typical over the entire wafer.

B.2.D. Development

An optional post-exposure bake (PEB) can be performed for 90 s at 110 °C and may help to reduce standing wave effects which can cause scallops or ripples in photoresist sidewalls. An AZ 300-MIF (metal ion free) developer is used either in a breaker with manual agitation for 45-60 s, or in the ACS200 where it is directly sprayed for 30 s. Spray developing helps maintain even developing of smaller features. An optional hardbake, identical to the PEB can be performed to increase the resistance of photoresist to subsequent chemical etches. Wafers are inspected for opening of small features, where poor exposure or developing can be observed by the difference in color relating to the interference filter effect of the semitransparent photoresist. Specific exposure-development times can be obtained again by performing a matrix test and selecting a safer operating point.

B.2.E. Descum

Depending on the process, a very shallow layer of photoresist may remain adhered to the surface of developed areas. These perhaps monolayers can effectively block the subsequent etching processes, so a method is needed to

remove only a small amount of photoresist. This is accomplished with a low power 50 W oxygen RF plasma (March Asher) at a pressure of 250 mTorr and gas flow of 17% (17 sccm) for 30 s. Without this step, very small features can have the tendency to be left completely untouched by the subsequent wet etch steps.

B.2.F. Etching

The interactions between many etchants and materials has been systematically studied [Wil96, Wil03] and should be consulted before any chemical process. The primary use for photoresist in this work is in the patterning of silicon dioxide; either thermally grown dry and wet oxides, or silane deposited low temperature oxides (LTO). Hydrofluoric (HF) acid is the primary etchant for oxides, since it does not etch silicon or organic materials like photoresist. The exhaustion of HF occurs rapidly with etching, so a buffering agent (typically ammonium fluoride) is added to stabilize the relative pH value. Typically etch rates are ~800 Å/min, however oxide quality and chemical storage, lifetime, use can all affect the absolute etch rate by 50% or more. For this reason, it is best to replace any chemical directly before use. Bare and patterned monitor wafers are used to test the actual etch rate just prior to etching process wafers. Oxide thicknesses are measured with a Nanometrics Nanospec 6100 film characterization system. A 10 μm minimum spot of white light is reflected off the surface of a wafer and is spectrally analyzed. The unique signature of the resulting interference pattern is then used to estimate the thickness, given appropriate model selection.

B.2.G. Lift-Off

While some metal layers are able to be selectively etched with suitable etchants, there are times when these etchants would attack the underlying layers. For “gentle” processing, a lift-off process can be used, wherein metal deposited over photoresist is removed in a photoresist etchant. The challenge is to prevent the readhesion of metal films once floating in solvent. This can be achieved with the addition of surfactants or use of photoresist removers like Microposit 1112A. A unique single layer photoresist solution is achieved by hardening the top of a

layer of photoresist with a pre-exposure develop. A 20 s spray with AZ 300 MIF is applied to a non-softbaked wafer just prior to exposure. This step provides a suitable photoresist step profile for lift-off to be successful for even conformal coating as from sputtering.

After metal or thin-film layers are sputtered or evaporated, the wafers are placed faced down in a plastic cassette inside a beaker of solvent: acetone, PRS-2000, or 1112A. Additional heating on a hotplate can accelerate the liftoff process, so long as no boiling or rapid fluid flow occurs. The beaker should remain as undisturbed as possible until the majority of the layer is lifted off. The bulk of the liftoff metal should be carefully filtered out, while keeping the wafers wetted with solvent. An ultrasonic bath is useful in removing any stubborn pieces.

B.2.H. Removal

All photoresist is removed for 10 min in a heated bath of PRS-2000 specifically set aside for "no metals" processing. A 2 min DI rinse and SRD follow each photoresist strip. Wafers are finally inspected to ensure no residue remains.

B.3. Diffusion Masking

One of the key benefits of silicon over other potential semiconductors is the high quality silicon dioxide (known simply as oxide) that can be easily grown in an oxygenated atmosphere at temperatures above ~800 °C. This *thermal* oxide is very efficient at tying off dangling Si bonds at the surface, and is one of the best methods for electrical passivation. Thermal oxides of sufficient thickness also act as very good barriers to diffusion of common semiconductor dopants, which makes patterned doping straightforward. Patterning of silicon dioxide is achieved predominantly through photolithography and hydrofluoric (HF) acid wet etching techniques.

