CHARACTERIZATION OF SILICON GEIGER-MODE AVALANCHE PHOTODIODES WITH NOVEL DEVICE ARCHITECTURE

by

Kimberly E. Kolb

B.S. Rochester Institute of Technology

(2008)

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Chester F. Carlson Center for Imaging Science
of the College of Science
Rochester Institute of Technology

July 15, 2011

Signature of the Author ____

Accepted by ____
Coordinator, M.S. Degree Program Date

CHESTER F. CARLSON
CENTER FOR IMAGING SCIENCE
COLLEGE OF SCIENCE
ROCHESTER INSTITUTE OF TECHNOLOGY
ROCHESTER, NEW YORK

CERTIFICATE OF APPROVAL

M.S. DEGREE THESIS

The M.S. degree Thesis of Kimberly E. Kolb
has been examined and approved by the
thesis committee as satisfactory for the
thesis requirement for the
Master of Science degree

Dr. Donald F. Figer, Thesis Advisor

Dr. Karl Hirschman, Committee Member

Dr. Zoran Ninkov, Committee Member

Date

CHARACTERIZATION OF SILICON GEIGER-MODE AVALANCHE PHOTODIODES WITH NOVEL DEVICE ARCHITECTURE

by
Kimberly E. Kolb

Submitted to the
Chester F. Carlson
Center for Imaging Science
College of Science
in partial fulfillment of the requirements
for the Master of Science Degree
at the Rochester Institute of Technology

ABSTRACT

Geiger-mode avalanche photodiode (GM APD) detectors are capable of counting single photons, measuring arrival times with high resolution, and generating zero read noise (when operated with a CMOS digital readout circuit) due to their unique internal gain characteristics. These capabilities make them exceptionally suited to tasks that require precise arrival time measurements or characterization of faint signals (low photon flux). Laser ranging systems use their arrival time measurement capabilities to build three-dimensional images, while adaptive optics applications have recently begun to capitalize on their low noise and high-speed operation for correcting wavefront imperfections due to atmospheric interference. There is now growing interest in using GM APDs for imaging applications where accurate measurements of faint signals are necessary, such as in astronomy. MIT Lincoln Laboratory and the RIT Center for Detectors have developed silicon GM APDs with unique architecture, utilizing scupper regions to minimize detector noise.

This thesis investigates the performance of these detectors in terms of dark count rate (DCR). There are a number of mechanisms that produce dark counts, the most prominent being thermal excitation of carriers. Thermal carrier generation rates are generally only dependent on the temperature of the diode and may be constant under certain controlled conditions. Afterpulsing results from the release of carriers trapped in intermediate energy states (states with energy in the band gap of the material). Unlike thermal carrier generation, afterpulsing is dependent on the quenching time of the device (during which the device is unable to detect a carrier). Another mechanism, called self re-triggering, occurs when relaxing carriers emit photons during an avalanche. These photons can be absorbed in the substrate and generate dark carriers. Self-retriggering is also dependent on the quenching time of the device.

Theories for afterpulsing and self-retriggering are discussed. Specialized test circuitry is used with a customized data acquisition technique, and the author develops a method for parameter extraction from the raw data. Device characteristics derived from experimental results are examined. The author also develops a simulation program to approximate the dark count rate (among other parameters) of a device based on semiconductor characteristics and testing conditions. This thesis makes conclusions about the dependence of DCR on device architecture and how individual carrier generation mechanisms affect device performance.

LIST OF FIGURES

Figure 1 – Energy Band Diagram for a Semiconductor 3

Figure 2 – PN Junction Band Diagram (Van Zeghbroeck 2007)........ 6

Figure 3 – Four Quenching Circuit Schematics: (a) passive quench; (b) active quench; (c) active recharge; (d) active quench and recharge (Kindt and de Langen 1998) 10

Figure 4 – Gain vs. Bias Voltage for silicon Geiger mode and standard operation APDs(PerkinElmer, Inc. 2006) 12

Figure 5 – Inverse of Avalanche Build-Up time ($1/\tau_{av}$) vs. Depletion Width (w_{av}) at M = 20 for a germanium substrate APD(Ando and Kanbe 1985) 14

Figure 6 – Data Rate (bit/s) vs. Minimum Received Optical Level (dBm)(Ando and Kanbe 1985) 15

Figure 7 – Noise Level vs Gain for Variations on F (noise ratio) according to Eq. 14........ 19

Figure 8 – Optical layout of a curvature wavefront sensor-based AO system (Craven-Bartle, Dorn and Beletic 2000) 22

Figure 9 – 3D LADAR Image of a Portion of the RIT Campus (Ardt 2008)........ 24

Figure 10 – Wavelength upconverter based on periodically-poled lithium-niobate (PPLN) (Grein, et al. 2010) 25

Figure 11 – Doping Profiles of LFF (left) and HFF (right) devices(Aull, et al. 2010)........ 26

Figure 12 – Electric field distribution of LFF device architecture(Aull, et al. 2010)........ 28

Figure 13 – Electric field distribution of HFF device architecture(Aull, et al. 2010) 29

Figure 14 – Mean DCR vs Quench Time for both HFF (red) and LFF (blue) devices 30

Figure 15 – Test device area geometry. The circled contacts show the devices that were tested; device architecture varies from low to high going left to right and top to bottom. For scale the metal pads are approximately 70 µm x 70 µm 32

Figure 16 – Passive Quenching Setup and Equivalent Circuit Diagram 33

Figure 17 – Equivalent Circuit Diagram for State 1 of Passive Quenching Circuit Operation.... 34

Figure 18 – Equivalent Circuit Diagram for State 2 of Passive Quenching Circuit Operation.... 35

Figure 19 – Equivalent Circuit Diagram for State 3 of Passive Quenching Circuit Operation with considerations for the Thévenin equivalent circuit 36
Figure 20 – PSPICE Equivalent Circuit RC Recharge Simulation Results................................ 38
Figure 21 – Oscilloscope Screen Capture of Passive Quenching Diode Output.......................... 39
Figure 22 – Diagram of Active Quenching Test Setup.. 41
Figure 23 – Sample Waveforms for Arm, Quench, and APD Signals Under Active Quenching Operation 43
Figure 24 – Visualization of the Band-to-Band Tunneling Mechanism...................................... 46
Figure 25 – Self-Retriggering Mechanism in stages.. 47
Figure 26 – Poisson Probability Distribution for Various λ Values 52
Figure 27 – Combined Poisson Processes in a “First Arrival” Scenario 53
Figure 28 – Representation of Chi-Squared Value Storage technique 54
Figure 29 – Passive Quench Screen Capture with Cursor Measurements.................................. 59
Figure 30 – Example of Arrival Time Distribution with various artifacts present...................... 62
Figure 31 – Example of Artificial Avalanche Recognition Due to Circuit Leakage................... 63
Figure 32 – Observed Dark Counts for Various Dead Time Values for the Same Device 65
Figure 33 – Experimental Results: DCR vs Quench Time for LFF Device 67
Figure 34 – Experimental Results: DCR vs Quench Time for MFF Device............................... 68
Figure 35 – Experimental Results: DCR vs Quench Time for HFF Device................................ 69
Figure 36 – Simulated Results: DCR vs Quench Time for LFF Device 70
Figure 37 – Histogram of Chi-Squared Results for LFF Simulation Matching 72
Figure 38 – Simulated Results: DCR vs Quench Time for MFF Device 73
Figure 39 – Histogram of Chi-Squared Results for MFF Simulation Matching 74
Figure 40 – Simulated Results: DCR vs Quench Time for HFF Device 75
Figure 41 – Histogram of Chi-Squared Results for HFF Simulation Matching.......................... 76
Figure 42 – Experimental Results: First Bin Residuals for LFF Device 77
Figure 43 – Experimental Results: First Bin Residuals for MFF Device 78
Figure 44 – Experimental Results: First Bin Residuals for HFF Device..................................... 79
Figure 45 – Simulated vs Experimental Results: First Bin Residuals for LFF Device 81

Figure 46 – Simulated vs Experimental Results: First Bin Residuals for MFF Device 82
Figure 47 – Simulated vs Experimental Results: First Bin Residuals for HFF Device 83
Figure 48 – Alternate Simulated vs Experimental Results: First Bin Residuals for HFF Device 84
Figure 49 – Overall Experimental Results for DCR vs Quench Time 85
Figure 50 – Overall Experimental Results for First Bin Residuals 87
Figure 51 – Gaussian Surface for Diode Capacitance Calculations B-1
Figure 52 – Frame of Reference for Electric Field Contribution Function Calculation B-3
Figure 53 – Derivative Section of the Outer Capacitive Area B-6
Figure 54 – Variations on DCR vs Quench Time with respect to Base Thermal DCR; 30 kHz (top), 100 kHz (middle), 500 kHz (bottom) E-1
Figure 55 – Variations on DCR vs Quench Time with respect to Number of Filled Traps; 10 (top), 100 (middle), 300 (bottom) E-3
Figure 56 – Variations on DCR vs Quench Time with respect to Detrapping Lifetime; 5 µs (top), 10 µs (middle), 30 µs (bottom)...E-4
Figure 57 – Variations on DCR vs Quench Time with respect to Carrier Lifetime;10 µs (top), 50 µs (middle), 100 µs (bottom)...E-5

LIST OF TABLES

Table 1: Settings for Passive Quenching Measurements 58
Table 2: Settings for Active Quenching Measurements 58
Table 3: Expected and Actual Values for Passive Quench Parameters 60
Table 4: Device Diode Geometry 60
Table 5: LFF Simulation Match Input Values 71
Table 6: MFF Simulation Match Input Values 73
Table 7: HFF Simulation Match Input Values 75
Table 8: Alternate HFF Simulation Match Input Values 84
Table 9: CV Measurements for Large-Area Diode C-1
Table 10: DCR Values - LFF D-1
Table 11: DCR Values - MFF D-1
Table 12: DCR Values - HFF D-1
Table 13: First Bin Residual Values - LFF D-2
Table 14: First Bin Residual Values - MFF D-2
Table 15: First Bin Residual Values - HFF D-2
Table 16: Input Values for Sensitivity Analysis – Variations on Base Thermal DCR E-2
Table 17: Input Values for Sensitivity Analysis – Variations on Number of Filled Traps E-3
Table 18: Input Values for Sensitivity Analysis – Variations on Detrapping Lifetime E-4
Table 19: Input Values for Sensitivity Analysis – Variations on Carrier Lifetime E-6

TABLE OF CONTENTS

Abstract vii

List of Figures ix

List of Tables xiii

Table of Contents xv

1 Introduction 1

  1.1 Background Theory 2

    1.1.1 The PN Junction Diode 2

    1.1.2 Standard Photodiodes 6

    1.1.3 Avalanche Mechanism 7

  1.2 GM APD Operation 9

  1.3 Advantages of GM APDs 11

  1.4 Considerations for GM APDs 13

    1.4.1 Avalanche Buildup Time 13

    1.4.2 Pixel to Pixel Crosstalk 16

    1.4.3 Afterpulsing 16

    1.4.4 Premature Breakdown 16

    1.4.5 Avalanche Initiation Probability 17

    1.4.6 Comparison to P-I-N Diodes 18

2 Previous Work 20

  2.1 Past and Current Uses of Silicon GM APDs 20

    2.1.1 Historical Development 20

    2.1.2 Current Applications of Note 21

2.1.2.1 Adaptive Optics and Wavefront Sensing 21
2.1.2.2 LADAR 23
2.1.2.3 Communications 24
2.1.2.4 Other Applications 25
2.2 Development of GM APD Structure 26
2.2.1 Doping Profile Design 26
2.2.1.1 LFF Device 27
2.2.1.2 HFF Device 28
2.2.2 Previous Testing Results 29
3 Experimental Approach 31
3.1 Passive Quenching 33
3.1.1 Theory of Operation 34
3.1.2 Measuring Passive Quench RC Recharge 38
3.1.3 Diode Capacitance 40
3.2 Active Quenching 41
3.2.1 Dark Count Rate Theory 43
3.2.1.1 Thermal Dark Count 44
3.2.1.2 Afterpulsing 45
3.2.1.3 Band-to-Band Tunneling 45
3.2.1.4 Photon-Induced Self-Retriggering 46
3.2.2 DCR vs Quench Time 48
3.2.3 First Bin Residuals 49
4 Simulation of DCR Behavior of GM APDs 51

4.1 Theory 51

4.2 Approach 53

5 Experimental Data 57

5.1 Experimental Setup 57

5.2 Passive Quenching Results 58

5.2.1 RC Recharge Circuit Analysis 59

5.2.2 Diode Capacitance 60

5.3 Active Quenching Results 61

5.3.1 Initial Measurements and Improvements 61

5.3.2 Data Reduction Method 64

5.3.3 DCR vs Quench Time 66

5.3.3.1 LFF 67

5.3.3.2 MFF 68

5.3.3.3 HFF 69

5.3.3.4 Simulation Matching 70

5.3.4 First Bin Residual Results 76

5.3.4.1 LFF 77

5.3.4.2 MFF 78

5.3.4.3 HFF 79

5.3.4.4 Simulation Matching 80

6 Conclusions 85

6.1 DCR vs Quench Time 85

6.2 First Bin Residuals 87

6.3 Simulation 88

7 Contributions 90

8 Future Work 92

References 95

APPENDIX A: List of Acronyms and Definitions A-1

APPENDIX B: GM APD P-I-N Diode Capacitance Derivation B-1

APPENDIX C: C-V Measurements C-1

APPENDIX D: Raw Measured Data D-1

APPENDIX E: Sensitivity Analysis E-1

1 INTRODUCTION

The purpose of this thesis project is to test and characterize the behavior of silicon Geiger-mode avalanche photodiodes (GM APDs) in the absence of light and to study the internal mechanisms that govern their behavior. Understanding the dark noise inherent to such a device provides data that informs subsequent design improvements and provides optimum operational parameters to acquire the best signal-to-noise ratio (SNR) possible. In a collaborative effort, MIT Lincoln Laboratory and the RIT Center for Detectors are developing arrays of GM APDs hybridized to all-digital CMOS readout circuits. These devices can detect single photons, enabling quantum-limited sensitivity. In this implementation, a logic circuit in the pixel senses and digitizes the pulse from the GM APD in response to an incident photon. This photon-to-digital conversion enables the detectors to operate with zero readout noise because the information about photon count or arrival time is digitized before readout. As such, there is no noise penalty for operating at high frame rates or with short integration times.

GM APDs became a prominent research topic in the 1960s and 1970s, mainly as a means for replacing the increasingly expensive photomultiplier tube (PMT). As research fueled development, the advantages of GM APDs over PMT and charge-coupled device (CCD) systems became clear. GM APDs offered digital-mode operation and compact focal plane arrays (unlike the PMTs) and overcame the high readout noise and subsequently poor SNR of CCDs at low fluency levels (very few photons). As an alternative, the p-i-n diode can convert a single photon into a photoelectron, but there is no gain associated with this type of detector (only one signal carrier is generated per photon). Without gain, the electrical signal due to photon flux is usually lost in readout noise. For most photon-starved and high-frame-rate applications, the performance of digitized GM APDs for single photon counting exceeds the performance of PMTs, CCDs, and p-i-n diode detectors (Renker 2006).

GM APDs are well suited to applications involving faint signals and those that would benefit from high frame rate capabilities. Such applications include laser ranging systems, adaptive optics, and, more recently, imaging tasks where accurate measurements of faint signals are crucial (such as astronomy, remote sensing, and medical imaging). The ability to accurately count single photons is essential in these applications.

Characterization of this GM APD device prior to hybridization (bonding to the readout circuitry) is critical to future development. Due to variation in internal device architecture, there are differences in performance between low fill-factor (LFF) and high fill-factor (HFF) devices. These structures vary not only by active area (the portion of the device that can accept and detect photons), but by noise mitigation capabilities as well. For imaging applications, a high fill-factor is desirable so that the detector collects all the photon information possible. With the current design, however, this increase in fill-factor comes at the cost of detector noise. Properly understanding and identifying the underlying causes of this increase in noise will lead to targeted design improvements and therefore better performance in future iterations of the device.

1.1 BACKGROUND THEORY

Understanding the basic concepts of semiconductor physics is essential to grasping the motivation, results, and conclusions of this thesis. This section outlines an overview of device physics, specifically intrinsic (pure) semiconductor material, pn junctions, and specific operational situations related to GM APDs.

1.1.1 *THE PN JUNCTION DIODE*

A diode is the junction of two oppositely doped semiconductor materials. For instance, doping silicon with boron will cause the material to become p-type, or have a majority of positive carriers (holes) in the material. If phosphorus were introduced instead, the material would become n-type, or have a majority of negative carriers (electrons). These extra carriers enable higher conductivity. Figure 1 shows a band diagram, which illustrates the available energy levels for a carrier in a crystalline structure (Pierret 1996).

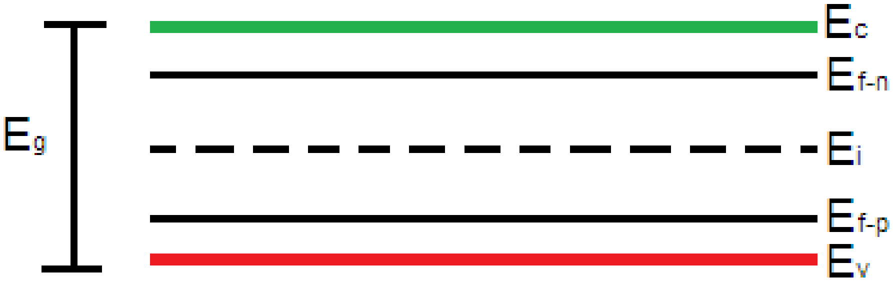


Figure 1 – Energy Band Diagram for a Semiconductor

E_c refers to the energy level of the conduction band (where electrons can freely move within the lattice), and E_v represents the energy of the valence band where electrons reside in their bound state. E_g is the band gap energy, or the amount of energy needed for an electron to move to the conduction band from the valence band, and is a constant for a semiconducting material. For silicon, E_g is 1.12 eV at 300 K. E_i refers to the average electron energy in the lattice, which is roughly 0.56eV (half the band gap) at room temperature. The probability that a carrier will occupy a certain energy state is a function of the Fermi energy level ($E_{f\text{-}n}$ for n-type dopants and $E_{f\text{-}p}$ for p-type dopants) and the temperature of the material. The Fermi function describes this probability (see Eq. 1) and the expressions for the corresponding energy levels are shown in the Fermi level expressions (Eq. 2 and Eq. 3) (Pierret 1996).

$$\boldsymbol{f(E) = \frac{1}{1 + e^{\frac{E - E_f}{kT}}}} \qquad \text{Eq. 1}$$

f(E) denotes the probability (under equilibrium conditions) that an electron will occupy an available state of energy E. k is the Boltzmann constant and T is the temperature of the substrate in Kelvin. E_f is the Fermi level, which can be calculated as in Eq. 2 and Eq. 3 (Pierret 1996).

$$\boldsymbol{E_{f-n} = E_i + kTln\left(\frac{N_D}{n_i}\right)} \qquad \text{Eq. 2}$$

$$\boldsymbol{E_{f-p} = E_i - kTln\left(\frac{N_A}{n_i}\right)} \qquad \text{Eq. 3}$$

Eq. 2 applies to n-type (donor) dopants, with N_D being the number of dopant atoms per unit volume. Eq. 3 applies to p-type (acceptor) dopants, with N_A having the same units as N_D. n_i is the intrinsic carrier concentration for the semiconductor (~1e10 cm^{-3} for silicon). For most applications, only one type of dopant is used in a particular volume of the semiconductor (either p-type or n-type). A mode of doping exists called compensated doping, which includes both n- and p-type dopants in the same region, but imaging detectors do not generally utilize these doping schemes (Pierret 1996).

The carrier densities n and p (of electrons and holes, respectively), shown in Eq. 4 and Eq. 5, represent the number of carriers per unit volume. These values are dependent on the material's intrinsic characteristics, the Fermi level, and the temperature.

$$\boldsymbol{n = n_i e^{(E_f - E_i)/kT}} \qquad \text{Eq. 4}$$

$$\boldsymbol{p = n_i e^{(E_i - E_f)/kT}} \qquad \text{Eq. 5}$$

Eq. 6, called the law of mass action, directly relates n and p to the intrinsic carrier concentration of the material. Using Eq. 4 - Eq. 6, one can calculate n, p, or E_f from only one of the variables in the group.

$$\boldsymbol{np = n_i^2} \qquad \text{Eq. 6}$$

Current is defined as the net motion of carriers through the semiconductor. This does not refer to one carrier traveling the full length of a device, but to the average carrier motion that favors one direction. For an electron to be free to move about the lattice, it must gain enough energy to exist in the conduction band. When the electron makes this jump, it leaves behind a hole, which is a positive carrier. Eventually, the electron will encounter another hole and recombine. The mean distance that the carriers can travel before recombining is the diffusion length, and the mean time the carrier takes to travel that length is the carrier lifetime.

The semiconductor physics of a common structure called the pn junction include the basic relationships already discussed. As its name indicates, a pn junction is a p-type layer adjacent to an n-type layer. The majority carriers (electrons in an n-type material and holes in a p-type material) on one side of the junction diffuse to the other side, giving rise to space charge layers on each side of the junction; the resulting electric field creates a built-in potential barrier to the further diffusion of carriers. This layer is also called a depletion region because the carrier concentrations in this region are negligible compared to the majority carrier concentrations outside of the layer. Eq. 7 gives an expression for the width of the depletion region (Pierret 1996).

$$\boldsymbol{W = \sqrt{\frac{2\epsilon_r\epsilon_0}{q}\left(\frac{N_A + N_D}{N_A N_D}\right)\left(V_{bi_-}V_{app}\right)}} \qquad \text{Eq. 7}$$

V_{app} is the voltage applied to the junction, and V_{bi} is the built-in potential of the junction, described in Eq. 8 (and seen in Figure 2 as $q\emptyset_i$).

$$\boldsymbol{V_{bi} = \frac{kT}{q} ln\left(\frac{N_D N_A}{{n_i}^2}\right)} \qquad \text{Eq. 8}$$

One side of the junction is usually much more heavily doped than the other. In that case, the larger value will fall out of the dopant term in Eq. 7, revealing a dependence on the dopant concentration on the lightly doped side (Pierret 1996).

The band structure across the depletion width changes for a pn junction as compared to a bulk material with no junction as in Figure 1. Figure 2 shows the altered band structure for a pn junction (Van Zeghbroeck 2007).