Due to the precise temperature budget of the potentially four diffusions in the processes developed herein, it is unreasonable to grow a thermal oxide mask for each diffusion. However, given the ultimate furnace doping conditions of 800 °C

for 10 min, reasonable thicknesses of oxide might be able to be grown. The question remains what combination of dry and wet oxides would effectively shield such diffusions. SUPREM modeling can give relative hints, but should not be trusted for absolute values. Instead, a spreading resistance profile should be taken for successively decreasing mask thicknesses until the appropriate limit has been achieved.

To avoid this task, a low temperature oxide (400 °C) is deposited via low pressure chemical vapor deposition (LPCVD). Silane gas dissociates and combines with oxygen to form silicon dioxide of higher quality than is found in plasma enhanced CVD (PECVD) films. There is an added benefit just prior to the diffusion process

B.4. Furnace Diffusion

Diffusion is achieved through a two-step process in which a solid-solubility concentration of dopants is introduced to a shallow depth. This controlled dose is then driven in to greater depths and lesser concentrations. Boron diffuses out of the bare silicon surface at elevated temperatures (above 600 °C), while phosphorus piles up just below the surface.

Phosphorus dopants are introduced by bubbling nitrogen through liquid $POCl_3$ along with additional oxygen. The phosphorus and oxygen combine on the surface, and the oxygen also combines with the silicon to grow a shallow thermal oxide. As the phosphorus diffuses through the P_2O_5 into the silicon, the growing dry oxide helps to prevent phosphorus silicides and precipitates from forming which would be difficult to remove later. The dry oxide acts to undercut the phosphorosilicate glass (PSG) and aids in removal in common BHF or HF etchants. Diffusion profiles were verified by spreading resistance profiling to be

Boron diffusion is performed with solid B_2O_3 glass sources (Tecneglas BoronPlus) which are designed to emit a maximum dopant concentration at 1150 °C. This high temperature is rather inappropriate for the more shallow junctions in this work. The sources were originally selected when the tool

Thermco TMX-9000 furnaces were first installed for deep (15 μm) boron diffusion at solid solubility.

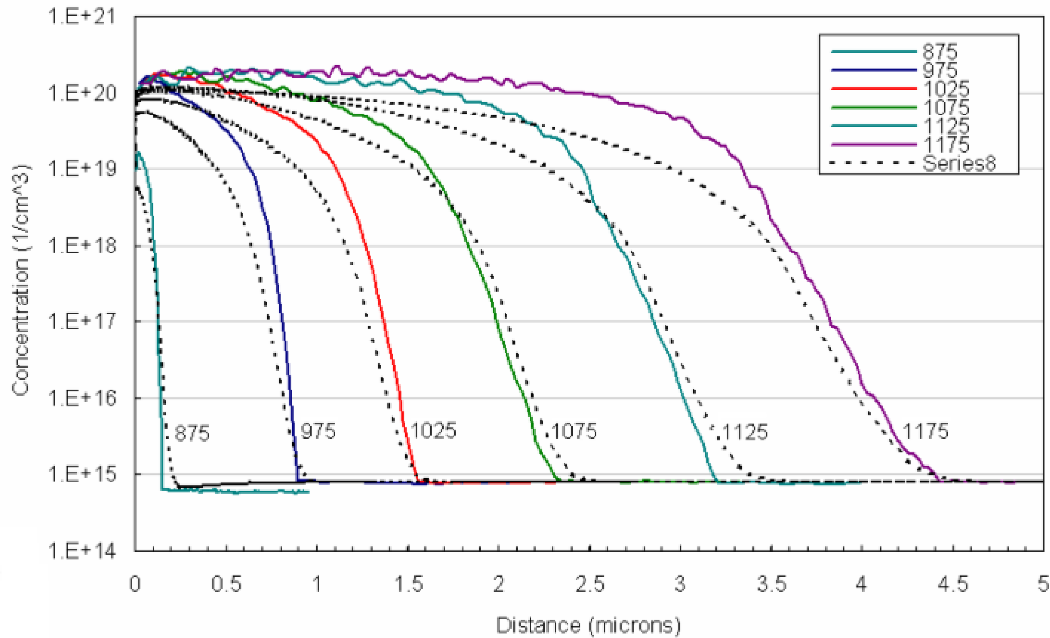
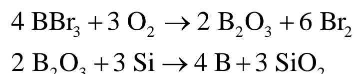


Figure B.1. Spreading resistance profiles performed by Solecon Labs for 10 min solid source diffusions at temperatures from 875-1175 °C. Dashed lines indicate best global fit with SUPREM simulations.