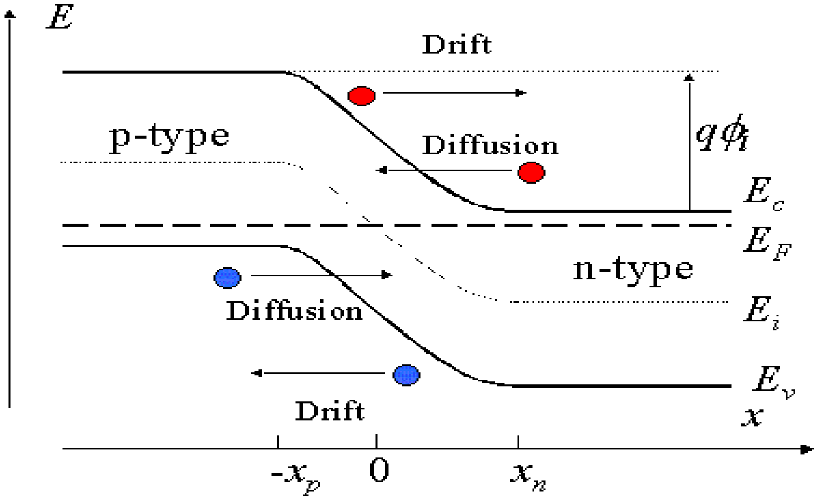


Figure 2 – PN Junction Band Diagram (Van Zeghbroeck 2007)

In Figure 2, x_n and $-x_p$ represent the bounds of the depletion width on the n- and p-sides of the junction, respectively (the zero point is the physical location of the pn junction). Electrons flow easily from left to right (drift current), but flowing from right to left (diffusion current) requires more energy; the inverse is true for holes. A positive voltage applied to the p-side of the junction in excess of the built-in voltage V_{bi} eliminates the potential barrier for diffusion current, turning the device on. Eq. 9 describes the current flowing through a diode.

$$\boldsymbol{I = I_0\left(e^{\frac{qV_{app}}{kT}} - 1\right)} \qquad \text{Eq. 9}$$

I_0 is the leakage or dark current associated with the reverse drifting of carriers for any device. It can be obtained by I-V measurements at a large reverse bias (Pierret 1996).

1.1.2 *STANDARD PHOTODIODES*

The case of the photodiode is a theoretical extension of the pn junction. In a regular, forward-biased diode, a positive voltage across the junction removes the potential barrier and free carriers will drift (current due to an electric field) across the depletion width. This is not the case for a photodiode, which operates in reverse bias (increasing the potential barrier for drift current). The only significant carrier movement is dark current (diffusion current due to a density or thermal gradient, independent of applied bias), which is usually very small (~1 nA) because it is limited by the number of minority carriers available. When reverse-biased, the depletion area behaves like a capacitor storing charge with capacitance defined by Eq. 10.

$$\boldsymbol{C_j = \frac{A_j}{2}\sqrt{\frac{2q\epsilon_r\epsilon_0}{V_{bi} - V_{app}}N_B}} \qquad \text{Eq. 10}$$

N_B is the smaller of the two doping levels and A_j is the cross-sectional area of the junction. This equation has a built-in calculation of the depletion width. If the depletion width at the set voltage is known, however (or if the voltage indicates maximum depletion width), then the equation reverts to the standard parallel plate calculation for a capacitor with silicon as the dielectric layer (Pierret 1996).

While the diode is in this reverse-biased state, a photon incident on the depletion region of the device excites a carrier to the conduction band, generating an electron-hole pair. Each carrier then drifts with the electric field across the junction: the electron goes to the n-side and the hole goes to the p-side. This alters the total charge contained in the junction and the capacitance value, which modifies the voltage across the junction. A readout circuit reports the voltage difference, which is directly proportional to the number of photons the device detected. The signal is then read out, digitized, and displayed by a computer.

1.1.3 *AVALANCHE MECHANISM*

When an electron has energy in the conduction band, it can move about the lattice structure until it recombines with a hole. Along its path, it may interact with other atoms in the lattice, causing some vibrations, which are generally of little significance. However, if the electron has sufficient energy (imparted by a strong electric field) it may dislodge another electron and the associated hole. These newly freed carriers immediately accelerate due to the strong electric field, dislodging more electron-hole pairs, causing an avalanche of charge to propagate through the lattice. This is the avalanche mechanism in a semiconductor. Once an avalanche has begun, a competition develops between the rate at which electron-hole pairs are generated and the rate at which they are collected at the device terminals (Pierret 1996).

At biases below some breakdown voltage (V_{BR}), collection dominates, causing the avalanche current to decay and ultimately stop. The gain (the number of carriers generated during an avalanche by a single initiating carrier) in this scenario is finite, and is determined by the statistics of the avalanche process. This type of operation is called linear-mode, since photocurrent is proportional to incident photon flux.

For linear-mode operation, Eq. 11 defines an empirical relationship between the multiplication factor (or gain) M and the reverse bias V_{app}.

$$\boldsymbol{M = \frac{1}{1 - \left[\frac{|V_{app}|}{V_{BR}}\right]^{m}}} \qquad \text{Eq. 11}$$

m can take on values between three and six, and is dependent on the substrate material. M corrects the ideal diode equation when the reverse bias voltage approaches the breakdown voltage (Pierret 1996).

In the case where the bias is above V_{BR}, multiplication outpaces collection. Initially, this causes exponential growth of the current. After some length of time, electrons and holes accumulate at the n- and p-sides of the depletion region, respectively. This creates an internal electric field that is in opposition to the applied bias and arrests the growth of the current. The device remains in an on state, however, until the circuit reduces the applied bias (quenches the device), allowing the APD to turn off. In this type of operation, known as Geiger-mode, the gain would be infinite if the bias were held above breakdown. With the use of quenching, however, the gain is determined by the circuit rather than by the avalanche statistics. In either case, an electrical event resulting from a single incident photon is indistinguishable from one initiated by a larger number of photons arriving simultaneously.

It is important to note that even in Geiger-mode, there is a probability that the avalanche may dwindle in its earliest stages and result in a non-detectable signal. While gain is an important metric for linear-mode operation, Geiger-mode operation is better characterized by the probability that the avalanche will become self-sustaining, referred to as the avalanche initiation probability.

Applications that utilize APDs take advantage of the avalanche mechanism to boost signal from a single incident photon. For example, an observation of a target emitting very low photon flux may have a low signal-to-noise ratio (SNR) if the photocurrent is similar in magnitude to the dark current or to the readout noise. When an avalanche multiplies the signal without multiplying the dark current, the SNR greatly improves and the target becomes discernable in the observation (Pierret 1996).

1.2 GM APD OPERATION

Geiger-mode operation is defined by a bias above breakdown that ensures a signal gain sufficient to count single photons. It is useful in applications where there is very little incident flux and provides a more reliable and predictable multiplication of signal than a photomultiplier tube (which GM APDs were originally developed to replace) for the same applications. The focus of this thesis is the testing and characterization of GM APDs with separate absorption and multiplication regions of the device to optimize the performance of each.

A Geiger-mode APD has the rare ability among detectors to count single photons. The bias voltage is set above the breakdown voltage so any carrier generated will immediately start an avalanche (Kindt and de Langen 1998). This is very useful for applications in which incident radiation is rare (e.g. muon detection). Whereas in a standard photodiode the signal would be lost amongst the noise, the high signal multiplication in Geiger-mode generally overcomes any noise present. Cooling the device augments this, reducing dark counts (a dark current carrier setting off an avalanche event in the absence of an incident photon) (PerkinElmer, Inc. 2006).

To extract useful information from detection events, photodetectors require some form of integrated circuit beyond a simple output. GM APDs require a number of specific operating conditions in order to take advantage of their unique characteristics. These requirements include a sufficient reverse bias and some form of quenching (and resetting) circuitry.

As a GM APD avalanches in the on state, the current through the device increases exponentially. The purpose of a quenching circuit is to detect that the GM APD has turned on and reduce the APD bias so that it turns off again (Kindt and de Langen 1998).

There is a variety of quenching circuit layouts, and Figure 3 shows examples of the most common forms: passive quenching, active quenching, active recharging, and active quenching and recharging (Kindt and de Langen 1998). Note that these circuits (especially the active quenching variety) can be complex and involve a number of ICs, and the schematics shown are only conceptual. The specific quenching systems used for the device tested in this project will be discussed in later sections.

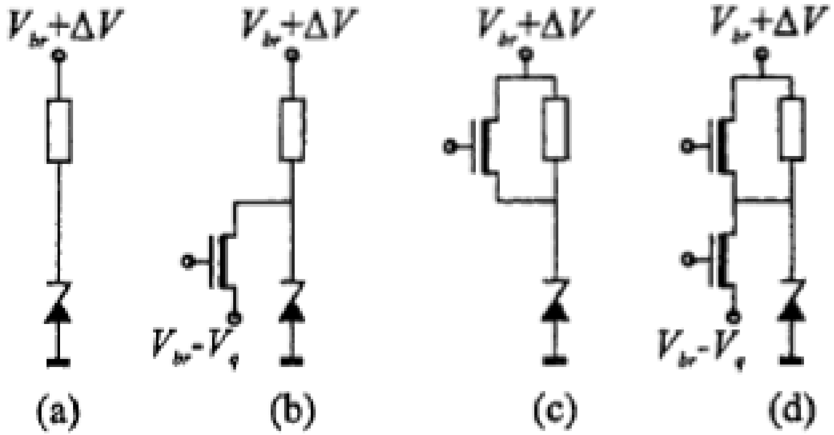


Figure 3 – Four Quenching Circuit Schematics: (a) passive quench; (b) active quench; (c) active recharge; (d) active quench and recharge (Kindt and de Langen 1998)

V_{BR} is the breakdown voltage of the diode, V_q is the quench voltage (a voltage below the breakdown voltage), and ΔV is the overbias for the diode (the applied bias minus V_{BR}). The quenching circuit in Figure 3a is known as a passive circuit. The block at the top of the diagram is a large resistor or current source, which prevents the current from rising above a certain level by Ohm’s Law. This slows the recharge of the junction, however, because it limits the speed at which charge can accumulate on the APD. Figure 3b is an active quenching circuit. A transistor sets the off-state voltage across the APD and a smaller resistor is placed in series with the APD. When the time comes to recharge the device (after the APD has entered breakdown), the transistor switches off, allowing current to pass through the small resistor. The reduction in resistance reduces the recharge time compared to the large resistance used in the passive quenching circuit. Figure 3c shows an active recharge structure. Here, the transistor is positioned across the resistor and is turned on during the recharge period to rapidly discharge the diode. The resistor limits the current flowing through the APD. Finally, Figure 3d shows a combination of approaches called active quench and recharge. The transistor positioned across the diode sets a voltage and prevents any afterpulsing after discharge (which will be discussed in a subsequent section), and the resistor (or current source) is in parallel with the upper transistor as in Figure 3c. The large parasitic capacitance across the diode is a drawback for this approach (as opposed to the smaller parasitic capacitance in Figure 3c) (Kindt and de Langen 1998).

1.3 ADVANTAGES OF GM APDS

There are differences between GM APDs and other detector types such as CCDs or even linear-mode APDs. In some circumstances, these differences are advantageous to the operation of a system or may offer a better way of implementing the system.

The avalanche effect in an APD can theoretically boost signal without boosting noise at all, because the only amplification is that of the original photon-generated carrier. In reality, there is an increase in shot noise (which is a function of the signal flux), and there is some propagation noise in terms of crosstalk (which will be discussed later). An example of the gains expected from a structure optimized for GM APD operation and a standard APD structure are shown in Figure 4 (PerkinElmer, Inc. 2006).

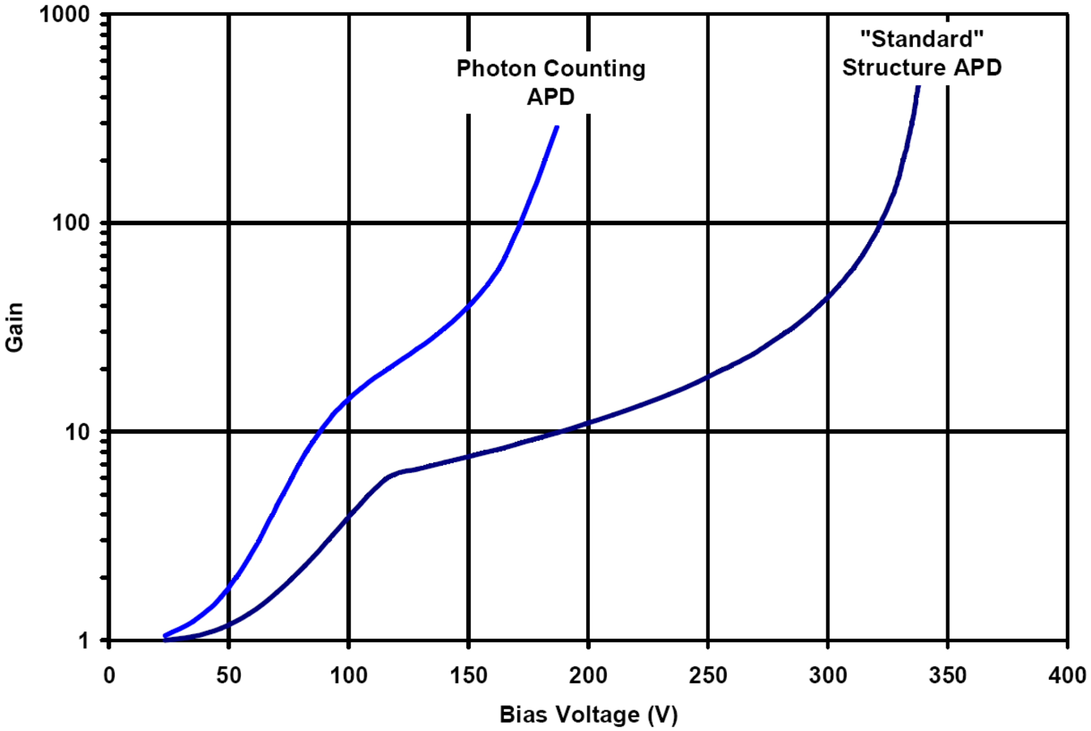


Figure 4 – Gain vs. Bias Voltage for silicon Geiger mode and standard operation APDs (PerkinElmer, Inc. 2006)

The scale of voltages necessary to achieve such gain for these devices is large – the chart in Figure 4 extends to 350 V for the gain measurements. An APD engineered to have a lower breakdown voltage for Geiger-mode devices will exhibit a steeper rise in gain at lower voltages than its standard structure counterpart for more efficient operation. Both structures approach a gain (M of Eq. 11) of 1000, which means that for every photon-generated electron-hole pair, one thousand carriers are collected. The chart in Figure 4 is for silicon, which generally has gain values between 50 and 1000 for linear-mode operation, while materials such as germanium or indium-gallium-arsenide generally only achieve M values of 10-40 (PerkinElmer, Inc. 2006).

In the case of the standard structure, linear-mode APDs are a good choice for use if the noise constraint of the system will be the preamplifier noise or other noise in the measurement system (dark current), not the shot noise of the signal itself. For applications in which this is not the case, a p-i-n diode (in which one photon generates only one carrier) may be a better choice (Gullikson, Gramsch and Szawlowski 1995).

1.4 CONSIDERATIONS FOR GM APDS

Although GM APDs have many advantages, some characteristics may be drawbacks in certain measurement scenarios. Avalanche buildup time, pixel to pixel crosstalk (both optical and electrical), afterpulsing, premature breakdown, avalanche initiation probability, and a GM APDs merits when compared to a p-i-n diode should all be considered before integrating a GM APD into a measurement system design.

1.4.1 *AVALANCHE BUILDUP TIME*

The avalanche mechanism is not instantaneous and requires time to build momentum, which affects the speed at which the device can react to light. The depletion width (Eq. 7) directly affects the build-up time, and Figure 5 illustrates this relationship (Ando and Kanbe 1985).

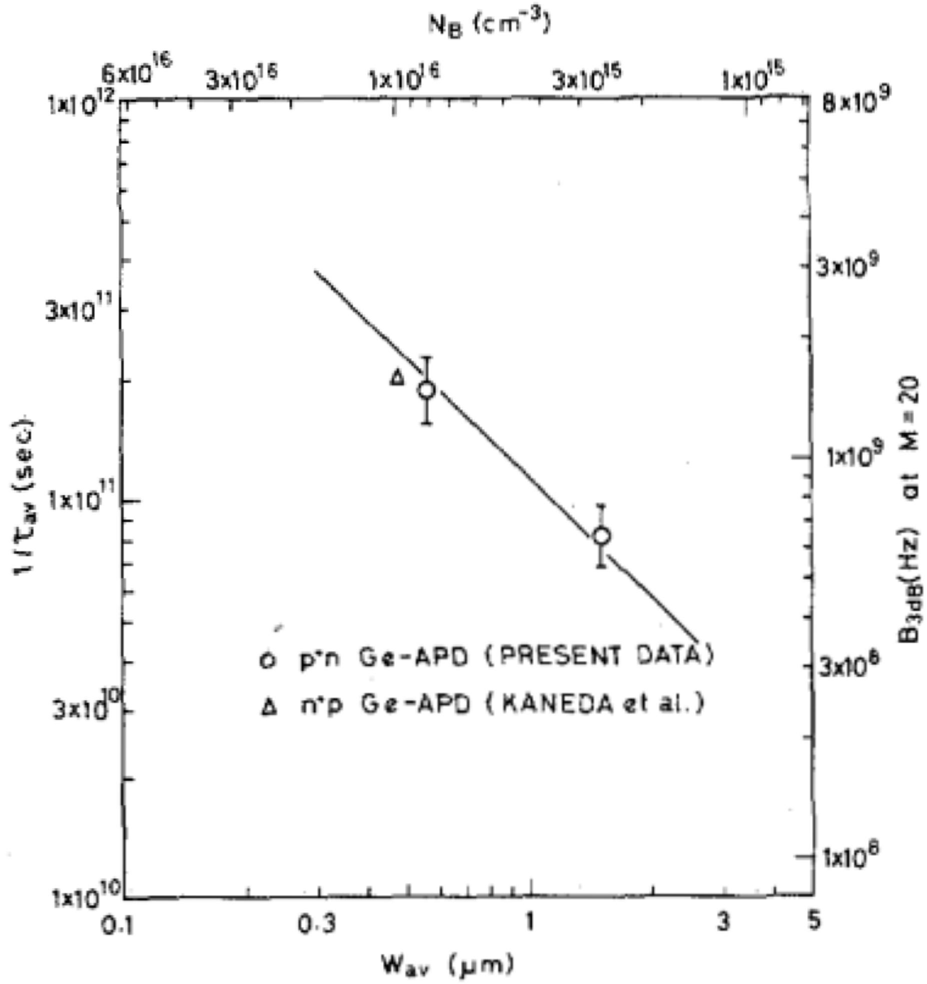


Figure 5 – Inverse of Avalanche Build-Up time ($1/\tau_{av}$) vs. Depletion Width (w_{av}) at M = 20 for a germanium substrate APD (Ando and Kanbe 1985)

The chart in Figure 5 contains a few points of interest. It reinforces the fact that the depletion width (w_{av}, lower x-axis) is proportional to the doping concentration (N_B, upper x-axis) of the lightly doped side of the pn junction, as mentioned earlier. It also shows that germanium has a lower gain potential than silicon (right y-axis, M = 20). The most important point is that an increase in the depletion width results in an increase of the avalanche build-up time (τ_{av}). This is intuitive given the theoretical discussion in previous sections – the carriers have less initial potential energy and so τ_{av} is increased (Ando and Kanbe 1985).

Geiger-mode does have an advantage over linear-mode in terms of the effects of avalanche build-up time, which introduces timing jitter in the latter. Figure 6 shows the minimum received optical level in dBm (incoming flux signal) as a function of the readout data rate in bit/s (x-axis) for constant SNR. Data for two different readout circuit impedance levels is presented. According to the plot, higher bit rates (x-axis), which imply shorter readout times, increase the necessary flux level (y-axis). This makes sense because more incoming flux means more signal per time interval, maintaining the SNR. The plot also indicates that as the avalanche build-up time increases from the theoretical zero (solid lines), the signal level must increase in order to maintain the same SNR at a specific data rate. For example, the minimum optical level required at a data rate of 1.1x10^9 bit/s (red points on the low-impedance curves) increases with increasing τ_{av} (Ando and Kanbe 1985).

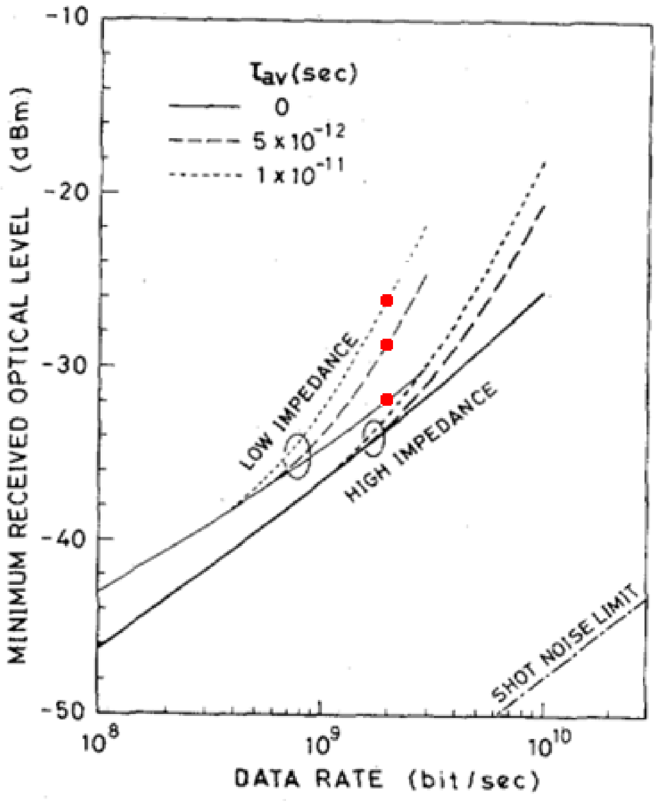


Figure 6 – Data Rate (bit/s) vs. Minimum Received Optical Level (dBm)(Ando and Kanbe 1985)

1.4.2 *PIXEL TO PIXEL CROSSTALK*

A GM APD has a relatively large crosstalk component caused both by parasitic light emissions from the large flux of electrons moving through the semiconductor and an increase in noise due to the number of carriers moving all at once. The noise due to parasitic light is caused by hot carriers in the high-field region emitting photons when they relax (from a high energy level to a lower energy level) which in turn causes carrier generation in a nearby pixel. Pixels are generally not well isolated physically, and electrically there is no barrier except an electric field to guide carriers to the nearest collection site. This also explains the noise due to the number of carriers. As the avalanche gets larger in scale, more and more carriers become free throughout the depletion area. Some of these carriers are generated at the edge of the pixel boundary and are collected in the adjacent pixel (Blazej, et al. 2006).