The sources were verified, via spreading resistance profiling (Solecon Labs), to emit boron at lower temperatures (Figure B.1), however the source manufacturer was unable to comment on the performance at temperatures below the original specification. An attempt was made to simultaneously fit all profiles with a set of diffusion parameters in the process simulator TSUPREM4.

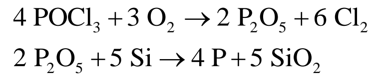
B.4.A. Tempress Diffusion Furnaces

While the majority of results in this thesis were obtained with the now-legacy Thermco furnaces previously mentioned, one run was performed with newer technology. In October of 2010, a new bank of ~20 Tempress furnaces was installed within the LNF which included tubes for liquid $POCl_3$ and BBr_3 diffusion sources. The bromine acts similarly to the chlorine to getter metallic impurities during diffusion. The chemical reactions during BBr_3 deposition are



$$4\ BBr_3 + 3\ O_2 \rightarrow 2\ B_2O_3 + 6\ Br_2$$
$$2\ B_2O_3 + 3\ Si \rightarrow 4\ B + 3\ SiO_2$$

while the reactions for $POCl_3$ are



$$\begin{aligned} &4\ POCl_3 + 3\ O_2 \rightarrow 2\ P_2O_5 + 6\ Cl_2 \\ &2\ P_2O_5 + 5\ Si \rightarrow 4\ P + 5\ SiO_2 \end{aligned}.$$

The sources provided by Air Products, Inc. are extremely pure at >99.99998% based on metals analyzed (the highest being <15 ppb Fe). This compares extremely favorably to the existing solid diffusion sources which contain metallic impurities at levels 100 to 1,000 times greater. Just as with the previous furnaces, a set of diffusion profiles had to be obtained for several temperatures of interest.

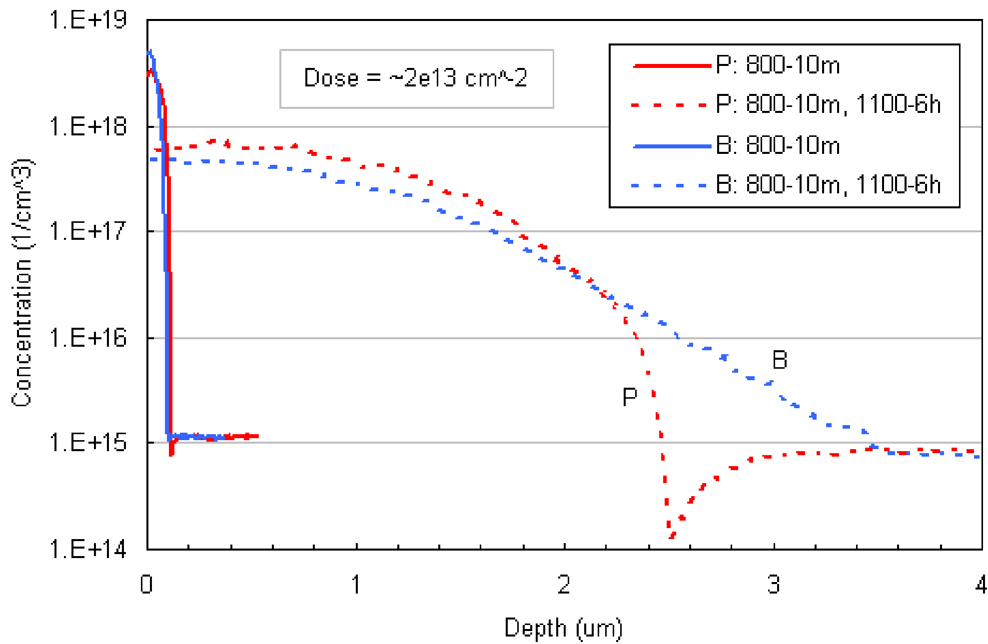


Figure B.2. Spreading resistance profiles for liquid bubbled phosphorus and boron sources predeposited at 800 °C for 10 min. Dashed profiles are the result of an 1100 °C drive in for 6 hr.

An initial set of profiles was obtained for 900 °C diffusions, however the concentration was determined to be close to or higher than the solid solubility. This indicated the possibility of incomplete dopant activation and thus a higher density of generation sites. A lower diffusion temperature of 800 °C was attempted, with surface concentrations an order of magnitude lower (Figure B.2) and an estimated dose of 2×10^{13} cm^{-2}.

B.5. Sputtered Contacts

The process for creating a Ti / TiN / Ti / TiN / Ti / Al:Si(2%) film stack begins with a lift-off recipe photoresist. Just prior to sputtering, wafers are held in DI water, then dipped for 3 s in BHF followed immediately by DI water and a SRD cycle. Wafers are immediately transferred to an Enerjet sputter tool and loaded face down. The goal is to minimize the amount of native oxide growth on the wafers before metal deposition. Sputtering is preferred over electron beam evaporation as the argon plasma has a cleaning effect on both the source and the wafer. The vacuum chamber is sealed and pumped to a base pressure of less than 10 μTorr. A flow of argon is applied until 7 mTorr pressure is achieved. Titanium is sputtered at room temperature from target location #1 at 500 W DC power. The exact voltage and current should not be compared outside this tool due to the inherent differences in target and plasma geometries between tools. Titanium nitride deposition uses the same titanium source, however 10-15% of the argon is replaced by nitrogen [Kaw96] in order to achieve a stoichiometric TiN film upon annealing. The final silicon-doped aluminum layer is deposited with a 3A DC current from target location #3. The deposition times for the entire film stack are: 4 / 2.5 / 0.5 / 4.5 / 4 / 25 min.

Annealing of contacts is necessary in order to improve the rough as-deposited microstructured interface between metal and silicon. In this process, any native silicon oxide is also taken up by the metal, thus improving the ohmic charge transport characteristics. Wafers are cleaned in acetone, IPA, DI, and spin-rinse-dried just prior to loading in a 400 °C furnace. An ambient of 2.7 SLPM N_2 and 300 SCCM H_2 provides a 10% hydrogen "forming gas" which is known to terminate bonds at the silicon-oxide interfaces.

B.6. Pre-Furnace Clean

The pre-furnace clean (PFC) adopted by the Lurie Nanofabrication Facility at the University of Michigan has been adapted rather directly from the original RCA clean [Ker70]. Wafers are cleaned of photoresist residue in a heated tank of PRS-2000, followed by spin rinse drying. A 6:1:1 mix of deionized (DI) water to hydrogen peroxide to ammonium hydroxide is prepared to 80 °C. Wafers are

introduced for 10 minutes and subsequently rinsed in cold flowing DI water for 5 minutes. The chemical oxidation built up on the surface during the first “organic clean” step is removed for ~30 s in either a 10:1 or 100:1 solution of DI to hydrofluoric acid. Another 5 minute DI rinse precedes a 6:1:1 DI to hydrogen peroxide to hydrochloric acid bath at 80 °C for 10 minutes. Another 5 minute rinse is followed by a 10 minute rinse in a faster N_2-bubbled rinse tank. Resistivity of the DI water is observed to achieve a given level before proceeding to a dedicated PFC spin-rinse-dryer unit, which performs several rinse cycles, followed by successive spin-drying cycles assisted by hot nitrogen. Wafers are then typically transferred to furnaces for immediate loading with dedicated quartz wands.

References

[Aki98] N. Akil, S. E. Kerns, D. V. Kerns, et. al., “Photon generation by silicon diodes in avalanche breakdown,” *Appl. Phys. Lett.*, 73-7 (1998) 871.

[Ant78] Antonaidis, D. A., et. al., *Stanford Univ. Elect. Lab. Tech. Rpt.* SEL 78-020, June 1978.

[Bal08] J. Baliga. *Fundamentals of Power Semiconductor Devices*. New York: Springer Science, 2008.