1.4.3 *AFTERPULSING*

Afterpulsing occurs when a trap releases a carrier that initiates an avalanche in the absence of an incident photon. Traps are energy states that exist in the band gap (see Figure 1) and can indicate the presence of impurities other than the major dopants. Other structural imperfections such as lattice mismatch, lattice damage from processing, or dangling bonds at material boundaries can also cause trap states. Unwanted impurities, frequently metal atoms, exist in the bulk due to contamination during growth of the silicon crystal or device fabrications. They provide states in which carriers become trapped; in other words, these carriers do not go on to create more electron-hole pairs. The carriers can remain in these states longer than the readout or dead time of the device, becoming free again during the next detection cycle. The electric field immediately accelerates the newly mobile electron or hole, which may initiate an avalanche of carriers in the absence of photon signal (Kindt and de Langen 1998).

1.4.4 *PREMATURE BREAKDOWN*

The high voltage levels applied to an APD induce large electric fields at the surface of the device, particularly where the junction has high curvature, such as at the periphery of the diode where the junction bends up to the surface. Premature breakdown can occur (due to the nonuniform distribution of voltage across the diode) at these high curvature sites, of which only a small portion is the junction area (Squillante, et al. 2003).

1.4.5 *AVALANCHE INITIATION PROBABILITY*

For each carrier generated by an incident photon in a GM APD, there is a corresponding avalanche initiation probability that determines the probability that the carrier will initiate a self-sustaining avalanche. This probability is a function of the carrier generation location within the depletion region and the overbias of the diode.

Carriers generated at the full length of the depletion region have more potential energy than carriers generated at the mid-point of the region and will generate more kinetic energy during the path to the collection node. This indicates a higher probability of ionizing impacts with other carriers, which is the driving mechanism behind the formation of an avalanche. The amount of overbias defines the electric field inside the depletion region, which affects the potential energy of a carrier generated at any depth. The following set of differential equations (Eq. 12 and Eq. 13) can be used to calculate $P_e(x)$ (or $P_h(x)$), the probability that an electron (or hole) generated at position x in the depletion layer will initiate an avalanche (McIntyre 1973).

$$\frac{d\boldsymbol{P_e}}{d\boldsymbol{x}} = (\mathbf{1} - \boldsymbol{P_e})\boldsymbol{\alpha_e}[\boldsymbol{P_e} + \boldsymbol{P_h} - \boldsymbol{P_e}\boldsymbol{P_h}] \qquad \text{Eq. 12}$$

$$\frac{d\boldsymbol{P_h}}{d\boldsymbol{x}} = -(\mathbf{1} - \boldsymbol{P_h})\boldsymbol{\alpha_h}[\boldsymbol{P_e} + \boldsymbol{P_h} - \boldsymbol{P_e}\boldsymbol{P_h}] \qquad \text{Eq. 13}$$

α_e and α_h are the electron and hole ionization rates. When these parameters are known, the calculation becomes relatively simple with the boundary conditions that the probabilities of avalanche initiation at $x = 0$ and $x = w$ (the depletion width) are zero for P_e and P_h, respectively. Most often, however, these parameters are unknown, and calculation of them is intricate. The avalanche initiation probability is a component of the overall photon detection probability, the measurement of which is usually sufficient for testing and characterization purposes (McIntyre 1973).

1.4.6 *COMPARISON TO P-I-N DIODES*

The functionality of a p-i-n diode is distinctly different from that of a GM APD. The device consists of a p+ layer, an intrinsic layer, and an n+ layer. In integration mode, the diode is reverse-biased and the intrinsic layer in the middle of the device facilitates a large depletion region in which carriers can travel with relatively long lifetimes. The bias is not sufficient to give avalanche gain, and so one photon can only generate one electron-hole pair. This necessitates a pristine bulk material and very sensitive electronics in the readout circuitry to accurately count the collected carriers. In thin p-i-n devices, the dark current is usually very low and these devices are useful for low-flux imaging where the detector integrates for a prolonged period (high dark current means saturation at long integration times). The SNR must be high in order to discern faint objects from noise.

Previous sections mention p-i-n diodes as an alternative to APDs for their lower-noise qualities. To quantify the extent to which the two devices differ, Eq. 14 illustrates the total noise of a linear-mode APD in reference to a comparable p-i-n diode (PerkinElmer, Inc. 2006). Note that gain discussion is not pertinent to Geiger-mode operation, but is included as a basis for operational differences between p-i-n diodes and APDs in general.

$$\boldsymbol{i_n = \sqrt{2qB\left(I_{DS} + F(I_{DB}M^2 + R_0(\lambda)M^2P_S)\right)}} \qquad \text{Eq. 14}$$

i_n is the total noise, B refers to the electrical bandwidth of the detection circuit, I_{DS} and I_{DB} refer to the surface and bulk dark noise (respectively), M is the multiplication factor addressed previously (see Eq. 11), $R_0(\lambda)$ is the responsivity of the device at the specified wavelength, and P_S is the optical signal power. F is the noise factor, or noise ratio, of an APD with respect to a comparable p-i-n diode. The equation reduces to shot, bulk (I_{DB}), and surface (I_{DS}) noise when F = 1 for a p-i-n diode (PerkinElmer, Inc. 2006).

Figure 7 shows the noise level (normalized to shot, bulk, and surface noise for a p-i-n diode) as a function of the gain M of a linear-mode APD over multiple values of the noise ratio F between a linear-mode APD and a p-i-n diode. All other values (B, I_{DS}, I_{DB}, $R_0(\lambda)$, and P_s) are considered constant for both devices. The log-log plot shows a steady increase in noise level with increases in the gain and noise ratio. Note that when M and F are equal to 1, the noise level is also equal to 1 as expected due to the normalization.

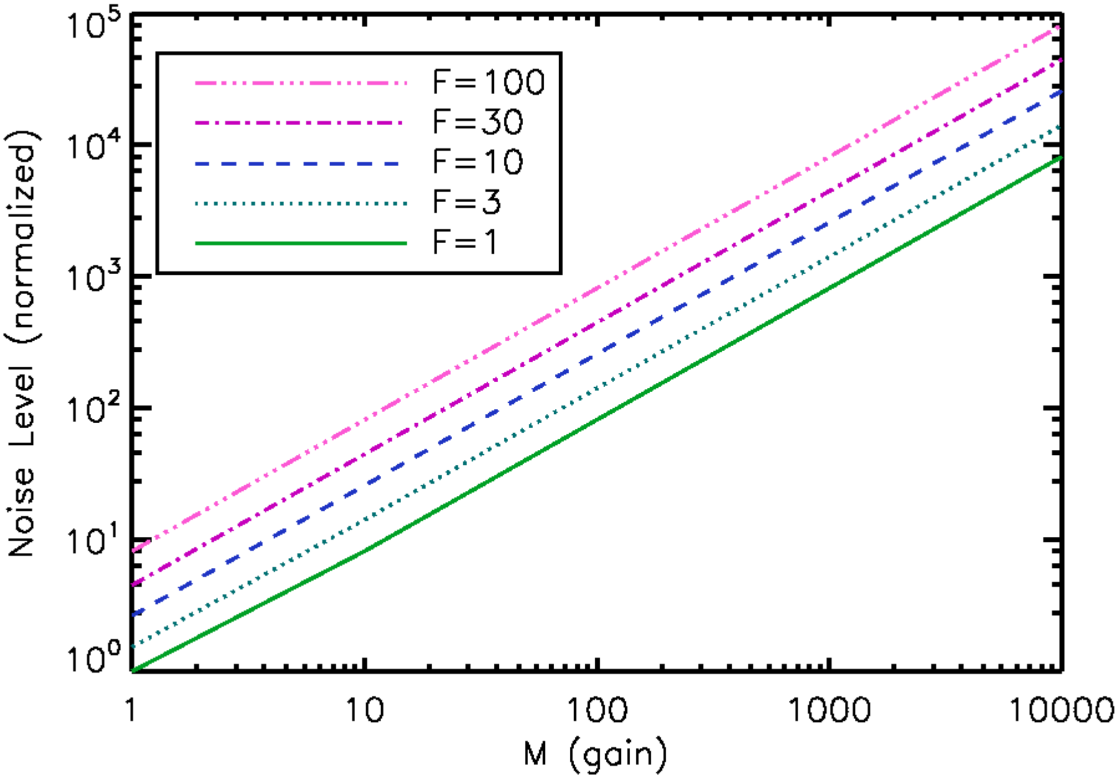


Figure 7 – Noise Level vs Gain for Variations on F (noise ratio) according to Eq. 14

2 PREVIOUS WORK

Exploring previous work and development of GM APDs and their uses along with the development of the current device architecture (and the intended purposes of its features) is critical to understanding the context of the motivation that drives this thesis. Understanding past applications brings greater awareness to the versatility of these detectors and their role in the future.

2.1 PAST AND CURRENT USES OF SILICON GM APDS

From first inception to current cutting-edge applications, GM APDs have been a distinctive detector type and have been important to many scientific endeavors. Their gain characteristics have made them an ideal solution to signal amplification without an amplification of noise. As GM APD designs and implementations improve, more applications utilize them, and these devices are often on the cutting edge of measurement system evolution.

2.1.1 *HISTORICAL DEVELOPMENT*

Avalanche photodiodes became a popular area of research in the 1960s and early 1970s, with both linear- and Geiger-mode functions being topics of interest. The motivation for this interest was the replacement of photo-multiplier tubes (PMTs), as modern experiments were likely to contain high electric fields (which degrade PMT performance). High PMT fabrication costs were also a factor; much of the internal architecture had to be handmade. GM APDs could replace PMTs as both a signal amplifier for low flux imaging applications and as a means to detect single photons (Renker 2006).

A notable benchmark in first linear-mode designs was produced by Perkin-Elmer, called SLIK^{TM}, which (like other contemporary devices) utilized passive quenching and were therefore rather slow (100 kHz cycle time) compared to modern, actively quenched devices (1 MHz cycle time). The next generation of single photon-counting APDs came from Rockwell International Science Center in 1987 when they developed the Solid State PhotoMultiplier (SSPM). The device design utilized very high donor concentrations, creating an impurity band level a mere 50 meV below the conduction band, which allowed the device to be sensitive even into the near

IR. The structure was later modified to have less sensitivity in the longer wavelengths and exists now as the Visible Light Photon Counter (VLPC) (Renker 2006).

One of the last radical improvements to the APD design occurred in Russia around 1990, called the Metal-Resistor-Semiconductor (MRS) APD. A thin metal layer coats either SiC or Si_xO_y (the resistor layer), allowing for a local reduction of the electric field, which in turn limits the magnitude of the diode breakdown. A smaller breakdown magnitude leads to faster quenching, which in turn leads to faster cycle times (Renker 2006).

2.1.2 *CURRENT APPLICATIONS OF NOTE*

The gain characteristics of GM APDs make them exceptional amplifiers of single photons (as an alternative to PMTs, for example), but GM APDs have many other potential applications. From simple amplifiers to cutting-edge systems like adaptive optics, LADAR (LAser Detection and Ranging), and communications, the GM APD's distinct capabilities offer both precision and high SNR.

2.1.2.1 Adaptive Optics and Wavefront Sensing

Adaptive Optics (AO) uses wavefront sensing to correct for atmospheric aberrations in a scene, primarily in ground-based astronomy. Over the course of an observation period, even while accounting for the rotation of the Earth, a star's signal will shift in the focal plane due to interaction with the Earth's atmosphere. Wavefront sensing systems can measure the tilt or curvature of the incoming wavefront using optical components and a detector. In a standard setup, a beam splitter divides the signal into two portions: one part of the signal goes to the wavefront sensing detector and the other to the imaging detector. The information collected by the wavefront-sensing detector feeds back into the system controlling the imaging optical components. These optical components manipulate the image signal so that the scene remains stationary on the focal plane. Figure 8 shows an example of the wavefront sensing portion of an adaptive optics system (Craven-Bartle, Dorn and Beletic 2000).

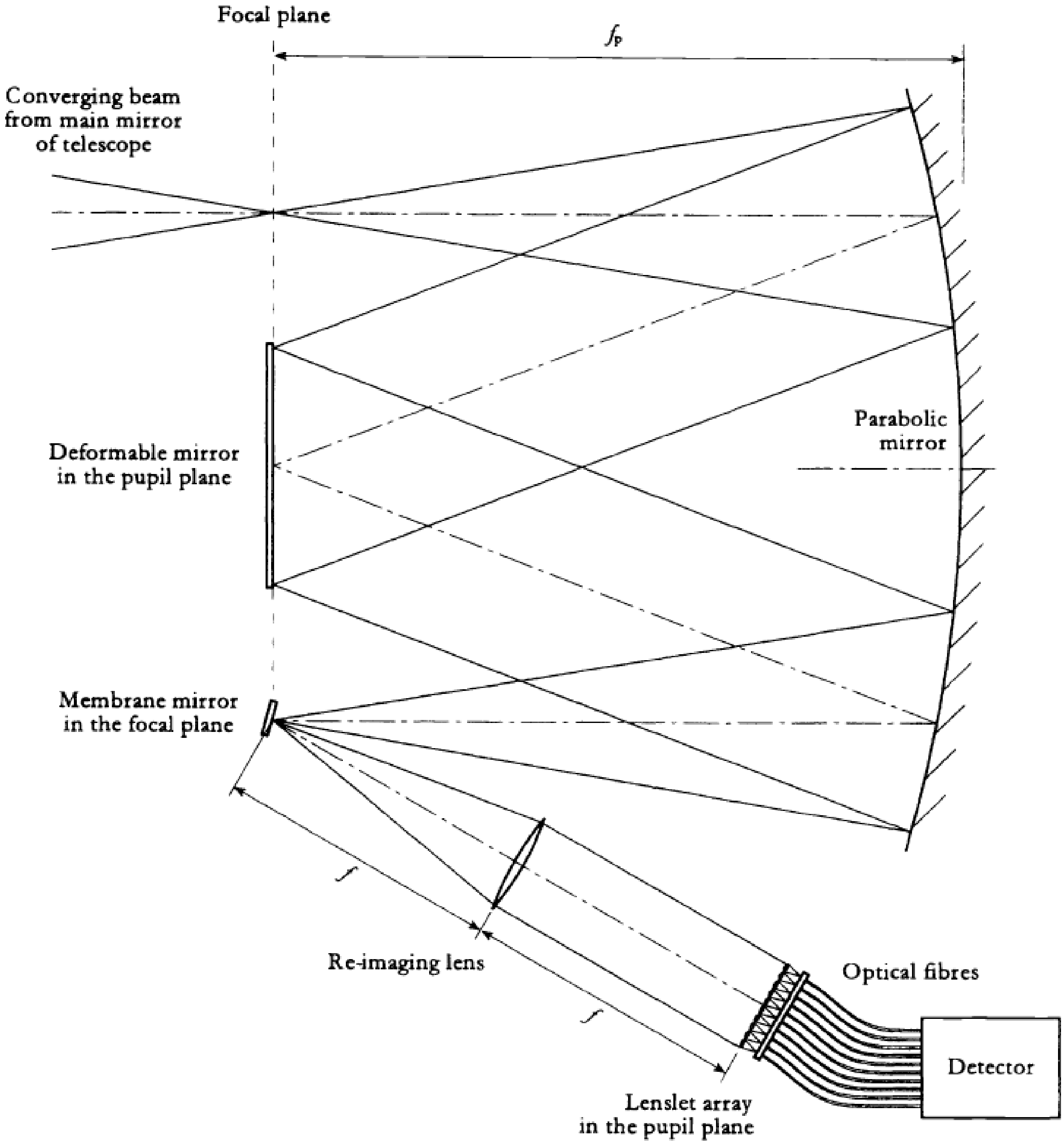

Focal plane
f_P
Converging beam from main mirror of telescope
Deformable mirror in the pupil plane
Parabolic mirror
Membrane mirror in the focal plane
f
Re-imaging lens
f
Optical fibres
Detector
Lenslet array in the pupil plane

Figure 8 – Optical layout of a curvature wavefront sensor-based AO system (Craven-Bartle, Dorn and Beletic 2000)

AO applications benefit from detectors with many pixels, high frame rates, and low noise. More pixels means finer spatial sampling; high frame rates mean finer temporal sampling; lower noise means more precise optical adjustments. Unfortunately for CCD imagers (the main alternative for AO applications), these three traits are generally mutually exclusive. A higher frame rate decreases the image latency but increases the read noise of a CCD. There are alternative methods, e.g. a multiple readout architecture (small sections of the larger array are read out individually and then combined), but this increases the complexity of a system as well as its power usage.

Because of the gain characteristics of GM APDs, they are capable of sending a digital signal to the readout circuitry by in-pixel digitization, eliminating read noise. When read noise is inconsequential, high frame rates and high pixel counts do not affect the noise level of measurements. This ability makes GM APDs ideal for wavefront sensing applications (Aull, et al. 2010).

2.1.2.2 LADAR

LADAR uses the time of flight from a transmitter laser to calculate the distance to and placement of distant objects in three-dimensional space, most commonly in airborne applications. The system must be capable of fine time resolution, high frame rate, and benefits from high spatial resolution (high pixel count). Since the detector is not imaging the scene but rather the laser reflection off the scene, the detector must be able to detect low flux with high efficiency. The lasers generally output short wave (near) IR signals (e.g. 1.06 µm) with a very narrow filter over the detector, ensuring that the only photons collected are those originally from the laser (Yuan, et al. 2010). Critical operational parameters for LADAR applications include dark count rate (DCR) and photon detection efficiency (PDE), which establish the noise floor of the system, as well as crosstalk between pixels (spatial resolution constraint) and timing jitter (temporal resolution constraint). Additionally, because weight and size matter more in airborne applications than in others, higher operating temperatures and smaller sizes (as well as low power usage) are always welcome improvements. Another term, LIDAR (LIght Detection and Ranging) is commonly used to describe the same system function, though it implies a broader use of the electromagnetic spectrum than the near infrared or infrared. Figure 9 shows an example of 3D LADAR imaging (Ardt 2008). In this particular image, longer wavelength (redder) shades represent objects with more height (close to the aerial detector). Scenes imaged using LADAR generally take advantage of multiple capture perspectives to form a point cloud and render a 3D representation of the scene.

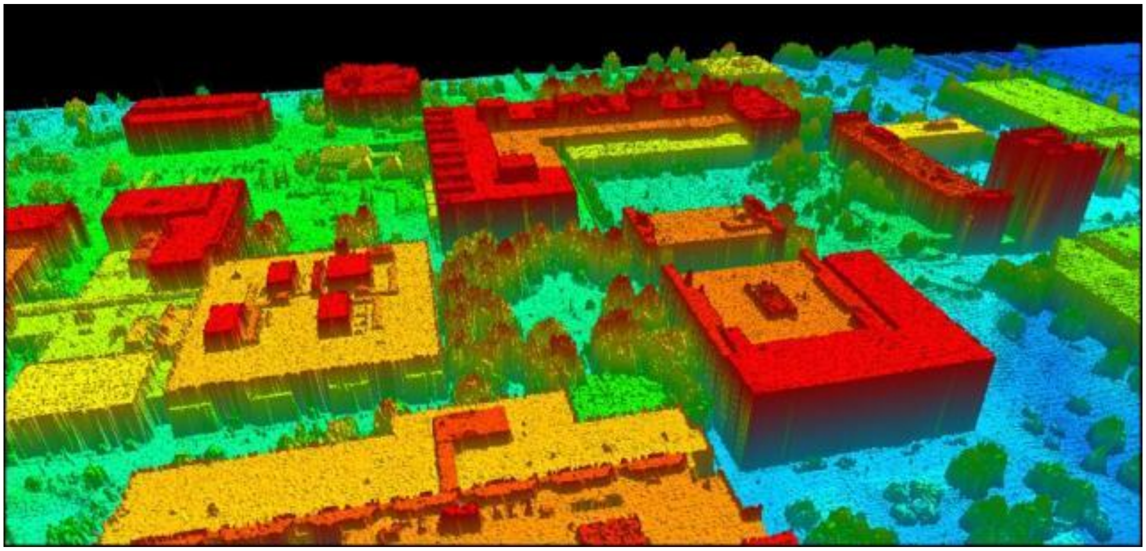


Figure 9 – 3D LADAR Image of a Portion of the RIT Campus (Ardt 2008)

2.1.2.3 Communications

Communication at telecom wavelengths for deep space and near-Earth applications is essential for extra-terrestrial exploration. Difficulties with the technology include the challenge of sending a signal over long distances in space (loss of information and flux from diffraction) as well as the tradeoff between lower energy signals (long wavelengths) and the detection efficiency of those signals. Lower energy photons (IR) travel well through space, but they are generally more difficult to detect and the materials required to do so are more expensive. These obstacles are compounded by the fact that detectors are often the limiting factor in a communication system's performance, making low noise and high efficiency desirable. Wavelength manipulation via waveguides offers a solution by allowing for the use of low-energy signal photons desired for space-based telecom with silicon-based detectors. An example of such a system is depicted in Figure 10 (Grein, et al. 2010).

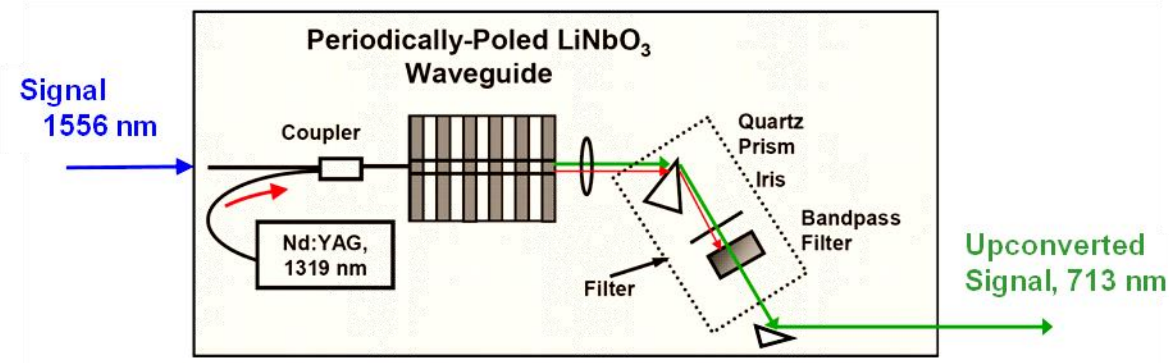


Figure 10 – Wavelength upconverter based on periodically-poled lithium-niobate (PPLN) (Grein, et al. 2010)

The waveguide (with Periodically-Poled $LiNbO_3$ – PPLN), which is only 48 mm long and temperature-stabilized, facilitates the conversion of 1556 nm photons to 713 nm photons using a 1319 nm pump (the original 1319 nm pump signal is filtered out via a bandpass filter before the signal reaches the detector). A silicon GM APD detects the up-converted signal at 713 nm, where silicon has high detection efficiency. The system has the potential to reach greater than 90 % efficiency, and the noise produced from the pump (evidenced in an increase in dark counts) was not the limiting noise factor of the system. Detector parameters to consider in such a system include the DCR, PDE, and timing jitter (which in this application constrain the data rate) (Grein, et al. 2010).

2.1.2.4 Other Applications

Aside from the applications already listed, GM APDs can function as photographic imagers (imaging a scene rather than a discrete signal). They offer low noise, precise photon counting for the most accurate representation of a scene, and anti-saturation capabilities with CMOS readouts (and accompanying coded logic during integration). GM APDs can also be fabricated with resolution as high as any other available detector.

2.2 DEVELOPMENT OF GM APD STRUCTURE

Understanding the internal device architecture of the GM APDs under study is vital to the characterization of the influence of various dark count contributions. There are distinctive elements of the internal structure designed to improve the detector's SNR, including separate absorption and multiplication regions and scupper regions. The challenge has been to successfully replicate the low-noise characteristics of the LFF (low fill factor) device in the HFF (high fill factor) device. So far, this task has been met with difficulty due to an unexpected and uncharacterized source of dark counts that is particularly dominant in the HFF device.

2.2.1 *DOPING PROFILE DESIGN*

The doping profile in these devices refers to the doping layer separating the absorption and multiplication regions, allowing for separate biasing conditions over the two regions, and varies between the LFF and HFF designs. The LFF doping profile is the shape of a small disk, while the HFF doping profile is a stepped plane, meaning that the layer is not at the same depth throughout the device, but is continuous. Figure 11 shows a vertical cross-section of each of the designs (Aull, et al. 2010). Two adjacent pixels are shown in each case.

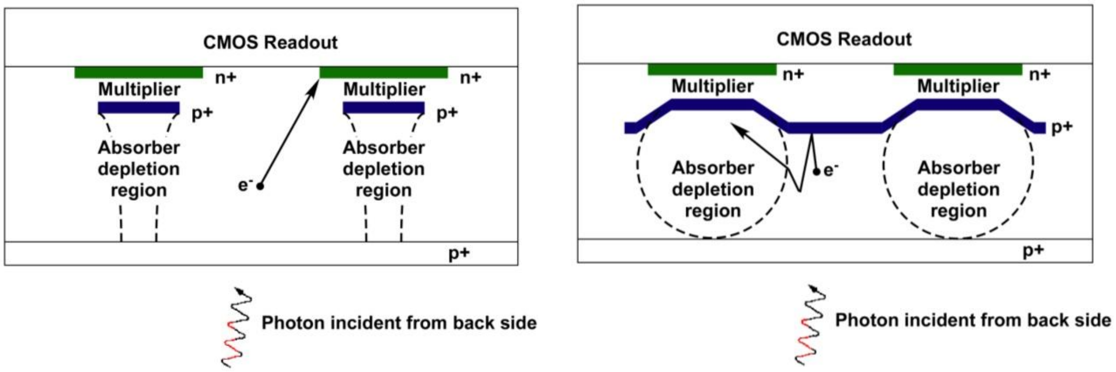

CMOS Readout
n+
Multiplier
p+
Absorber depletion region
e^-
n+
Multiplier
p+
Absorber depletion region
p+
Photon incident from back side
CMOS Readout
n+
Multiplier
n+
Multiplier
p+
Absorber depletion region
e^-
Absorber depletion region
p+
Photon incident from back side

Figure 11 – Doping Profiles of LFF (left) and HFF (right) devices (Aull, et al. 2010)

In both cases, a positive voltage is applied to the green n+ doping region (the areas in between the doping layers may be considered intrinsic) and the white p+ region at the bottom of the diagrams is grounded – this reverse-biases the APD. The arrows in both diagrams represent the path a free carrier might take if generated at the origin point denoted by e^-. In the LFF device, the carrier is likely to travel to the collection node without entering the multiplier region (avoiding an avalanche initiation event). The absorber region is very small, however, which means an array of these pixels would only use a small portion of the incoming signal. This feature does well in reduction of dark carrier noise, but it also limits the detector's ability to count the signal carriers. In the HFF device, there is a continuous doping step that makes it unlikely that the carrier would travel anywhere but the absorber region. The carrier is likely to be swept into the multiplier and may cause an avalanche. Though the absorber region allows an array of HFF pixels to capture a large percentage of the incoming signal, it comes at the cost of higher dark count rates due to the loss of the large scupper region, discussed below (see Figure 12).

2.2.1.1 LFF Device

The electric field distribution of the LFF device is shown in Figure 12. The medium low field around the periphery of the device acts as the scupper region, collecting carriers generated outside the absorber region without allowing them to pass into the multiplier region. The low field area in the absorber region moves carriers to the multiplier region, where the high field reaches the critical level necessary for a free carrier to initiate an avalanche (potential difference is above V_{BR}). For scale, the diameter of the n+ region is 60 µm and the diameter of the p+ region is 46 µm for the device characterized here (though these dimensions are much larger than a typical imaging device).

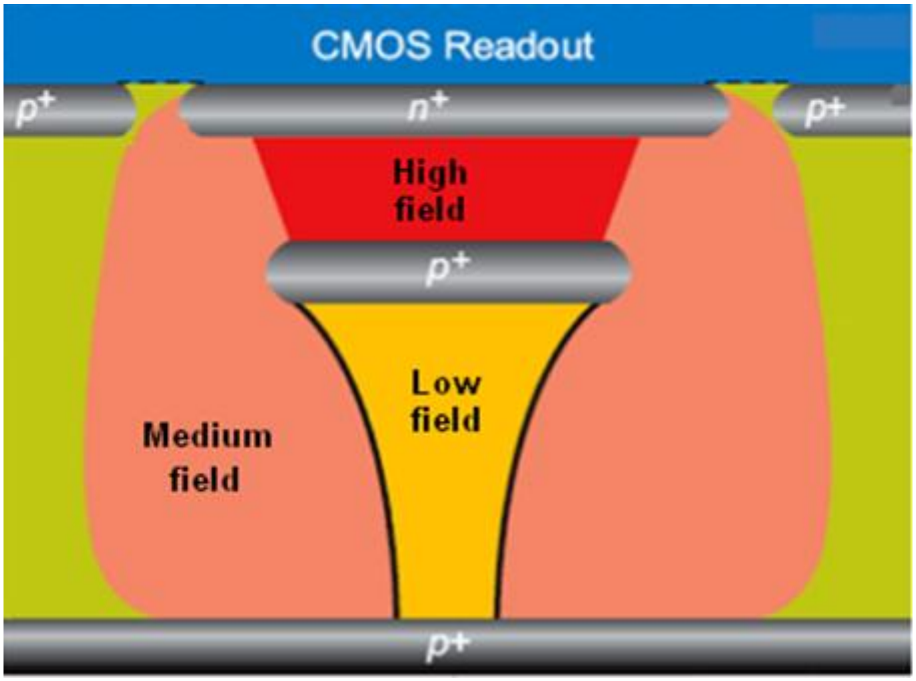


Figure 12 – Electric field distribution of LFF device architecture (Aull, et al. 2010)

2.2.1.2 HFF Device

Figure 13 shows the electric field distribution of the HFF device. The diameter of the n+ region is still 60 μm, as in Figure 12, but the scupper region seen in the LFF device architecture is no longer present since the dividing p+ doping layer is continuous. As shown in Figure 11, this leads to more free carriers entering the multiplier (high field) region, causing more dark counts than would be expected even if the LFF absorber area were scaled to the size shown in the HFF device. Another potentially problematic feature is the extremely high field found at the edge of the device region (denoted in Figure 13 by field lines). The field is not high enough to reach the critical level necessary for avalanche initiation, but it is high enough to facilitate tunneling (which will be discussed later), a source of dark counts. Carriers that become free due to the effects of tunneling may enter the multiplier region and initiate an avalanche (Aull, et al. 2010).

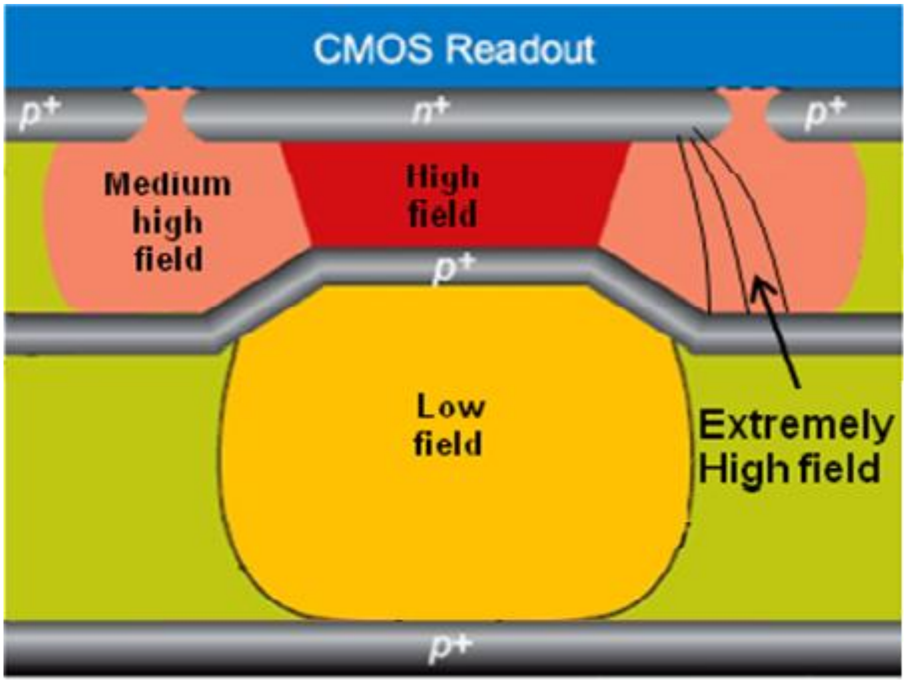

CMOS Readout
p+
n+
p+
Medium high field
High field
p+
Low field
Extremely High field
p+

Figure 13 – Electric field distribution of HFF device architecture (Aull, et al. 2010)

2.2.2 *PREVIOUS TESTING RESULTS*

Prior to the work described in this thesis, testing at MIT Lincoln Laboratory characterized the difference in DCR over varying quench time values (methodology discussed in Chapter 3) for different device architectures. The results are shown in Figure 14.

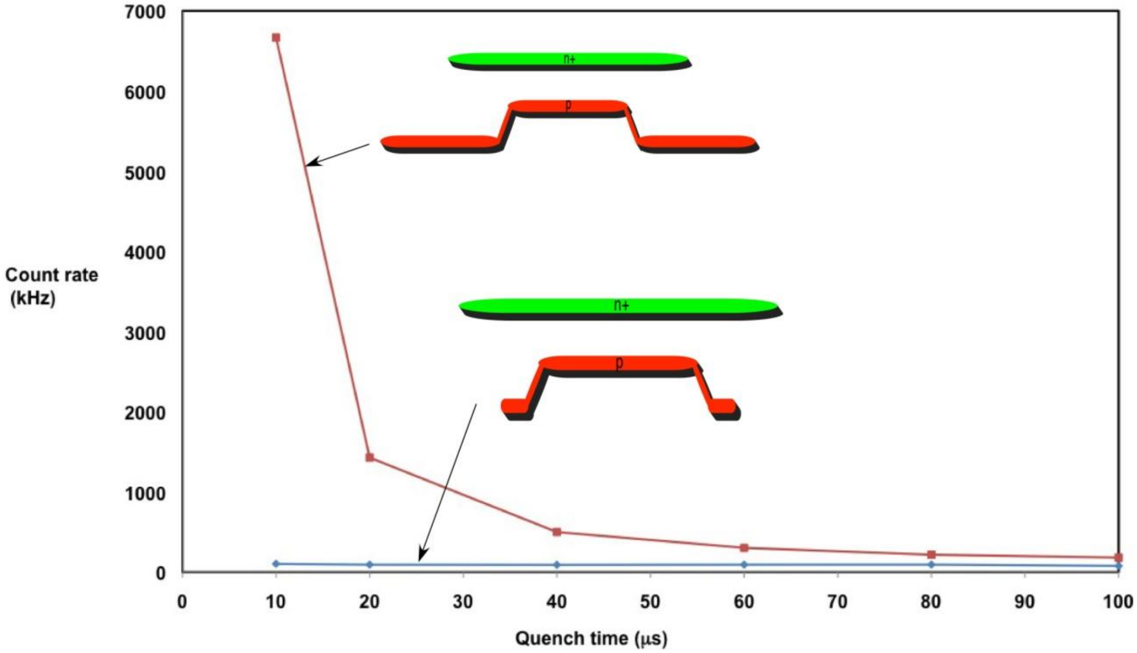


Figure 14 – Mean DCR vs Quench Time for both HFF (red) and LFF (blue) devices

On this scale, the LFF device (blue curve) appears to have a DCR that is independent of quench time, but it does have a shallow increase at the shortest quench times. This effect is dwarfed, however, by the extreme increase in DCR for the HFF device (red curve) at shorter quench times. Although afterpulsing does have a dead-time-dependent dark rate, the measured trend must incorporate contributions from more than just the afterpulsing mechanism: self-retriggering (discussed later) must be significant. This hypothesis is the main motivation behind the deeper investigation discussed here.

3 EXPERIMENTAL APPROACH

Characterization of the self-retriggering mechanism is the main motivation of this thesis. The previously measured increase in DCR at short quench times for the HFF devices cannot be explained by afterpulsing alone. Self-retriggering (discussed later in more depth) describes a mechanism in which avalanching carriers emit photons, and these photons generate free carriers that can cause an avalanche some time later. Since the exact nature and magnitude of contribution of this mechanism is the goal of characterization, any experimental approach that attempts to characterize it must yield results that are both statistically relevant and informative to the hypothesis at hand. This end requires careful planning of the data acquisition and analysis methods.

The common measurement setup for all experiments includes a 6” un-thinned wafer with many testing structures and devices of various architectural designs. This project focuses on one specific set of test devices, chosen based on the fill factor variations within the test area and the central location on the wafer (which is generally the best-behaved region). A representation of the test area is shown in Figure 15.

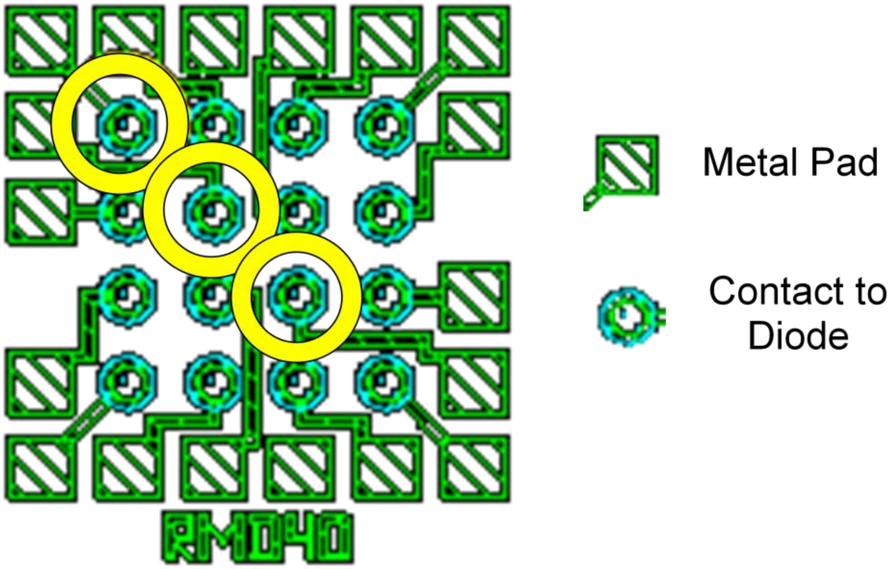


Figure 15 – Test device area geometry. The circled contacts show the devices that were tested; device architecture varies from low to high going left to right and top to bottom. For scale the metal pads are approximately 70 µm x 70 µm

The LFF device is in the top left corner, while the HFF device is located at the bottom right. The device in the center has a medium fill factor (MFF). These devices vary by the shape of the doping profile separating the multiplication and absorption regions. Recall from Figure 11 that the HFF device has a stepped layer of dopant, whose deep portion extends beyond the edge of the junction, while the LFF device lacks the step. The MFF device tested here has a stepped layer of dopant, but it does not extend beyond the junction. This results in a present but weakened scupper region. Each diode has its own metal connection and metal pad for contact with a probe needle. The back of the wafer (substrate contact) is grounded via the probe station stage, and the probe needle provides the diode biasing (a positive voltage applied to the diode's metal pads applies a reverse bias to the diode junction).

The mode of bias and measurement of the diode's output are specific to the testing mode. Passive quenching allows the external testing circuit to quench the avalanche and recharge the diode (via an in-series resistor in parallel with the cable capacitance). The circuit requires a single voltage source, and the oscilloscope samples the output between the diode and the in-series resistor (see Figure 16). Active quenching requires a significantly more sophisticated external circuit with logic ICs and multiple constant voltage source inputs. The quenching circuit monitors the output from the diode and produces digital pulses that are captured by an oscilloscope and transferred to an IDL program for display and analysis (see Figure 22).

3.1 PASSIVE QUENCHING

Practically applied, a passive quenching setup does not provide much information about the diode's performance. As a preliminary test, however, it confirms that the device is in fact a GM APD, and it facilitates data extraction of certain parameters via the recharge curve of the device. Figure 16 shows a representation of the passive quenching test setup. The diode is covered so that no stray light is collected and all signals are due to internally generated carriers (thermal, trap-induced, or photon self-retriggering – all of which will be discussed in detail).

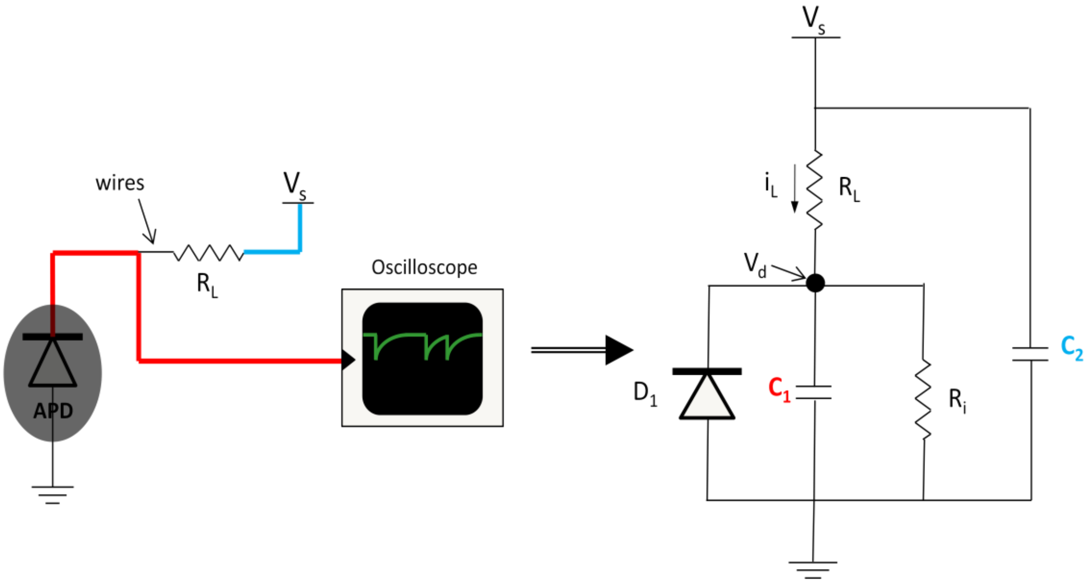


Figure 16 – Passive Quenching Setup and Equivalent Circuit Diagram

R_L represents the load resistor that recharges the APD after it has fired and turned off, C_1 represents the capacitance associated with the cable from the APD to the oscilloscope, and C_2 represents the capacitance associated with the cable from the constant voltage supply V_s to the load resistor R_L. R_i is the input impedance associated with the oscilloscope, i_L is the current through the load resistor at any time, and v_d is the diode voltage sampled by the oscilloscope.

3.1.1 *THEORY OF OPERATION*

There are three states of interest in the cycle of a passive quench circuit. In the first state, the GM APD is fully charged, or armed to a voltage above its breakdown voltage. The circuit is in stasis until an internally generated carrier triggers an avalanche (the next state). Until that happens, the equivalent circuit in Figure 17 describes the passive quenching circuit.

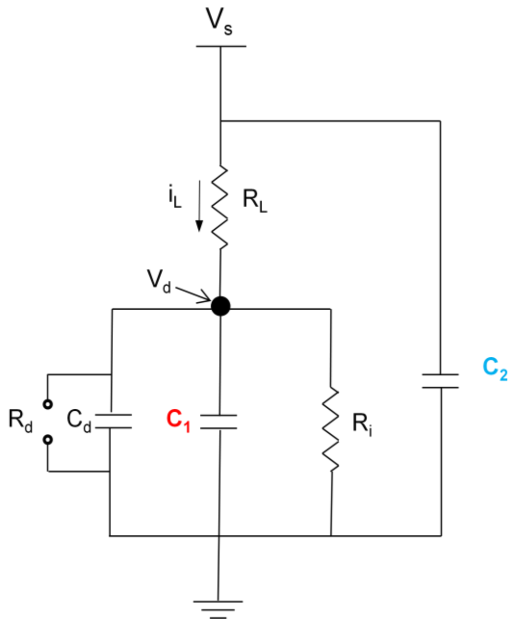


Figure 17 – Equivalent Circuit Diagram for State 1 of Passive Quenching Circuit Operation

C_d is the depletion capacitance of the diode. R_d is the equivalent resistance of the diode, which is ideally infinite (open circuit). In actuality, there is some leakage through the diode because of edge effects and dark current collected in the scupper areas, but the resistance is high enough to render the current through the diode's branch of the circuit negligible for the purposes of this analysis. Taking this into account, the voltage v_d and the current i_L can be described by Eq. 15 and Eq. 16, according to Ohm's Law.

$$\boldsymbol{v_d = \left(\frac{R_i}{R_i + R_L}\right) V_s} \qquad \text{Eq. 15}$$

$$\boldsymbol{i_L} = \frac{\boldsymbol{V_s} - \boldsymbol{v_d}}{\boldsymbol{R_L}} \qquad \text{Eq. 16}$$

The second state begins when a dark carrier triggers an avalanche. At this point, the GM APD draws exponentially increasing current through the load resistor until the carrier accumulation discussed previously causes the current to saturate. However, the supply voltage Vs divided by the load resistance R_L also limits this current (by Ohm's Law). The current i_D is instantaneously dynamic as the GM APD avalanches, and the GM APD itself becomes the equivalent of a high-valued current source (very little resistance) in parallel with a capacitor. Figure 18 shows an equivalent circuit model for the second state of passive quenching.