[Bar09] P. Barton, C. Stapels, E. Johnson, et. al., “Effect of SSPM surface coating on light collection efficiency and optical crosstalk for scintillation detection,” *Nucl. Inst. Meth. A* 610-1(2009) 393.

[Bas98] S. Bashar, “Study of ITO for Novel Optoelectronic Devices,” (Doctoral Dissertation) University of London, 1998.

[Bat60] R. Batdorf, A. Chynoweth, G. Dacey, et. al., “Uniform Silicon p-n Junctions. I. Broad Area Breakdown,” *Journal of Applied Physics* 31 (1960) 1153.

[Ber72] H. Berger, “Models for Contacts to Planar Devices,” *Solid Sate Elec.* 15 (1972) 145.

[Bir64] J. Birks, *The theory and practice of scintillation counting*, Macmillan, New York, 1964.

[Bur07] K. Burr and G. Wang, “Scintillation detection using 3 mm × 3 mm silicon photomultipliers,” *IEEE Nucl. Sci. Symp. Conference Record* (2007) 975.

[Cha05] W. Chang and Y Huang, “A new pre-etching pattern to determine 110 crystallographic orientation on 100 and 110 wafers,” *Microsystem Technologies* 11 (2005) 117.

[Chy56] A. Chynoweth and K. McKay, “Photon Emission from Avalanche Breakdown in Silicon,” *Physical Review* 102-2 (1956) 369.

[Chy60] A. Chynoweth, “Uniform Silicon p-n Junctions. II. Ionization Rates for Electrons,” *Journal of Applied Physics* 31 (1960) 1161.

[Con72] F. Conti and M. Conti, “Surface Breakdown in Si Planar Diodes Equipped with Field Plate,” *Solid-state Elec.* 15 (1972) 93.

[Dav64] R. Davies and F. Gentry, “Control of Electric Field at The Surface of P-N Junctions,” *IEEE Trans. Elec. Dev.* 11 (1964) 313.

[Din10] N. Dinu, C. Bazin., V. Chaumat, et. al., "Temperature and Bias Voltage Dependence of the MPPC Detectors," *IEEE Nucl. Sci. Symp. Conference Record* (2010).

[Erl09] A. Erlang, "The Theory of Probabilities and Telephone Conversations", *Nyt Tidsskrift for Matematik B* 20 (1909) 33.

[Fra09] T. Frach, G. Prescher, C. Degenhardt, et. al., "The digital silicon photomultiplier – Principle of operation and intrinsic detector performance," *IEEE Nucl. Sci. Symp. Conference Record* N28-5 (2009) 1959.

[Ger09] M. Gersbach, J. Richardson, E. Mazeleyray, et. al., "A low-noise single-photon detector implemented in a 130 nm CMOS imaging process," *Solid-state Elec.*, 53 (2009) 803.

[Gha78] P. Ghate, J. Blair, C. Fuller, et. al., "Application of Ti:W Barrier Metallization for Integrated Circuits," *Thin Solid Films* 53 (1978) 117.

[Goe64] A. Goetzberger, B. McDonald, R. Haitz, et. al., "Avalanche Effects in Silicon p-n Junctions. II. Structurally Perfect Junctions," *J. Appl. Phys*. 34 (1963) 1591.

[Gos07] M. Gosalvez, K. Sato, A. Foster, et. al., "An atomistic introduction to anisotropic etching," *J. Micromech. Microeng.* 17 (2007) S1.

[Gos09] M. Gosalvez, B. Tang, P. Pal, et. al., "Orientation- and concentration-dependent surfactant adsorption on silicon in aqueous alkaline solutions: explaining the changes in the etch rate, roughness and undercutting for MEMS applications," *J. Micromech. Microeng.* 19 (2009) 125011.

[Gre95] M. Green and M. Keevers, "Optical properties of intrinsic silicon at 300 K," *Progress in Photovoltaics* 3-3 (1995) 3-3.

[Gro84] D.E. Groom, "Silicon photodiode detection of bismuth germanate scintillation light," *Nucl. Instr. and Meth.* 219 (1984) 141.

[Gro85] D. Gross. *Fundamentals of queueing theory*. New York: John Wiley & Sons, Inc., 1985.