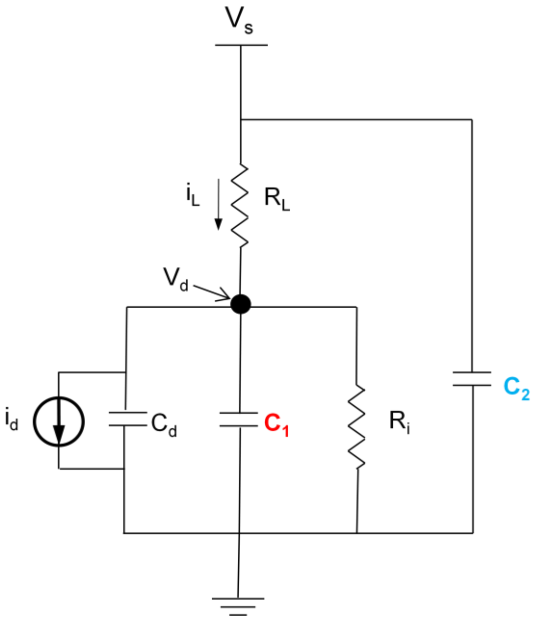


Figure 18 – Equivalent Circuit Diagram for State 2 of Passive Quenching Circuit Operation

i_d is the current through the diode during avalanche. This state represents the circuit in the time when the diode voltage at v_d is still greater than the breakdown voltage. The value of the voltage v_d can be described as in Eq. 17, where i_d grows exponentially during the avalanche, decreasing v_d because the increase in current through the load resistor also increases the voltage drop across the load resistor.

$$\boldsymbol{v_d = V_s - i_d R_L} \tag{Eq. 17}$$

Once v_d decreases below the breakdown voltage (or the current through the APD is no longer sufficient to replace the collected carriers), the free carriers in the multiplication region of the diode will no longer have enough energy for a self-sustaining avalanche. This brings the circuit into the third stage of operation, when the diode recharges due to the current flowing through R_L and the cable capacitance C_1. The voltage across the depletion region of the diode increases to the steady level in the first state. The equivalent circuit for third state is the same as for the first state (see Figure 17), but the initial conditions differ. Thévenin's Theorem simplifies the analysis of the circuit via the creation of a Thévenin equivalent circuit from the perspective of the output (Jaeger and Blalock 2004). Figure 19 shows the equivalent circuit and the relevant portions of it used for Thévenin equivalent analysis.

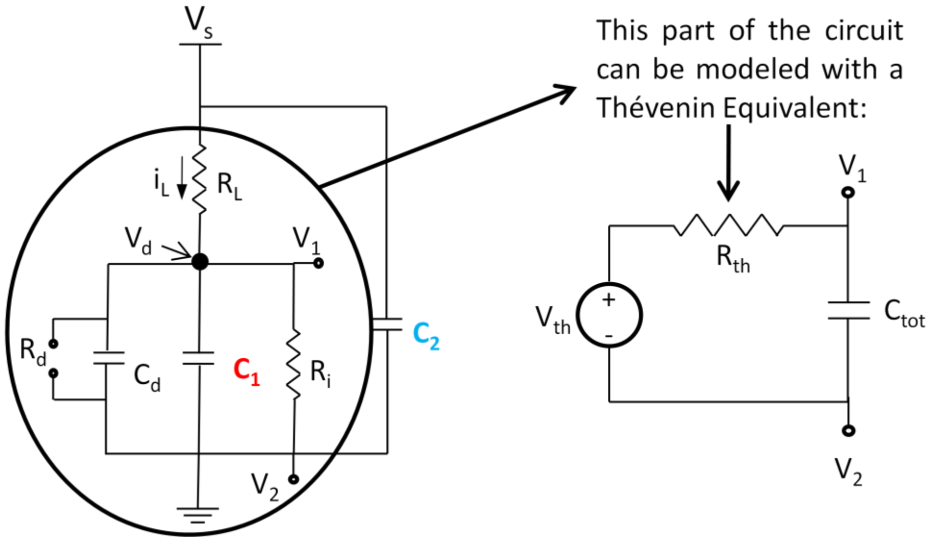

This part of the circuit can be modeled with a Thévenin Equivalent:

Figure 19 – Equivalent Circuit Diagram for State 3 of Passive Quenching Circuit Operation with considerations for the Thévenin equivalent circuit

R_{th} and V_{th} are the Thévenin equivalent supply voltage and resistance, respectively, for the circuit portion included in the circle in Figure 19. Because the supply voltage is constant throughout the experiment, the model ignores C_2. Though it charges and introduces a dynamic when the supply voltage turns on, the supply voltage is constant throughout the experiment.

Since current through a capacitor is equal to the change in voltage over time, C_2 does not influence the behavior of the diode. C_{tot} is the sum of C_d and C_1 – two capacitors in parallel add linearly. The components of the Thévenin equivalent circuit are described in Eq. 18 - Eq. 20.

$$\boldsymbol{V_{th} = V_s * \left(\frac{R_i}{R_i + R_L}\right)} \qquad \text{Eq. 18}$$

$$\boldsymbol{R_{th} = \left(\frac{1}{R_i} + \frac{1}{R_L}\right)^{-1}} \qquad \text{Eq. 19}$$

$$\boldsymbol{C_{tot} = C_d + C_1} \qquad \text{Eq. 20}$$

Eq. 21 describes the behavior of the Thévenin equivalent circuit, which is conveniently in the form of a simple RC circuit (Jaeger and Blalock 2004).

$$\boldsymbol{v_d(t) = V_{BR} * e^{\frac{-t}{C_{tot} \cdot R_{th}}} + V_{th}\left(1 - e^{\frac{-t}{C_{tot} \cdot R_{th}}}\right)} \qquad \text{Eq. 21}$$

V_{BR} is the breakdown voltage of the diode. Known boundary conditions can verify the resulting equation, based on the passive quenching circuit component values. At time $t = 0$, the voltage is equal to V_{BR}, as the diode will have just stopped avalanching because the voltage no longer sustains it. Conversely, at $t = \infty$, the voltage returns to the steady-state voltage described in Eq. 15, which is also the definition of V_{th} (Eq. 18). In both boundary condition cases, Eq. 21 gives the expected results and so it is valid.

A PSPICE software circuit simulation validated the equation with respect to the equivalent circuit (see Figure 17). Since the form of Eq. 21 is a known solution to the RC circuit differential equation, the crux of validation is whether the calculated Thévenin equivalent RC value is correct. Figure 20 shows the simulation output for a circuit with $C_d = 30$ pF, $C_1 = 35$ pF, $V_{BR} = 29$ V, $V_s = 36$ V, $R_L = 200$ kΩ, and $R_i = 1$ MΩ.

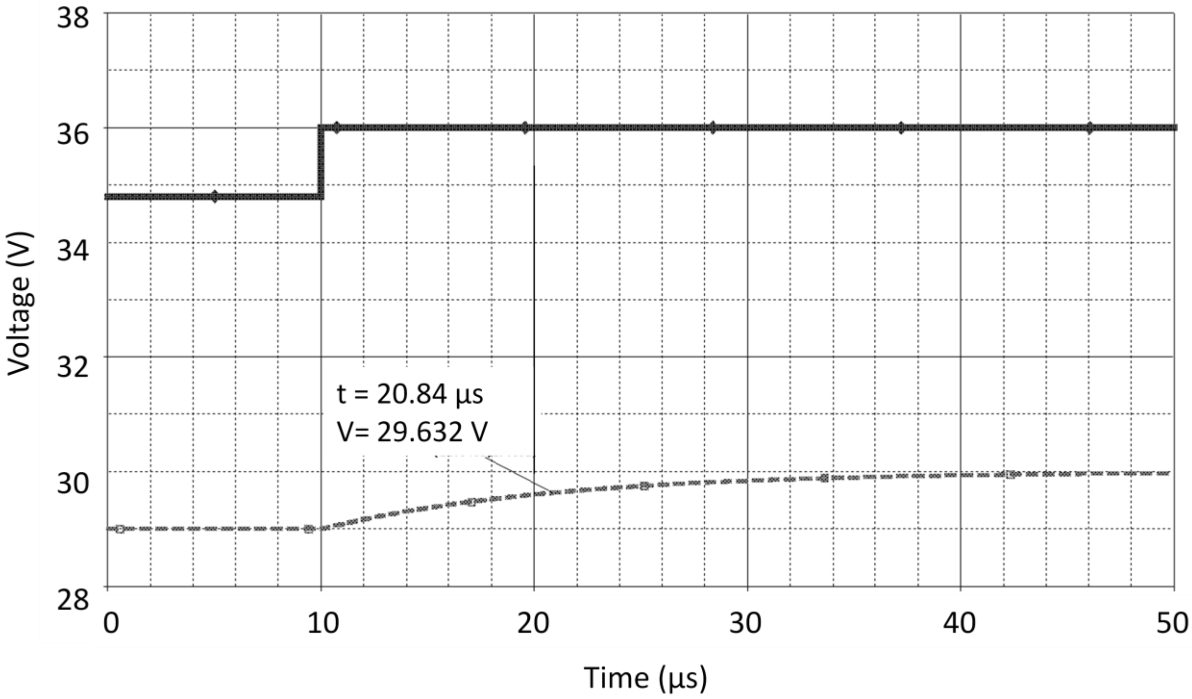


Figure 20 – PSPICE Equivalent Circuit RC Recharge Simulation Results

The stepped supply voltage waveform simulates the voltage over the capacitor without specifying the initial current through the GM APD during avalanche, but the resulting effect on the circuit is the same. The initial supply voltage setting is such that the initial voltage across the diode is equal to the breakdown voltage, while the step voltage brings the voltage across the diode to the steady-state level. At time t = RC, the voltage is equal to the initial diode voltage plus 63.2 % of the maximum change in voltage (1 V) since the exponential argument reduces to $e^{-1} = 0.632$. According to Eq. 21, RC = 10.83 µs and $v_d(RC) = 29.632$ V. Since there is a built-in delay of 10 µs in this simulation, a voltage of 29.632 V is expected across the diode at time t = 20.83 µs, which is precisely the result.

3.1.2 *MEASURING PASSIVE QUENCH RC RECHARGE*

Since the only unknown variable in Eq. 21 is C_d (all other elements are measureable), calculations based on measured data can estimate its value. Figure 21 shows a screen capture from the oscilloscope used to monitor diode voltage as shown in Figure 16.

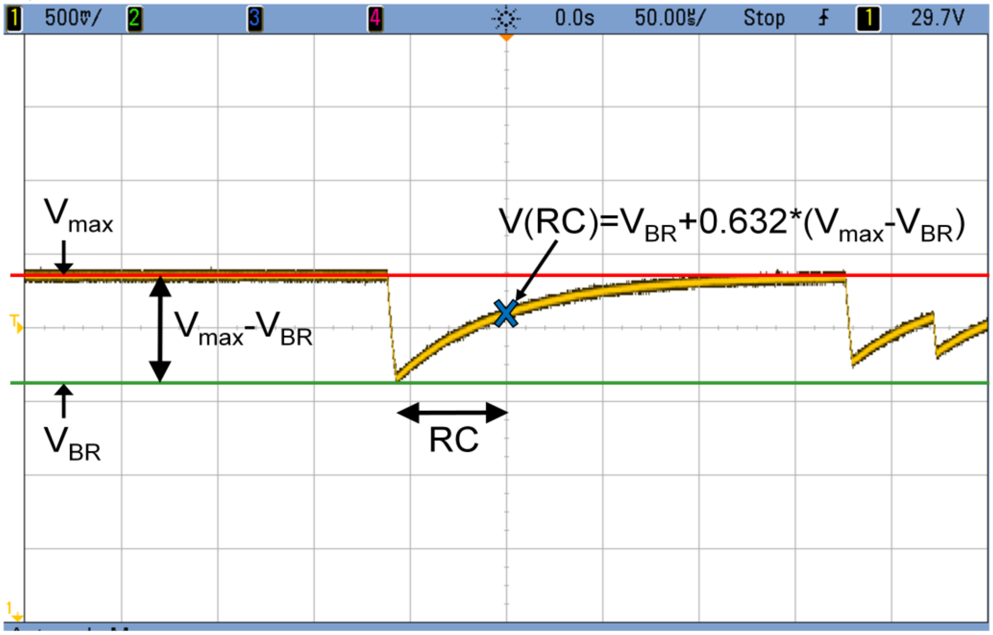


Figure 21 – Oscilloscope Screen Capture of Passive Quenching Diode Output

The “x” marks the point where the voltage is equal to the base (breakdown) voltage plus 63.2 % of the maximum change in voltage. Since RC is measurable on the waveform, and R is a known parameter of the testing setup, the sum of C_1 and C_d (C_{tot}) must be equal to RC/R. Since the capacitance of the device is very small, the cable capacitance C_1 will dominate the RC constant, and so a simple calculation based on the circuit model and measured or chosen values will yield the correct time constant.

The recharge curve shown in Figure 21 is ideal because it follows a long period of steady-state recharge and precedes a full recharge curve. The portion of the waveform immediately following that curve illustrates one that is unfit for this analysis. It shows an avalanche after the preceding condition was not yet in a steady state, resulting in a slightly higher minimum voltage. As the diode is recharging in the second curve, another avalanche occurs before full recharge completes and so the voltage drops again.

3.1.3 *DIODE CAPACITANCE*

As a basis of comparison for any diode capacitance measurements made, a theoretical calculation of the GM APD's junction capacitance is necessary. Often, a parallel place capacitor sufficiently describes a p-i-n diode depletion capacitance. Calculation of the capacitance between two plates is generally simple if certain assumptions hold. First, the plates must be assumed to be infinitely large (or large enough so that edge effects are not significant to the total capacitance), and secondly the plates must be of equal size so that the cross-sectional area of the capacitor is constant throughout its depth. Eq. 22 shows the standard formula for the cases where these assumptions hold.

$$\boldsymbol{C = \frac{\varepsilon_0 \varepsilon_r A}{d}} \qquad \text{Eq. 22}$$

ε_0 is the permittivity of free space (a constant), ε_r is the relative permittivity of the capacitive material (equal to the ratio of the material's permittivity to that of free space), A is the cross-sectional area of the capacitive region, and d is the distance between the two parallel plates.

Unfortunately, neither of the assumptions above is truly valid for the architecture of the depletion region that exists in the GM APDs in this project. A more correct expression is derived in Appendix B, though the details are spared here. The resulting expressions are shown in Eq. 23 - Eq. 25.

$$\boldsymbol{C = \frac{\varepsilon_0 \varepsilon_r \left(\pi {R_B}^2\right)}{d} + \varepsilon_0 \varepsilon_r \int_{\emptyset=0}^{2\pi} \int_{\theta=0}^{tan^{-1}\left(\frac{R_A - R_B}{d}\right)} \frac{a(\theta, \emptyset)}{h(\theta)} d\theta d\emptyset} \qquad \text{Eq. 23}$$

$$\boldsymbol{a = \emptyset (R_B + d \tan\theta) d \tan\theta} \qquad \text{Eq. 24}$$

$$\boldsymbol{h = \frac{d}{\cos\theta}} \qquad \text{Eq. 25}$$

R_B is the radius of the dopant implant separating the multiplication and absorption regions (see Figure 11), $\emptyset$ is the angle around the capacitor in cylindrical coordinates, θ is an angle that describes the cross-sectional area of a component portion of the capacitor, and the upper limit of integration for θ is based on fixed the geometry of the diode.

Even at the roughest level of approximation, however, it makes sense that the actual capacitance of the diode falls between the two capacitances as calculated with Eq. 22 using (respectively) the areas of the p+ and n+ implants alternately as the cross-sectional area values (see Figure 11).

3.2 ACTIVE QUENCHING

The active quenching setup provides the basis of data collection that measures dark count rates and, by changing specific settings in the setup, aids in the characterization of certain dark count mechanisms. Figure 22 shows a diagram of the active quenching setup. As in the passive quenching setup, a covering over the diode ensures that no stray light affects the dark count rate as measured.

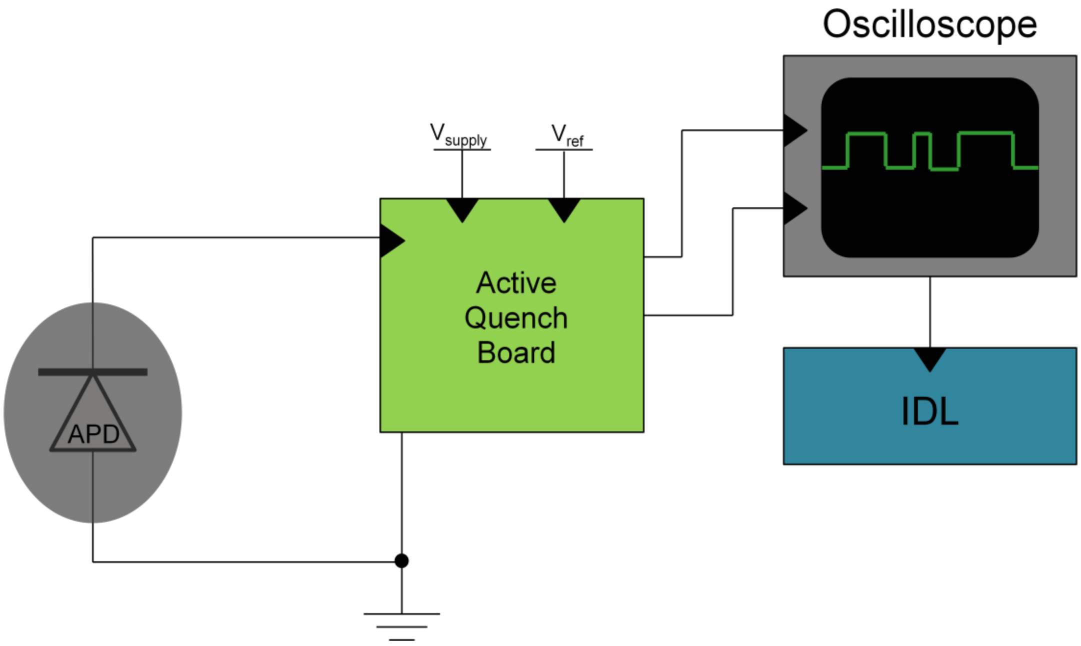


Figure 22 – Diagram of Active Quenching Test Setup

The active quench board (AQB) is a logic circuit that both monitors and sets the voltage on the APD depending on the state of the APD (stasis, avalanching, quenching). A comparator inside the AQB actively checks the voltage on the APD against a pre-set threshold voltage (V_{th}). As soon as the diode voltage drops below V_{th}, the circuit actively clamps the voltage to a low level called the quench voltage (V_q). The circuit holds the voltage at V_q for some pre-set duration called the quench time (t_q). At the end of t_q, the circuit enables a re-arm signal and brings the voltage up to some pre-set level called the arm voltage (V_a). In order to steady the circuit elements, the arm pulse holds the APD voltage at V_a for some pre-set duration called the arm time (t_a). At the end of t_a, the diode voltage is no longer actively set, and the circuit begins monitoring the voltage for a decrease below V_{th} and the cycle repeats. V_{th}, V_q, t_q, V_a, and t_a are tunable by the adjustment of five separate potentiometers located on the AQB. The circuit has two digital outputs, called quench and arm. Each produces a logic pulse when the APD is being quenched and armed, respectively.

The oscilloscope monitors the output from the quench and arm signals. During setup, the oscilloscope samples the APD signal to make sure the circuit is working properly, but does not monitor it during active measurements. The APD sampling probe introduces a leakage drain on the APD voltage, which causes the AQB circuit to prematurely quench the APD (even if no avalanche has occurred).

A computer connects to the oscilloscope via USB. IDL interfaces with the oscilloscope and collects the buffered data (the data shown on the oscilloscope screen) as a series of time and voltage coordinate pairs. A higher-level program (also coded in IDL) uses this coordinate array to calculate the time at which the first avalanche occurred after the APD was re-armed. The IDL routine records that time and repeats the process as many times as the user defines. Once the measurement set is complete, another IDL routine uses the histogram of arrival times to characterize the performance of the APD.

Figure 23 shows a screen capture where all three signals (V_a, V_q, and V_{APD}) are viewable. The purple waveform is the diode voltage signal, the quench voltage signal is in orange, and the arm signal is in green.

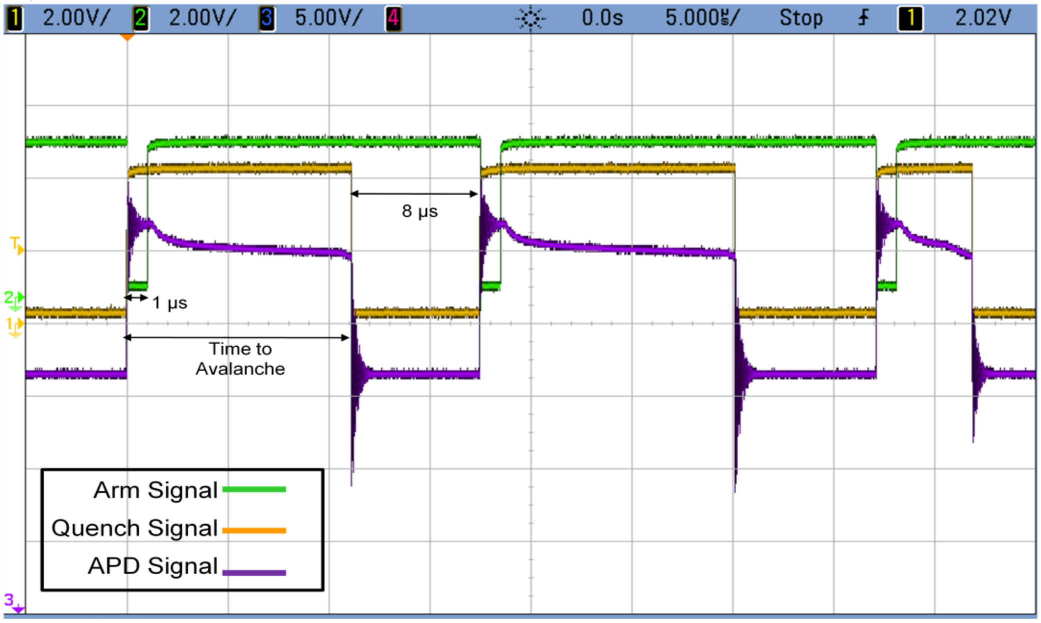


Figure 23 – Sample Waveforms for Arm, Quench, and APD Signals Under Active Quenching Operation

The figure contains a few important measurements. The time to avalanche (the measured arrival time) is the length of the quench signal waveform from the rising edge (which marks the start of re-arm) to the next falling edge (which marks the time at which the circuit senses the diode is in avalanche). The figure also shows that t_q is 8 µs for this experiment, and that t_a is 1 µs. As stated previously, the oscilloscope does not sample the APD signal during active measurements due to leakage effects.