[Hai64] R.H. Haitz., "Model for the electrical behavior of a microplasma," *J. Appl. Phys.* 35 (1964) 1370.

[Hay98] B. Hayes, "How to Avoid Yourself," *American Scientist* 86-4 (1998) 314.

[Hol89] S. Holland, "An IC-Compatible Detector Process," *IEEE Trans. Nucl. Sci*. 36 (1989) 283.

[Hu08] J. Hu, X. Xin, X. Li, et. al., "4H-SiC Visible-Blind Single-Photon Avalanche Diode for Ultraviolet Detection at 280 and 350 nm," *IEEE Trans. Elec. Dev*. 55-8 (2008) 1977.

[Joh09] E. Johnson, P. Barton, K. Shah, et. al., "Energy resolution in CMOS SSPM detectors coupled to an LYSO scintillator," *IEEE Trans. Nucl. Sci.* 56-3 (2009) 1024.

[Joh10] E. Johnson, C. Stapels, X. Chen, et. al., "Large-area CMOS solid-state photomultipliers and recent developments," *Nucl. Instr. Meth. A* (article in press) (2010).

[Kaw96] M. Kawamura, Y. Abe, H Yanagisawa, et. al., "Characterization of TiN films prepared by a conventional magnetron sputtering system: influence of nitrogen flow percentage and electrical properties," *Thin Solid Films* 287 (1995) 115

[Kay93] S. Kay. *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory*, Pearson Education, 1993.

[Ker70] W. Kern and D. Puotinen, "Cleaning solutions based on hydrogen peroxide for use in silicon semiconductor technology," *RCA Review* 31 (1970) 187.

[Key01] P. Keys, "Phosphorus-Defect Interactions During Thermal Annealing of Ion Implanted Silicon," (Doctoral Dissertation) University of Florida, 2001.

[Kin99] W. Kindt, "Geiger Mode Avalanche Photodiode Arrays – For Spatially Resolved Photon Counting," (Doctoral Dissertation) Technische Universiteit Delft, 1999.

[Kno10] G. K. Knoll. *Radiation Detection and Measurement* (4th edition), New Jersey: John Wiley & Sons, 2010.

[Lag97] Y. Laghla and E. Scheid, "Optical study of undoped, B or P-doped polysilicon," *Thin Solid Films* 306 (1997) 67.

[Liś03] M. Liśkiewicz, M. Ogihara and S. Toda, "The complexity of counting self-avoiding walks in subgraphs of two-dimensional grids and hypercubes," *Theo. Comp. Sci.* 304 (2003) 129.

[Loe55] L. Loeb, *Basic Processes in Gaseous Electronics*, University of California Press, Berkeley, 1955.

[Log95] M. Loghmarti, K. Mahfoud, J. Kopp, et. al., "High Phosphorus Gettering Efficiency in Polycrystalline Silicon by Optimisation of Classical Thermal Annealing Conditions," *Phys. Stat. Sol.* 151 (1995) 379.

[Lut07] G. Lutz, *Semiconductor Radiation Detectors*, Springer, 2007.

[Mae90] W. Maes, K. De Meyer, and R. Van Overstraeten, "Impact Ionization in Silicon - A Review and Update," *Solid-state Elec.* 33-6 (1990) 705.

[McC71] J. McCaldin and H. Sankur, "Diffusivity and Solubility of Si in the Al Metallization of Integrated Circuits," *Appl. Phys. Lett*. 19-12 (1971) 524.

[Mci61] R. McIntyre, "Theory of microplasma instability in silicon," *J. Appl. Phys*. 32 (1961) 983.

[McK54] K. McKay, "Avalanche breakdown in silicon," *Phys. Rev*. 94 (1954) 877.

[Nic07] C. Niclass, M. Gersbach, R. Henderson, et. al., "A Single Photon Avalanche Diode Implemented in 130 nm CMOS," *IEEE J. Quant. Elec.*13-4 (2007) 863.

[Nin10] J. Ninkovic, L. Andricek, C. Jendrisyk et. al., "The first measurements on SiPMs with bulk integrated quench resistors," *Nucl. Instr. Meth*. (article in press, available July 2010).