3.2.1 *DARK COUNT RATE THEORY*

In a standard photodiode under reverse bias in the dark, there is a measurable leakage current, which is the flux of charged carriers (on average) traveling through a device per second. This same principle applies to a GM APD. Instead of inducing a steady current, however, some of these dark carriers induce avalanches. On very small time scales, an avalanche marks the arrival of each dark carrier as long as the APD has time to reset before the next carrier induces an avalanche. Therefore, the noise contribution from dark carriers in a GM APD is the rate of avalanches triggered in the device under dark conditions.

A few different mechanisms can produce dark carriers. The most prominent is thermal excitation, the rate of which is altered only by the temperature of the diode substrate. Other mechanisms include afterpulsing, (the release of carriers trapped in intermediate energy states), tunneling, and self re-triggering. In the case of the latter, photons produced when the APD avalanches generate carriers in the device.

Fortuitously, Poisson statistics (discussed later) govern the combination of dark carrier generation mechanisms inside of a device. This allows the experiment to sample avalanche arrival times without regard to order or time lapse between the measurements because, by definition, the events of a Poisson process are memoryless, or statistically independent of any prior event.

3.2.1.1 Thermal Dark Count

As long as the temperature of a substrate is not absolute zero, the carriers inside of a device with a pn junction will have thermal energy resulting in minority carrier diffusion, or random motion. Under reverse bias conditions, minority carriers from both sides (p-side electrons and n-side holes) can wander into the depletion region and accelerate across it. At a bias level below V_{BR} this amounts to reverse bias leakage current, some of which can be multiplied, but in a GM APD this can result in false detection events (Pierret 1996).

The amount of minority carrier drift across the junction is essentially independent of the reverse bias across the junction, since its source is limited (Pierret 1996). However, even if thermal dark carriers are the only contributors to the dark count rate (DCR), the as-measured DCR will increase with an increase in bias over V_{BR} because the avalanche initiation probability will increase (see Eq. 12 and Eq. 13) (McIntyre 1973). While the number of carriers is not changing, the probability that those carriers will initiate an avalanche is. The thermal dark carrier contribution of a given volume is only dependent on temperature. The carriers have less energy at lower temperatures and therefore their random motion decreases. In this particular instance, however, increasing the bias across the device increases the area of the absorber and multiplier regions, thereby increasing the overall volume where thermal dark carriers can be immediately collected.

3.2.1.2 Afterpulsing

Imperfections in semiconductor manufacturing or molecular contamination of the substrate, along with other sources, can generate energy states that exist in the band gap (see Figure 1). These trap states can occur at physical boundaries between materials in a device, in bulk regions if a foreign molecule like sodium (a common contaminant) has diffused there, or at a crystal lattice imperfection. In the case of a GM APD in avalanche, some carriers will become trapped in these intermediate energy states and then released later on (Pierret 1996).

The average time it takes a carrier to leave a specific trap (become free) is the detrapping lifetime. Afterpulsing occurs when carriers become trapped during the avalanche state of the APD and are released later on. If the circuit does not sustain the quench for a sufficient length of time (in the case of active quenching), then the carrier can de-trap after the APD is re-armed, inducing another avalanche. The time intervals between avalanches caused by afterpulsing are generally much shorter than those associated with thermal dark carriers, and so the measured DCR is much higher when afterpulsing is significant.

The contribution of afterpulsing decreases significantly with longer quench times because any carriers released while the APD voltage is low pass through the multiplication region without initiating an avalanche. As long as the quench time is significantly longer than the detrapping lifetime associated with the type and number of traps in the device, the afterpulsing contribution will be negligible.

3.2.1.3 Band-to-Band Tunneling

Band-to-band tunneling is feasibly a contributing factor in HFF devices where V_{BR} around the edges of the diode is smaller than at the center and the electric field is much higher (see Figure 13). In this quantum-mechanical process, a carrier combines with an empty state on the other side of the band gap of the material. The tunneling phenomenon is completely quantum in nature. There is some probability that the carrier will “jump” to the energy state on the other side of the band gap, which is a function of the energy of the particle, the band gap energy, the number of available states on the other side, and the amount of band-bending present due to the electric field (Pierret 1996). Figure 24 shows a visual representation of the band-to-band tunneling mechanism. Carriers from one side of the band gap can tunnel through to the other side and recombine. The structure pictured represents a (p+)-i-(n+) diode, which is not exactly the same as the actual structure being tested (which has a slightly p-doped region instead of an intrinsic region), but the principle is the same.

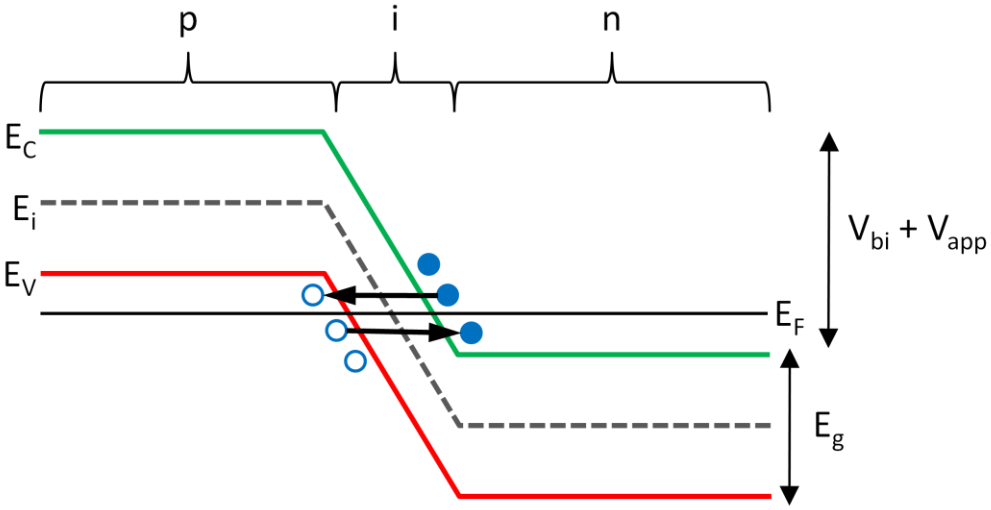


Figure 24 – Visualization of the Band-to-Band Tunneling Mechanism

Recall the band diagram for a single material type in Figure 1 and the various energy levels present. As shown in Figure 2, when a reverse bias is applied to a diode junction, the total voltage across the junction is equal to the built in voltage V_{bi} plus the applied voltage. For band-to-band tunneling to be significant, the band structure must undergo extreme bending (very high electric field), and the n+ and p+ sides of the diode must be very heavily doped (Pierret 1996).

3.2.1.4 Photon-Induced Self-Retriggering

While thermal dark count and afterpulsing contributions to DCR are well studied and generally understood, the mechanism here called self-retriggering is uniquely noticeable in the specific architecture of the diodes studied here. This necessitates that special attention be paid to characterization of the effect and its causes, so that its contributions may be limited in future iterations of the device architecture.

Recall the diode structure from Figure 12 that depicts the multiplication, absorption, and scupper regions of the device. The LFF device architecture will be used to illustrate the effects of photon-induced self-retriggering. Figure 25 shows the steps necessary for one of these carriers to induce a self-retriggering event. Stage 1 describes the detector during an avalanche. Stages 2 and 3 occur while the device is being quenched by the external circuitry, and stage 4 represents the initiation of another avalanche and the cycle repeats.

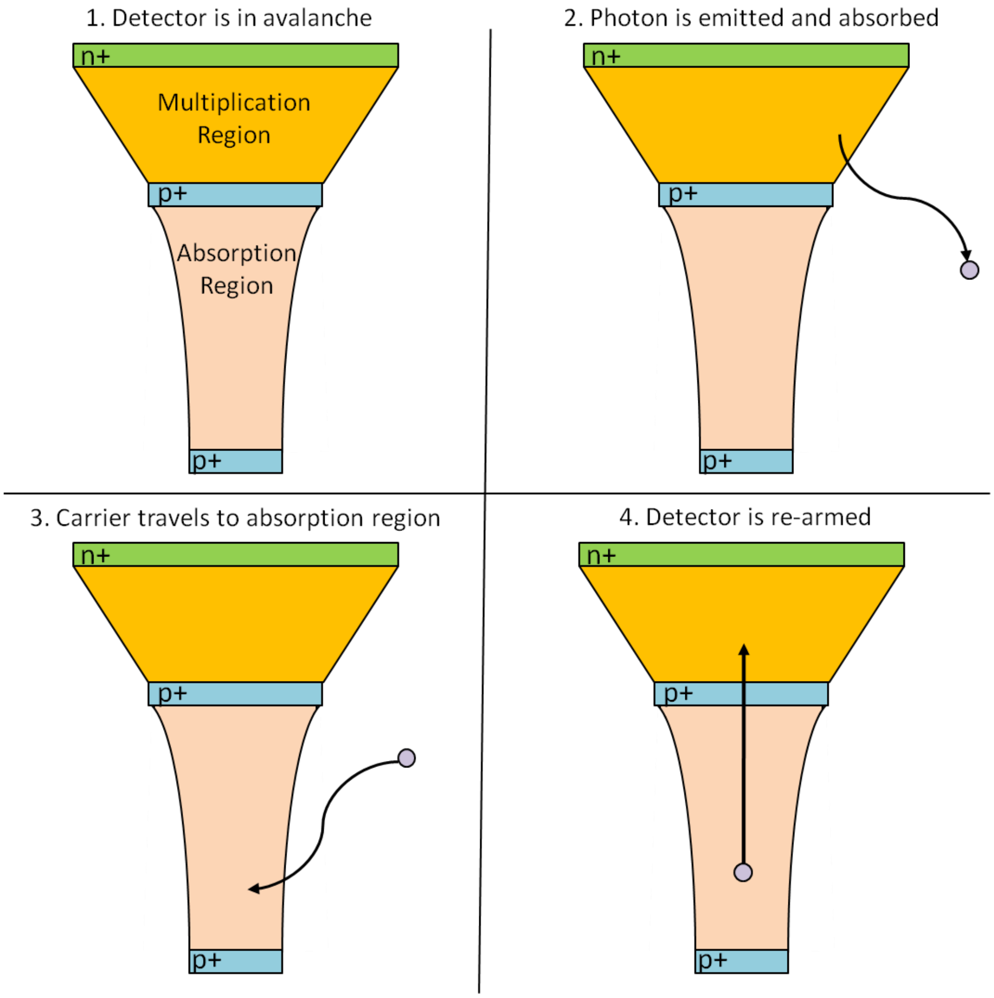

1. Detector is in avalanche
n+
Multiplication Region
p+
Absorption Region
p+
2. Photon is emitted and absorbed
n+
p+
p+
3. Carrier travels to absorption region
n+
p+
p+
4. Detector is re-armed
n+
p+
p+

Figure 25 – Self-Retriggering Mechanism in stages

The photon-induced self-retriggering theory states that, when an avalanche occurs, many excited carriers move through the device, and the subsequent relaxation (or recombination) of some carriers results in the emission of some number of photons. The photons emitted by the avalanching carriers (stage 1) can be absorbed by the surrounding substrate (stage 2). The substrate material and the energy of the photon determine the absorption depth, or the distance from photon generation to carrier generation. If these self-retriggering carriers are free in the multiplication region during avalanche, the electric field sweeps them to the collection node with the rest of the carriers and they are inconsequential. If, however, they are free in the absorption region or nearby while the device is armed (stage 3), they will wander about in the low field area until they reach either the scupper region or the multiplication region, or until the carrier recombines. If these carriers exist long enough (determined by the carrier lifetime associated with the device) to be present in the absorption region during or after the re-arm signal, the electric field will sweep them into the multiplication region and they could initiate an avalanche (stage 4). This type of avalanche would always register immediately, inflating the DCR by decreasing the mean arrival time of observed avalanches. Other carriers that are generated by self-emitted photons outside of the pixel area could re-enter the pixel in a subsequent cycle if the carrier lifetime is long enough, but since these carriers do not necessarily have a unique arrival time window, they are not as easy to characterize.

3.2.2 *DCR VS QUENCH TIME*

Given the distinct mechanisms governing the contribution of each type of carrier generated in an APD under dark conditions, a plot of the measured DCR vs the quench time associated with the active quenching setup helps to characterize the magnitude of contribution from each mechanism. For instance, if the DCR does not change significantly as a function of the quench time, then afterpulsing and self-retriggering are not significant contributors to the DCR. Conversely, a steep rise in DCR at short quench times indicates that afterpulsing is a problem in that particular device. When that rise is not consistent with afterpulsing alone, self-retriggering must be a contributor. In addition, the DCR measured at very long quench times will approach (and eventually equal) the thermal and tunneling dark count contributions, as any afterpulsing and self-retriggering will have subsided.

3.2.3 *FIRST BIN RESIDUALS*

Another means of characterization is the first bin residual calculation. In a histogram of arrival times, there are a large number of counts in the first bin histogram because of the way the AQB functions. The circuit holds the arm pulse for some time t_a, as described previously. Any carriers entering the multiplication region during this time will be subject to a field that surpasses the critical level necessary for avalanche initiation (by design), but the voltage across the APD is not allowed to decrease because the circuit is actively holding the voltage at V_a, which sustains the avalanche. If an avalanche initiated during the arm pulse, the voltage immediately begins to decrease at the end of the arm pulse. The time it takes for the voltage to decay plus the built-in delay of t_a equals the minimum arrival time that the active quenching setup can measure. Some carriers from both thermal and afterpulse contributions will arrive during the arm pulse, but the self-retriggering mechanism contributes heavily to immediate events, providing an opportunity to characterize its effects.

The first bin residual is calculated by determining the number of counts expected based on the trend associated with the rest of the distribution. Extrapolating the function fit across the arm pulse duration, then integrating over that time period, gives the number of counts that are expected in the first bin (if afterpulsing and thermal carrier generation were the only contributors). The residual is the actual value less the expected value. A positive residual indicates counts that must be due to the self-retriggering mechanism (with some error due to loss of resolution of very short detrapping lifetimes because of the length of the arm pulse), since many self-retriggering carriers make their contribution during the arm pulse according to the theory presented here. There are a few points that may be raised as to the error associated with the first bin measurement. These points are discussed to acknowledge their potential implications, but it should be noted those implications have been deemed small enough to ignore in the scope of this project.

First, there is unquantifiable error associated with the first bin residual measurements. There is no way to quantify any afterpulsing contributions from traps with detrapping lifetimes much less than the arm pulse. Since these traps may also induce immediate avalanches, but might have no measureable contribution after the arm pulse, they contribute to the first bin calculation even though they are not due to self-retriggering. The impact of these traps, however, would likely be small due to the zero time of trapping mechanism (the previous avalanche) and the nature of the detrapping time distribution. Any trap with a very short lifetime will also have a small carrier detrapping distribution because in a Poissonian process, the mean is equal to the standard deviation (full discussion below). On the other hand, any trap who's lifetime is long enough to register an avalanche after or during the arm pulse will have a standard deviation sufficient to populate at least some later arrival times as well (allowing the full histogram of the dark counts to account for that trap's contribution). While this point is important for any future work to address, the effects noted are considered negligible or small sources of error in the calculation of the first bin residual.

Secondly, there is no correlation recorded between the incidence of immediate avalanches. Since an immediate avalanche generally indicates that an event occurred during the arm pulse, the detector is allowed to avalanche for the rest of the duration of the arm pulse. This produces many more carriers than a non-immediate avalanche and so the self-retriggering generation of carriers is higher. Thus, it stands to reason that the next avalanche is more likely to be immediate if the previous avalanche was, and the strength of that correlation would be dependent on the length of the quench time. This relationship would be an interesting study in any future, more sophisticated iterations of data collection, but is not characterized here.

4 SIMULATION OF DCR BEHAVIOR OF GM APDS

In order to more fully understand the interaction of all dark count contributions, a program was written in IDL that simulates the arrival time of avalanches given various physical material characteristics and operational parameters based on the AQB circuitry and the measurement techniques used.

4.1 THEORY

For this type of analysis, it is important to completely grasp the statistics that govern carrier arrival times and how to manipulate raw data in a statistically significant way. Dark current carriers are governed by Poisson statistics (based on the law of rare events), which means that the probability distribution of the number of events per time window is governed by Eq. 26 (Kay 2005).

$$\boldsymbol{f(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}} \qquad \text{Eq. 26}$$

λ is the expected number of occurrences in a time window (some steady-state characteristic rate multiplied by the observation time) and k is an integer representing the number of occurrences seen in that same time window. For small values of λ, the distribution behaves almost like an exponential function, but for larger values of λ, the probability distribution approaches that of a normal distribution function. Figure 26 shows the Poisson probability function for different values of λ.

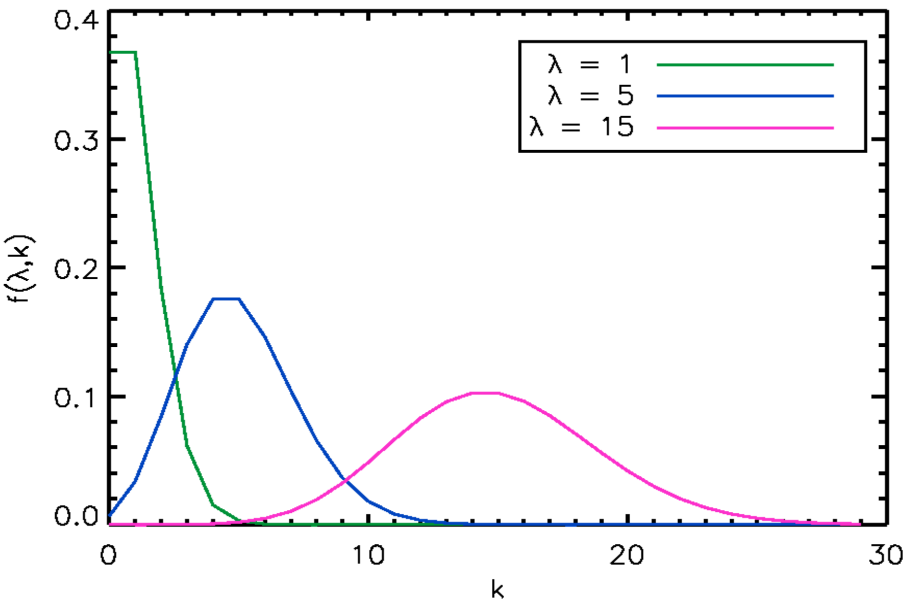


Figure 26 – Poisson Probability Distribution for Various λ Values

For the purposes of this simulation, the output must reflect the raw data as seen by the experiment. Therefore, from the Poisson distribution the analysis must go a step further to characterize the distribution of arrival times for these events. The probability distribution for Poisson-governed rare events is an exponential function with characteristic time 1/ λ, as shown in Eq. 27 (Kay 2005).

$$\boldsymbol{f(\lambda) = e^{-\lambda t}} \qquad \text{Eq. 27}$$

A random distribution of arrival times for any dark carrier contributor follows an exponential function with a specific characteristic rate (λ). Since the process is memoryless, a program can generate every arrival time from zero, as if it were the first. These statistical characteristics form the basis for the simulation of all types of carrier arrival times.

Note that two constant rate (time-independent) Poisson processes contributing simultaneously to a dark count rate, as is sometimes the case, do not result in a Poisson process with characteristic arrival time equal to the average or sum of the two original characteristic times. Dark count rate measurements record the first arrival time. Due to the shape of the exponential probability distribution, the observed rate will be higher than either of the original contributing rates. This occurs because there is an overall increase in the probability of early events, so the combined probability that one of the contributing Poisson processes will introduce a dark count early increases. Figure 27 shows an example of such an interaction. Using the exponential distribution, two constant rate processes were simulated using a Monte-Carlo setup, generating 1 million arrival times according to each process's characteristic rate. Each set of arrival time generations is treated as one trial. The smaller (first) arrival time for each trial is the observed arrival time, and the distribution of those times was plotted as the distribution for the resulting process.

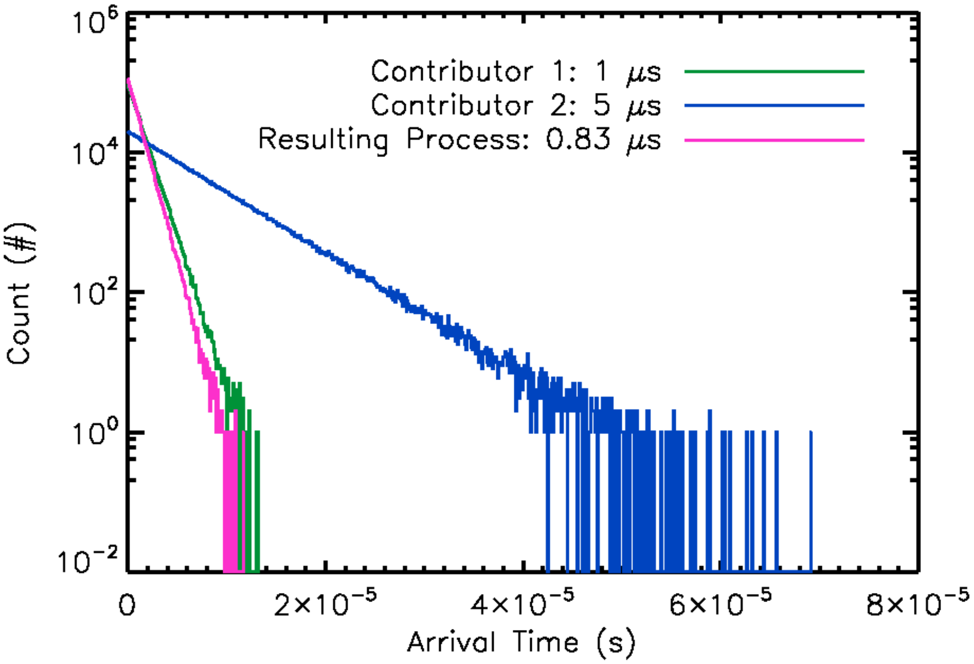


Figure 27 – Combined Poisson Processes in a "First Arrival" Scenario

The plot in Figure 27 shows the histograms for each set of arrival times, which are exponentially distributed (which appears linear on a log-linear plot). The two contributing Poisson processes have characteristic (mean) times of 1 μs and 5 μs, but the resulting combined process (as would be measured in a dark count scenario) has a mean arrival time of 0.83 μs.

4.2 APPROACH

To start, the structure of the simulation is such that the user may input any combination of parameter values. For instance, any number of thermal dark carrier generation rates, number of traps and their characteristic detrapping lifetimes, and the relative level of photon generation during an avalanche could be combined in every possible combination to generate a DCR vs Quench Time plot based on the results. This plot is then compared to the experimental results and evaluated by a chi-squared fit, the result of which is stored in an n-dimensional cube (where n is the number of input parameters) that stores the fit at coordinates corresponding to the input values that rendered it. Figure 28 shows the concept of this data storage design. By cataloging the fit data in this way, it is easy for the user to ascertain the combination of inputs that gave the best approximation of the experimental data. The simulation is Monte-Carlo style in that it generates many scenarios and calculates parameters based on the output, instead of the alternative of raw calculation based purely on equations.

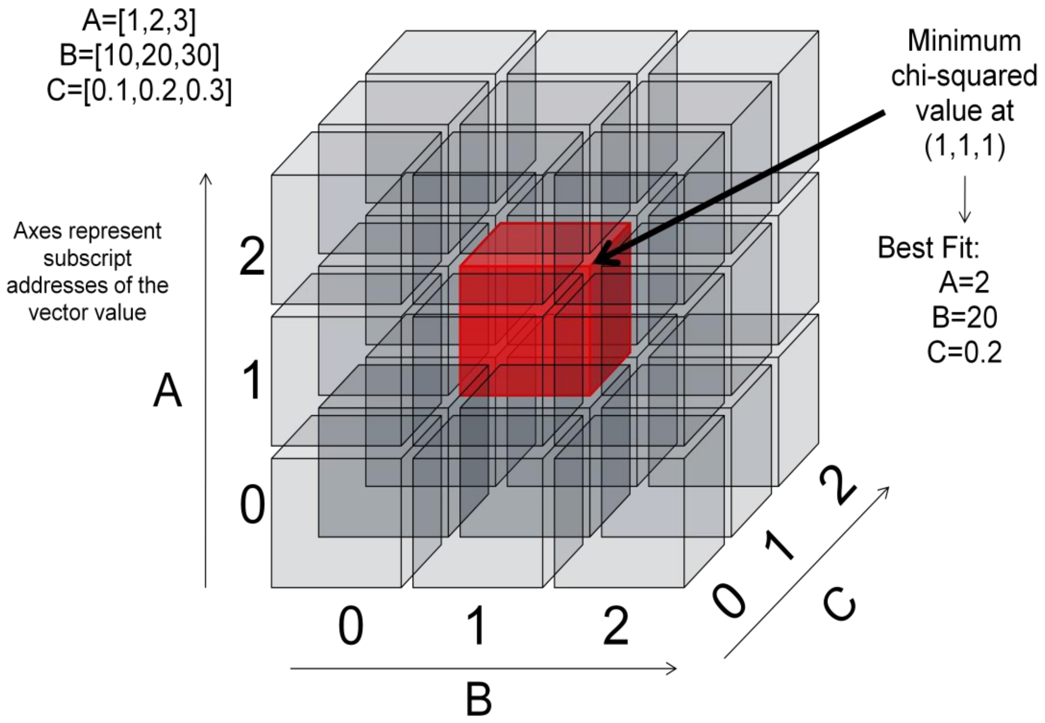

A=[1,2,3]
B=[10,20,30]
C=[0.1,0.2,0.3]
Axes represent subscript addresses of the vector value
A
2
1
0
0
1
2
B
0
1
2
C
Minimum chi-squared value at (1,1,1)
Best Fit:
A=2
B=20
C=0.2

Figure 28 – Representation of Chi-Squared Value Storage technique

For each type of carrier simulated, the same general rules apply based on the operability of the AQB (so that the simulated data is comparable to the experimental data).

1) The program assumes the electric field sweeps any carrier generated during the quench time to the collection node, and therefore it has no contribution. In the code language, these arrival times become a very large number, so the program does not count them as the shortest (and therefore first) arrival time.
2) Any carrier generated during the arm pulse will generate an avalanche with some probability defined as the avalanche initiation probability. These arrival times become equal to the length of the arm pulse, which implies that they cause an avalanche immediately.
3) The program assumes that any carrier generated after the arm pulse will initiate an avalanche at the generation time, with some probability defined as the avalanche initiation probability.

The first carrier contribution simulated is from the thermal dark carrier generation. Since this rate is independent of any other input parameters and will be the same over all values of quench time, it is a straightforward calculation. A random number generator is used with an exponential distribution based on the input thermal dark count rate. The zero time (point in time from which the arrival time is measured) for thermal dark counts is the rising edge of the quench signal (see Figure 23).

The second carrier contribution simulated is from afterpulsing. The simulation generates a set of arrival times for each individual trap defined by the user by an exponential function based on the detrapping lifetime (also defined by the user). Any traps specified represent the average number of filled traps per period, not the total number of traps in the device. The zero time for these distributions is at $-t_q$ because the traps are filled during the previous avalanche. This frame of reference makes the resulting afterpulsing contribution dependent on t_q, which means that the overall mean DCR will increase at shorter quench times if the device contains traps. This is in contrast to the constant rate observed from the thermal contribution with respect to the dead time.

The third carrier contribution simulated is from the self-retriggering carriers. The actual user input represents the number of these carriers that are present in the absorption region upon re-arm. This number is proportional to the original number of photons produced during an avalanche, based on the probability that a photon is generated from a carrier participating in an avalanche and the wavelength of that photon, which determines where the carrier is generated and therefore if it can initiate an avalanche. Since the simulation assumes that the self-emitted photons generate carriers immediately, the distribution of interest derives from the time that the carriers exist inside the device, called the carrier lifetime. The user can input any number of possible carrier lifetime values for the simulation. The result of the unique nature of these carriers is an on/off type of contribution. Since the carriers already exist, they start an immediate avalanche or they do not based on where they are in the device and the avalanche initiation probability to which every carrier is subject. It is very unlikely that these carriers would remain free after a full active quenching cycle. The contribution of these carriers, like afterpulsing, is dependent on t_q due to the carrier lifetime's zero time, which is $-t_q$, when the photons generate carriers.

This simulation ignores tunneling current because it is not likely a dominant, or resolvable, contributor to the DCR. Tunneling carrier generation remains static with t_q behaving as the thermal carrier generation does (Pierret 1996). It is possible that the best-fit thermal dark carrier contribution for the HFF device would actually be a combination of thermal carrier generation and tunneling carriers, but since both are constant rate processes with respect to quench time they are irresolvable (see Figure 27).

Once the simulation has generated all of the potential arrival times, it compares them as individual sets to determine the first arrival time for a single trial. If all the individual series of arrival times were rows of an array (thermal contribution as one row, one trap scenario as the next, etc.), then each "trial" is represented by a single column. The shortest time in a column is the observed arrival time for that trial. The result is a single-row vector of a length equal to the number of trials, populated with as-measured first arrival times that are comparable to the experimental data.

5 EXPERIMENTAL DATA

MIT Lincoln Labs, the developer of the GM APD architecture found in the diodes characterized here, provided a device wafer for testing. Using a microscope-equipped wafer probe station in the RIT Center for Detectors cleanroom lab space, the wafer was tested under the following conditions in order to characterize the dark performance of various devices.

5.1 EXPERIMENTAL SETUP

The test setup secures the wafer to the probe station platform via vacuum and uses small probes to contact the metal pads for specific devices, as seen in Figure 15. A dark covering (two layers thick) was used to block any stray light, which would affect the measurements, and all light sources in the room were turned off (except for measurement and power supply equipment). The background light leakage was measured using a calibrated large-area photodiode. A simple calculation uses the voltage output from the diode along with the manufacturer's gain output for the mean wavelength of light in the room (assumed) to define the number of incident photons per second on the device. The result as measured on the wafer platform was 15,900 $^{\text{photons}}/_{\text{second}}$, which is valid for all experiments.

Table 1 shows the operational settings for the passive quenching measurements using a load resistor. V_s is the supply voltage, R_L is the load resistance, R_i is the input impedance of the oscilloscope, R_{TH} is the Thévenin equivalent resistance (see Figure 19 and Eq. 18), and C_1 is the cable capacitance from the diode to the oscilloscope as shown in Figure 16. R_L and C_1 are measured values, and R_i is the specification for the oscilloscope.

Table 1: Settings for Passive Quenching Measurements

Parameter	**Value**
V_s	36.00 V
R_L	198.7 kΩ
R_i	1 MΩ
R_{TH}	165.8 kΩ
C_1	355.5 pF

Table 2 shows the operational settings for the active quenching measurements. V_{th} is the threshold voltage to which the circuit compares the diode voltage, V_q is the quench voltage, V_a is the arm voltage (the voltage applied to the diode at re-arm), and t_a is the time for which the diode is held at the arm voltage before the voltage is allowed to decay.

Table 2: Settings for Active Quenching Measurements

Parameter	**Value**
V_{th}	33.9 V
V_q	25.7 V
V_a	36.0 V
t_a	1 μs

Based on preliminary testing results, a sample size of 30,000 is sufficient to provide statistically relevant data for even high DCR values, where the number and population of bins other than the first in the corresponding histogram would be minimal at smaller sample sizes. In order to compute the first bin residual, all the individual trials combine to form a single histogram for a more accurate calculation.

5.2 PASSIVE QUENCHING RESULTS

The passive quenching experiment aids in the understanding of the circuit as well as the functionality of the diode. An analysis of the diode capacitance is also discussed, including theoretical calculations and measurement-based comparisons.

5.2.1 *RC RECHARGE CIRCUIT ANALYSIS*

The oscilloscope monitored the diode voltage according to the setup shown in Figure 16. Figure 29 shows a selected section of the waveform that exhibits the RC recharge curve as described in Figure 21.

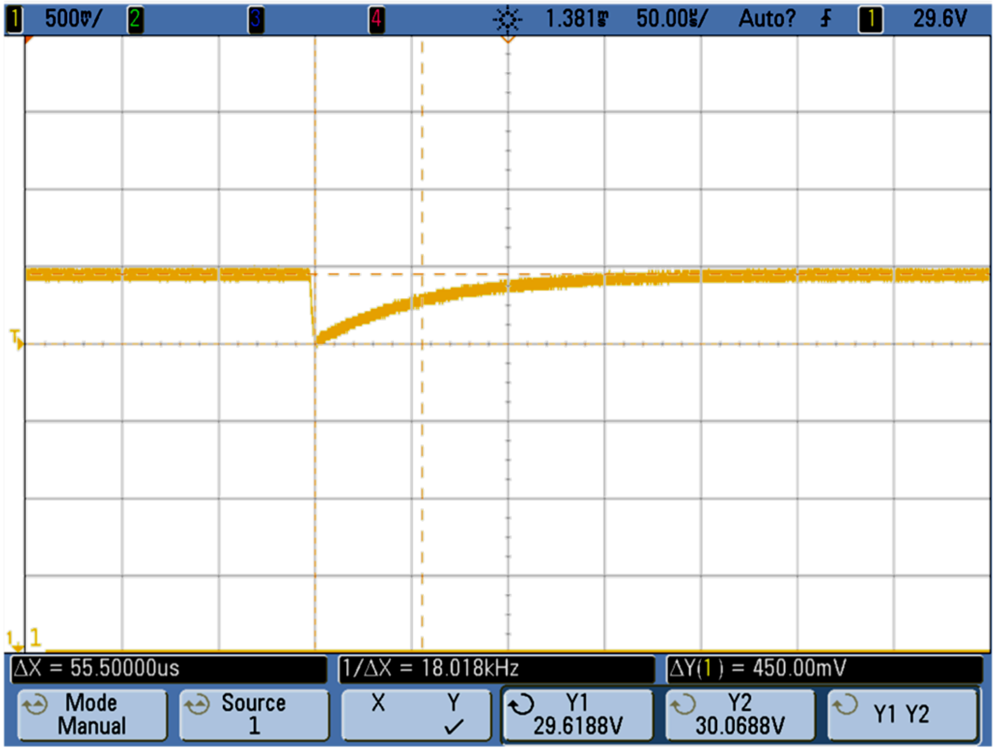

1 500mV/
2
3
4
1.381ms
50.00us/
Auto?
1 29.6V
ΔX = 55.50000us
1/ΔX = 18.018kHz
ΔY(1) = 450.00mV
Mode Manual
Source 1
X Y ✓
Y1 29.6188V
Y2 30.0688V
Y1 Y2

Figure 29 – Passive Quench Screen Capture with Cursor Measurements

The upper limit of the voltage is v_d (see Figure 19), and the expected value can be calculated with Eq. 15 and the values in Table 1. According to the passive quenching circuit model, V_{BR} is the lowest voltage level in the curve. The RC constant is equal to the time delay from the moment the curve reaches V_{BR} to when the curve crosses 63.2 % (or e^{-1}) of the change in voltage between the minimum and maximum values. Table 3 lists the expected and actual values for these parameters.

Table 3: Expected and Actual Values for Passive Quench Parameters

Parameter	Expected Value	Measured Value
V_{BR}	N/A	29.62 V
v_d	30.03 V	30.07 V
RC	58.9 μs	55.5 μs

5.2.2 *DIODE CAPACITANCE*

A high-level approximation for the diode capacitance is given by the bounds of Eq. 22 as calculated with the smaller and larger radius values (see Figure 51). Table 4 lists the geometric constants of the device. Using these values, the lower bound for the capacitance is 0.215 pF while the upper bound is 0.366 pF.

Table 4: Device Diode Geometry

Parameter	Value
R_a	30 μm
R_b	23 μm
d	0.8 μm

A better approximation of the capacitance comes from Eq. 42. Solving numerically using the values in Table 4 yields a diode capacitance value of 0.259 pF.

The passive quenching experimental results cannot characterize the diode capacitance because the cable capacitance is several orders of magnitude larger than the expected diode capacitance. In the absence of inductance-capacitance-resistance (LCR) meters with the required sensitivity, the best way to obtain an experimental value is to measure the capacitance of a larger-area diode and scale the results. Previous measurements of a large diode with top disk area 1210 times larger than the device diode (if the device diode area is calculated using the smaller disk radius) and equal depletion width (separation of the plates) gave a capacitance measurement of 259 pF (see highlighted line in Appendix C). Given that capacitance scales linearly with area (see Eq. 22), dividing the large-area capacitance by the area ratio gives an approximation of the capacitance expected for the device diode. This calculation yields a value of 0.214 pF if the parallel plate approximation holds and the smaller disk radius is used. The calculation does not take into account edge effects (for large areas, the edge effects are negligible, but for smaller areas they become more important). If the larger disk radius is used, the calculation yields 0.364 pF (the area would be only 711.1 times larger). The theoretical approximation calculated above falls between these two measurement-based values: $0.214 \text{ pF} < C_d < 0.364 \text{ pF}$. This range includes the calculated value from Eq. 42.

5.3 ACTIVE QUENCHING RESULTS

Using the AQB and an oscilloscope to capture the APD voltage over time, an IDL program interfaces with the oscilloscope and collects a set of first arrival times for a single value of t_q. The measurement sequence varies the quench time across the range of the AQB's setting capabilities, and then repeats the entire set to both improve measurement accuracy and characterize the associated error. IDL calculates the trend of mean DCR vs quench time based on this data. A discussion of the collection and analysis of the raw data follows.

5.3.1 *INITIAL MEASUREMENTS AND IMPROVEMENTS*

Initial measurement sets revealed that the circuit suffered from leakage problems, as shown in Figure 30.

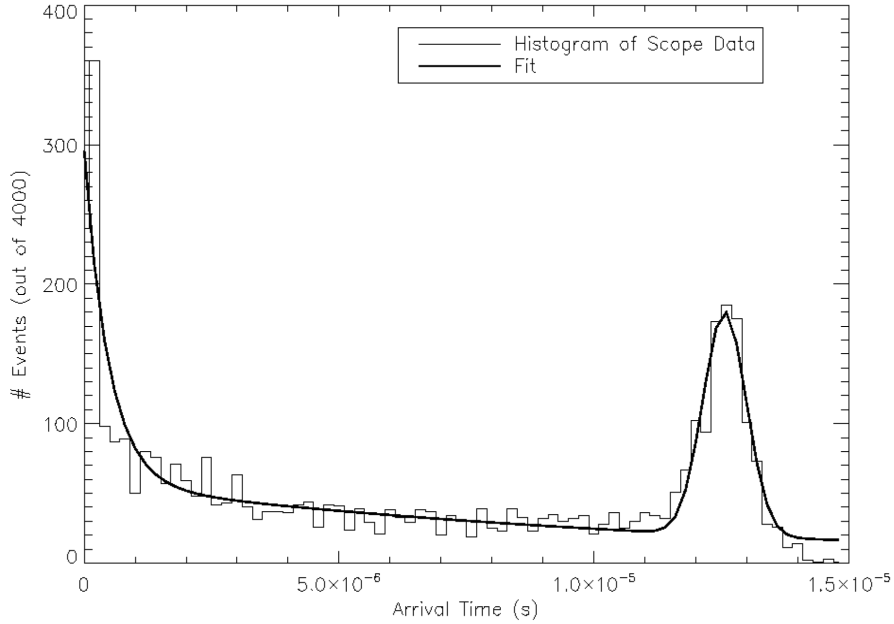


Figure 30 – Example of Arrival Time Distribution with various artifacts present

Note that the zero time axis in Figure 30 does not represent measured zero; the program is designed to shift the axis to show the smallest recorded time as zero. This shift simplifies the fit calculation and gives a more accurate description of the dark count rate. The events in the first bin occurred immediately (as soon as detectable), and so they should be counted as occurring at zero. If the program used the actual time stamp to calculate the rate, it would be much lower.

As previously discussed, the first bin is inflated when compared to the rest of the distribution if the exponential trend continued back in time (if the arm pulse were not artificially delaying the minimum arrival time). The Gaussian-like distribution centered at 12.5 µs is an artifact of circuit leakage, caused by the AQB circuit and measurement components leaking and drawing the voltage at the diode down over time. Figure 31 illustrates this mechanism as it manifests on the actual oscilloscope readings.

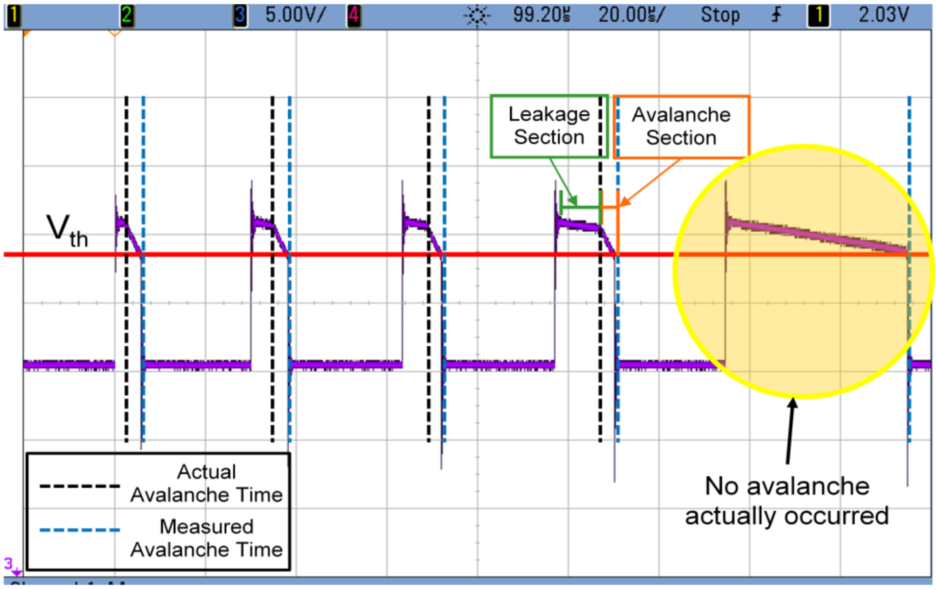


Figure 31 – Example of Artificial Avalanche Recognition Due to Circuit Leakage

The select sequence of avalanche events displays a variety of avalanche detection scenarios. Note that the progression of arrival time lengths shown is coincidental and does not represent a trend in the measurements in general.

As shown in Figure 31, there are two identifiable modes of voltage decay following the release of the arm pulse. The first is the voltage decay due to circuit leakage, which has a shallow slope. The second is the decay due to an avalanche in the APD, which has a markedly steeper slope. The first pulse in Figure 31 represents a case in which there is an immediate avalanche since no shallower voltage decay is evident. The next three pulses represent cases in which there was some time between the release of the arm pulse and the avalanche because the traces show two distinct sections of decay. The last pulse (highlighted) represents a case in which there was no avalanche but the leakage decay decreased the diode voltage to V_{th}. The circuit triggered the quench based on the voltage level, which means that there was a falling edge in the quench signal (see Figure 23), and so the measurement system recorded an avalanche time. Because the voltage decay due to the circuit leakage is a constant rate with some variation, the result in the arrival time histogram is as seen in Figure 30: a maximum time that the circuit is able to wait for an avalanche to occur, plus or minus some error that forms a normal distribution.

After adjusting some circuit elements and removing the monitoring probe for the APD during active measurements, the leakage artifact was no longer present in results and the experimental data collection could move forward.

5.3.2 *DATA REDUCTION METHOD*

Given a histogram with only one artifact (first bin), extraction of the mean DCR is desired. There are a few methods that may achieve this; depending on the characteristics of the data set, one may be more appropriate than another. In order to choose the best approach, a review of the exponential distribution's characteristics is required (the histogram should resemble this distribution for time-independent process contributions). The expected value of an exponential distribution is $1/\lambda$, as previously stated, but the standard deviation of the distribution is also equal to $1/\lambda$. For a constant rate process, fitting an exponential function to the data will yield the characteristic rate λ.

Using a fit function becomes more complicated, however, when there is a time-dependent rate as in the case of afterpulsing (leading to curvature even on a log-linear scale). Figure 32 shows the observed dark counts for the same device for various dead time values.