[Ous05] S. Oussalah, B. Djezzar and R. Jerisian, "A comparative study of different contact resistance test structures dedicated to the power process technology," *Solid-state Elec.* 49 (2005) 1617.

[Pap91] A. Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 1991.

[Pav05] N. Pavlov, G. Maehlum, and D. Meier, "Gamma Spectroscopy using a Silicon Photomultiplier and a Scintillator," *IEEE Nucl. Sci. Symp. Conference Record*, N9-3, (2005) 173.

[Pol88] M. Polignano, G. Cerofolini, H. Bender, et. al., "Gettering Mechanisms in Silicon," *J. Appl. Phys*. 64-2 (1988) 869.

[Ren09] D. Renker and E. Lorenz, "Advances in solid-state photon detectors." *Journal of Instrumentation*, Vol. 4, 2009.

[Roc03] A. Rochas, “Single Photon Avalanche Diodes in CMOS Technology.” (Doctoral Dissertation) École Polytechnique Fédérale de Lausanne, France, 2003.

[Rog74] T. Rodgers and J. Meindl, "Short-channel V-groove MOS (VMOS) logic," *Solid-State Circuits Conference*. 17 (1974) 112.

[Sci03] E. Sciacca, A. Giudice, D. Sanfilippo, et. al., “Silicon Planar Technology for Single-Photon Optical Detectors,” *IEEE Trans. Elec. Dev.* 50 (2003) 918.

[Sei90] H. Seidel, “The mechanism of anisotropic silicon etching and its relevance for micromachining,” in *Digest of Technical Papers, Transducers 1987, 4th Intl. Conf. Solid-State Sensors and Actuators* (1987) 120.

[Spi05] H. Spieler, *Semiconductor Detector Systems*, New York: Oxford University Press, 2005.

[Ste85] R. Stengl et. al., “Variation of Lateral Doping-A New Concept to Avoid High Voltage Breakdown of Planar Junction”, *IEEE International Electron Devices Meeting Digest*, 6-4 (1985) 154.

[Sze66] S. Sze and G. Gibbons, “Effect of Junction Curvature on Breakdown Voltage in Semiconductors,” *Solid-state Elec.* 9 (1966) 831.

[Sze07] S. M. Sze, *Physics of Semiconductor Devices* (3rd edition). New Jersey: John Wiley & Sons, 2007.

[Tya68] M. Tyagi, “Zener and Avalanche Breakdown in Si Alloyed PN Junctions (part I) – Analysis of Reverse Characteristics,” *Solid-state Elec.* 11 (1968) 99.

[TSU07] Taurus™ TSUPREM-4, version Z-2007.03, March 2007. A computer code from the Synopsis corporation.

[Van70] R. Van Overstraeten and H. De Man, “Measurement of the Ionization Rates in Diffused Silicon p-n Junctions,” *Solid Sate Elec.*13 (1970) 583

[Wil96] K. Williams and R. Muller, “Etch Rates for Micromachining Processing,” *J. Microelec. Sys.*, 5-4 (1996) 256.

[Wil03] K. Williams, K Gupta, and M. Wasilik, “Etch Rates for Micromachining Processing – Part II,” *J. Microelec. Sys.*, 12-6 (2003) 761.

[Wil10] G. Williams, “Limitations of Geiger-mode arrays for Flash LADAR applications,” *Laser Radar Technology and Applications XV* 7684 (2010) 768414.

[Wu96] W. Wu and C. Chiou, “Effect of oxygen concentration in the sputtering ambient on the microstructure, electrical and optical properties of radio-frequency magnetron-sputtered indium tin oxide films,” *Semi. Sci. Tech.* 11 (1996) 196.

[Wu05] W. Wu, K. Sai, C. Chao, et. al., “Novel Multilayered Ti-TiN Diffusion Barrier for Al Metallization,” *J. Elec. Mat.* 34 (2005) 1150.

[Yu00] D. Yu and J. Fessler, “Mean and variance of single photon counting with deadtime,” *Phys. Med. Bio.* 45 (2000) 2043.

[Zen34] C. Zener, “A Theory of the Electrical Breakdown of Solid Dielectrics,” *Proceedings of the Royal Society of London. Series A* 145-855 (1934) 523.