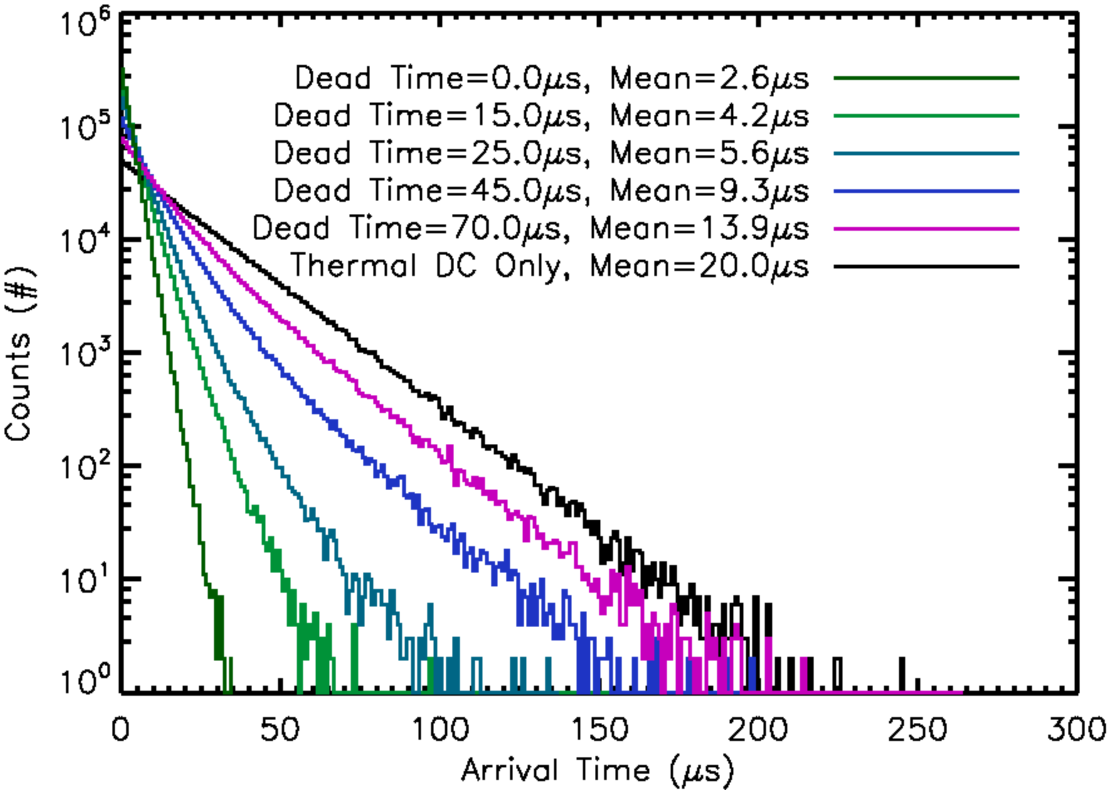


Figure 32 – Observed Dark Counts for Various Dead Time Values for the Same Device

Each curve represents a histogram of the dark counts observed for a device having 10 filled traps with detrapping lifetimes of 30 μs and a thermal carrier generation rate of 50 kHz. The only difference between the curves is the dead time simulated. The curvature in the histograms is due to the difference in zero reference for the thermal carriers and the afterpulse carriers. Visualize the change in zero reference as sliding the exponential probability distribution back in time to the negative dead time value. The thermal carrier probability distribution remains at zero. Any portion of the afterpulse distribution that falls before the zero mark has no affect on the measured dark count rate; only the portion of the distribution in positive time space will contribute. Because the remaining portion of the afterpulse distribution is not a pure exponential, the observed combined distribution will not be either.

Even though the sum of two different exponential functions does not literally describe a time-dependent process, such a model can closely describe the distribution. When the fit has characterized the distribution in terms of individual functions summed together, the observed characteristic rate can be obtained by calculating the weighted average of the arrival times based on the histogram's fit function. In some cases, it may be practical to try to capture the entire range of arrival time values by extending the observation window, thereby allowing for a very simple calculation of the mean of all the times collected. For low DCR sample sets, however, this is impractical because it greatly inflates the time needed to collect 30,000 samples (while retaining the same temporal resolution). An extrapolated fit function's weighted average correctly represents the mean as if the distribution were captured in its entirety.

5.3.3 *DCR VS QUENCH TIME*

As described above, the mean DCR is calculated using the weighted average of the histogram's fit function. Multiple sets of data are collected at each t_q value, and the combined mean DCR of those data sets represent one point on the DCR vs Quench Time plot. For the LFF devices, one expects a relatively flat relationship between the mean DCR and t_q, mainly because the device volume is small and the probability that traps exist inside the device is very low. The thermal dark carrier generation rate should be low because the device has less cross-sectional area and an effective scupper region around the active regions. Conversely, the HFF device should show a strong relationship between mean DCR and t_q as well as a higher thermal dark carrier contribution because of the increase in area and the decrease in the size and effectiveness of the scupper region. One would expect that the MFF device should fall between the two devices in both t_q dependence and base thermal dark carrier contribution.

The background light leakage level measured as 15,900 $^{\text{photons}}/_{\text{second}}$ is less than the minimum DCR measured in any experiment, and so it is assumed that the lower limit of the dark count rate is internally driven rather than noise-related.

Based on the range of settings available on the AQB, the quench time values range from 6 µs to 100 µs and all other values (as shown in Table 2) are constant throughout all active quenching experiments.

5.3.3.1 LFF

Figure 33 shows the DCR results at various quench times for the LFF device. The DCR is plotted on a log-linear scale and fit. This experiment utilizes the full range of the AQB quench time settings. Each data points represent three sets of 30,000 samples.

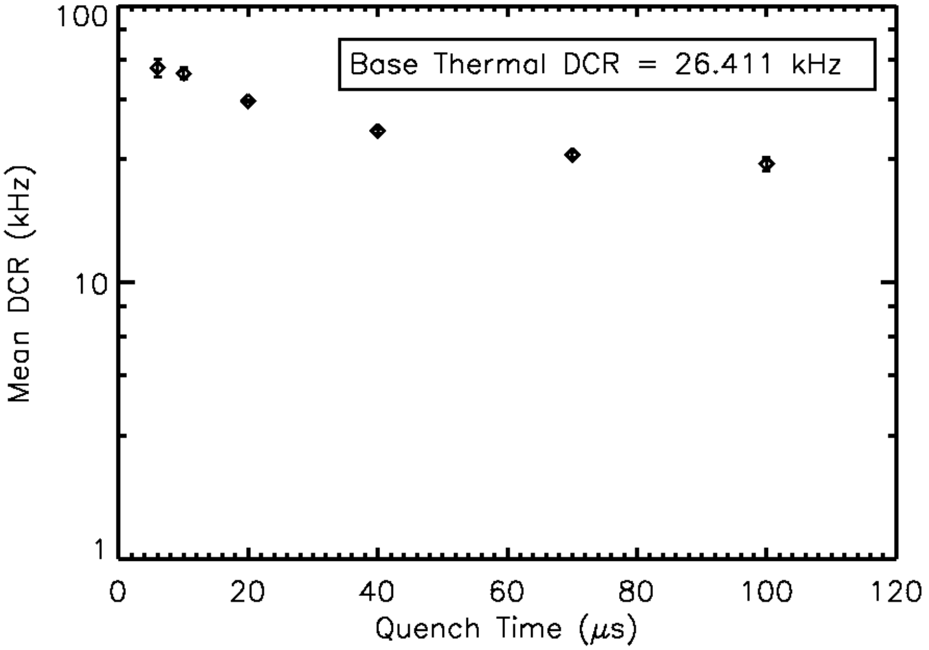


Figure 33 – Experimental Results: DCR vs Quench Time for LFF Device

The error bars shown are +/- one standard deviation of the data. The "Base Thermal DCR" is calculated based on a fit function of the form shown in Eq. 28.

$$\boldsymbol{DCR(t_q) = Ae^{-\lambda_1 t} + Be^{-\lambda_2 t} + C} \qquad \text{Eq. 28}$$

A and B represent the amplitudes of their respective exponential functions, while λ_1 and λ_2 are the corresponding rates. C represents the bias on the function, or the limit that the function approaches at infinity. Since the DCR measured at an infinite dead time represents only the thermal carrier generation rate, C is equal to the thermal DCR associated with the device.

The slope of the curve seen in Figure 33 is shallow, as anticipated. Since little to no afterpulsing or self-retriggering is expected due to the small volume of the diode and the surrounding architecture, this result agrees with expectations.

5.3.3.2 MFF

Figure 34 shows the experimental results for DCR at various quench times for the MFF device. The quench time ranges from 10 µs to 100 µs due to loss of resolution for calculations at shorter quench times. Each data points represent four sets of 30,000 samples.

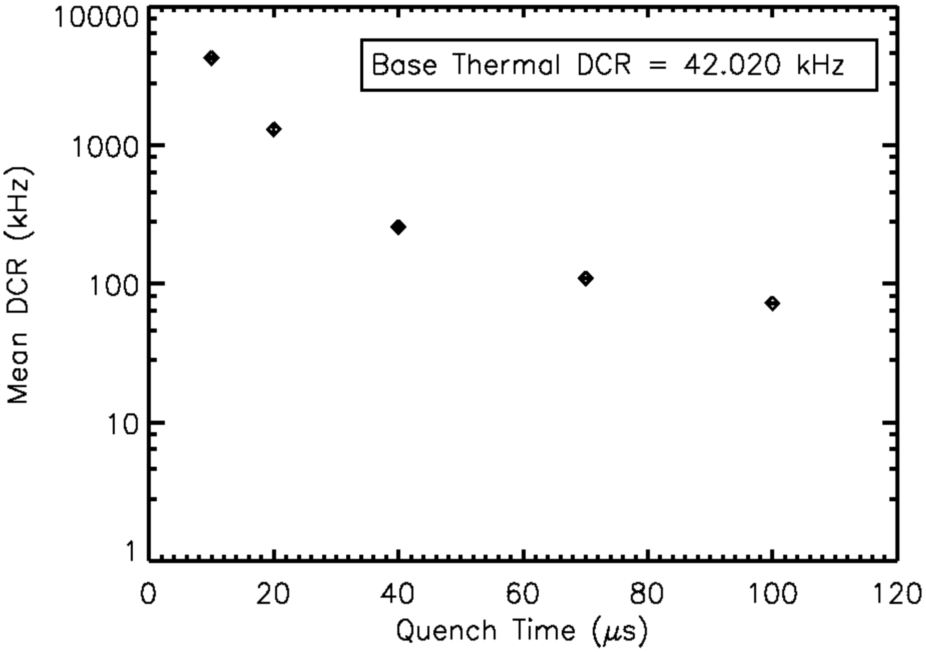


Figure 34 – Experimental Results: DCR vs Quench Time for MFF Device

Error bars are present but not distinguishable at the plot resolution. The thermal DCR associated with the MFF device is calculated according to Eq. 28, and is slightly higher than the same parameter for the LFF device. This makes sense due to the increase in cross-sectional area of the device as well as an increase in edge effects (i.e. tunneling) due to the decrease in influence of the scupper region.

The slope of the trend is also more pronounced than in the LFF device, which matches expectations. Due to an increase in volume, an increase in trap density (increase in afterpulsing contribution) is anticipated. Self-retriggering should also have more influence as the scupper region is smaller and less effective. These three trends should increase the thermal dark count rate and the slope at shorter quench time values, which is exactly the case.

5.3.3.3 HFF

Figure 35 shows the experimental results for DCR at various quench times for the HFF device. The quench time ranges from 25 µs to 100 µs due to loss of resolution for calculations at shorter quench times. Each data points represent four sets of 30,000 samples.

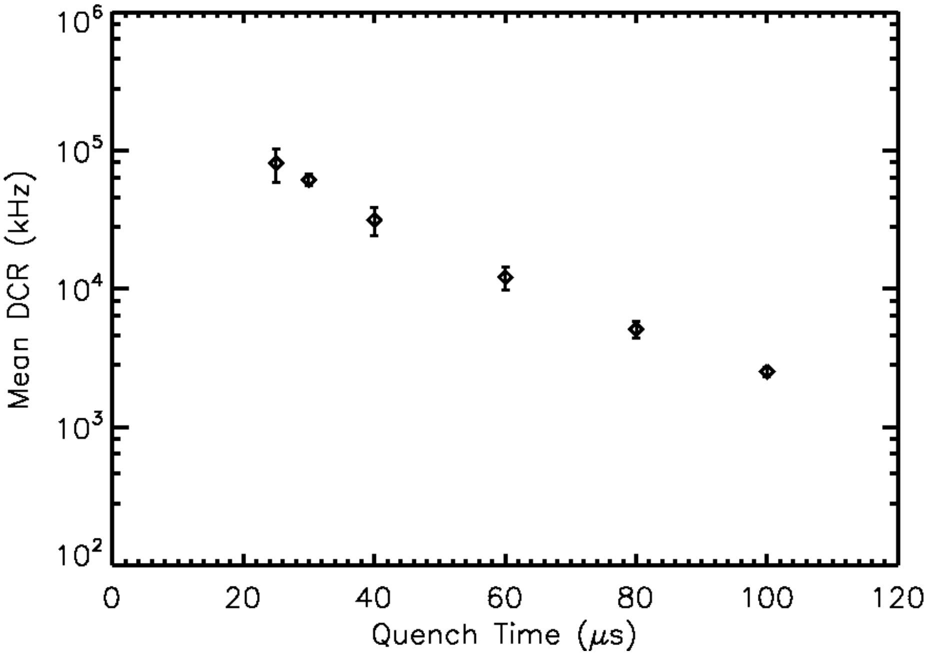


Figure 35 – Experimental Results: DCR vs Quench Time for HFF Device

There is no reported thermal DCR for the HFF device because the data does not show limit-reaching behavior in the available quench time window, and so a fit would not accurately describe the plateau in the trend.

Overall, the DCR values at each quench time are higher than the MFF values, and so it stands to reason that the thermal DCR would also be higher (the expected result). There is an absence of increasing slope at shorter quench times, but the data becomes indiscernible at quench times shorter than 25 µs. It is possible that the upturn in trend does not have influence at quench times so long. Both afterpulsing and self-retriggering are expected to have a more prominent contribution to the HFF DCR.

5.3.3.4 Simulation Matching

Based on chosen values for the avalanche initiation probability and set values for t_a and various t_q, the simulation program attempted to match the experimental data. The simulation varied the thermal dark carrier rate, but centered the input values on the value derived from Eq. 28 when available. IDL's built-in function XSQ_TEST calculated chi-squared values using the simulated values and measured (expected) values as matched pairs. The data (simulated and measured) was normalized so that the maximum experimental DCR point was equal to 10, making the chi-squared values for each device comparable, even though the nominal values of the matching data increases at higher fill factors (and therefore the chi-squared magnitude also increases). The results for matching the LFF data are shown in Figure 36 and the input values for the best-fit case are shown in Table 5.

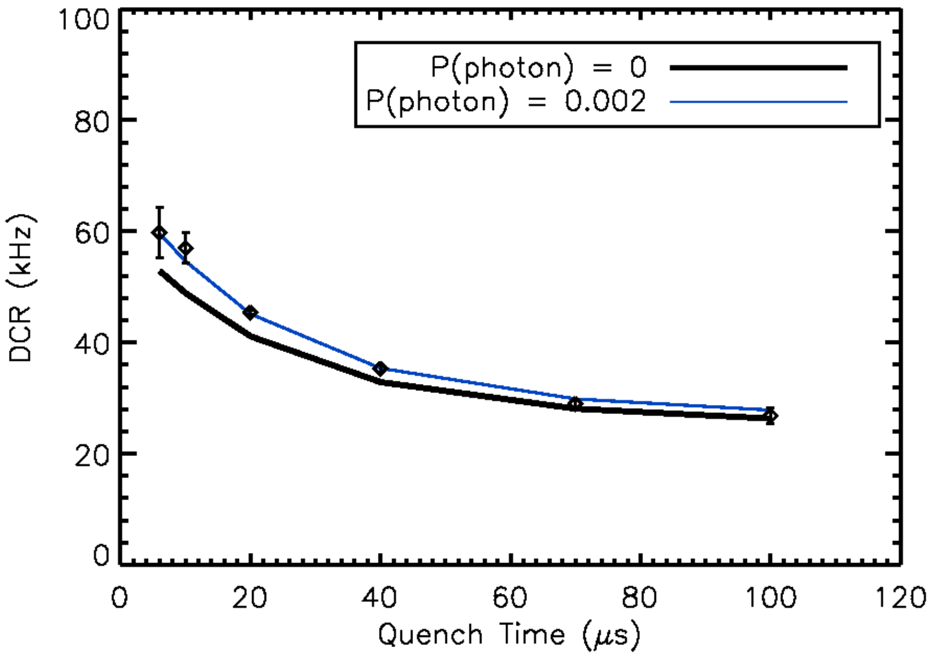


Figure 36 – Simulated Results: DCR vs Quench Time for LFF Device

Table 5: LFF Simulation Match Input Values

Parameter	**Value**
Avalanche Initiation Probability	0.8
Thermal DCR	25.75 kHz
Number of Filled Traps	2
Detrapping Lifetime	35 µs
Carrier Lifetime	10 µs
P_{phot}	0.002
χ^2 (% of minimum fit value)	0.024

The best-fit simulated thermal DCR value is very close to the extracted value of 26.4 kHz from the measured data. As expected, the best fit occurs when there are very few filled traps, and the carrier lifetime value of 10 µs is a reasonable value for a silicon device. P_{phot} represents the probability that a self-retriggering carrier will cause an immediate avalanche in any given arm period. It is a reduced term derived from a combination of individual probabilities. The first is the probability that a photon will be absorbed in an area facilitating its collection in the next arm cycle (small for the LFF device). The second is the probability that the carrier is collected, and the last is a scaled avalanche initiation probability based on the number of carriers likely available – the probability of an immediate avalanche is higher for two carriers present during the arm pulse than for one. The simulation defines these probabilities as separate inputs, but since the combined probability is the defining characteristic of the curve, the simulation plots show the reduced term. The chi-squared term is very small but not unique, as shown in Figure 38. Note that the chi-squared histograms show the values of the non-normalized results for greater resolution.

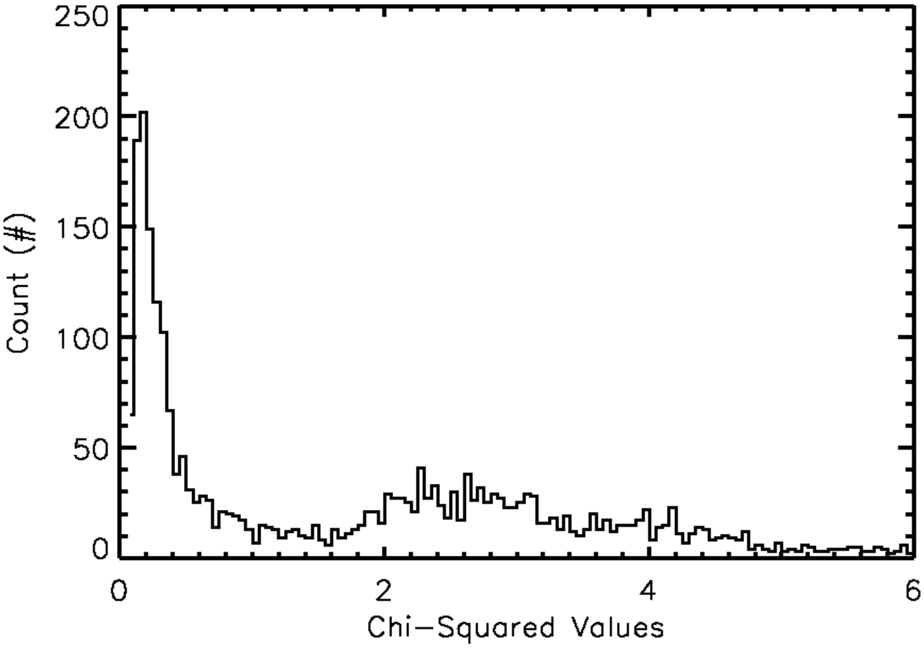


Figure 37 – Histogram of Chi-Squared Results for LFF Simulation Matching

The large congregations of similar values centered on different means indicates that one or more input variables is degenerate (does not significantly affect the outcome) or that the same fit can be generated with different input value combinations. In this particular case, the degeneracy seems to center around the contributors to P_{phot}, which is logical given that the self-retriggering photon contribution is very small and does not affect the DCR vs quench time trend significantly.

The results for matching the MFF data are shown in Figure 38 and Table 6. The thermal DCR extracted from the experimental data served as a starting point for the corresponding input to the simulation, while the avalanche initiation matched that for the LFF device since both devices are on the same substrate. t_a was set according to the measurement conditions at 1 µs and the simulation varied t_q according to the measured data. The simulation varied all other inputs to obtain the best fit.

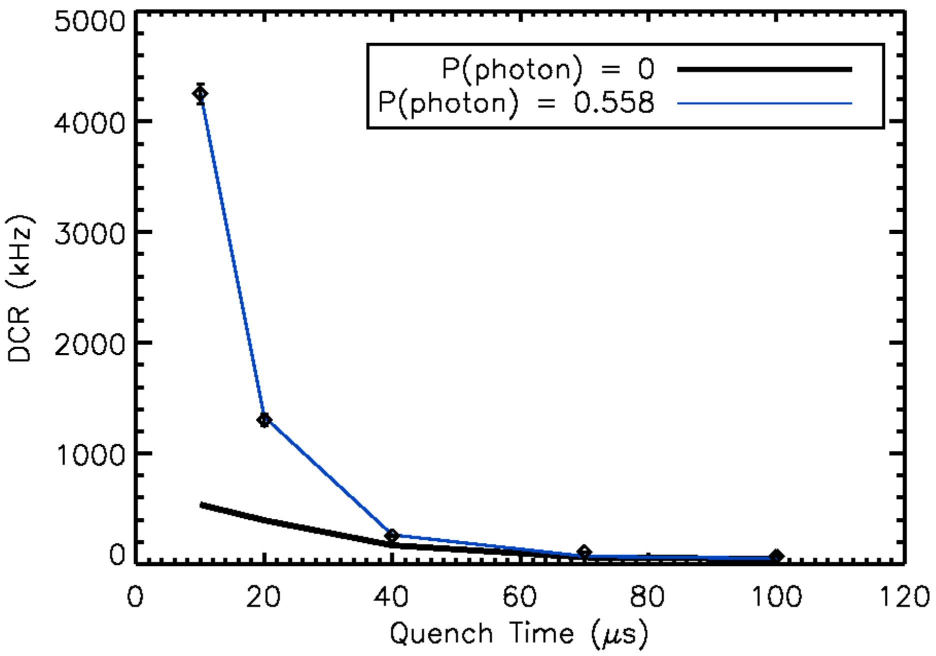


Figure 38 – Simulated Results: DCR vs Quench Time for MFF Device

Table 6: MFF Simulation Match Input Values

Parameter	**Value**
Avalanche Initiation Probability	0.8
Thermal DCR	44 kHz
Number of Filled Traps	46
Detrapping Lifetime	18 μs
Carrier Lifetime	9 μs
P_{phot}	0.558
χ^2 (normalized data)	5.6e-5

Again, the best-fit thermal DCR value closely follows the extracted value from the experimental data (42 kHz). The number of traps necessary for this match is more than for the LFF match, but this makes sense given that the area and volume of the diode are increased. The detrapping lifetime is slightly shorter in this simulation. Different detrapping lifetimes indicate different types of traps, but because the simulation tries to match only one type of trap, it forces the different detrapping lifetimes to combine and alias as one type. Therefore, it is not necessarily that the MFF device has a different trap species, only that the ratio of different species in the bulk or at the surface (present in all devices) is slightly different. P_{phot} is also higher for the MFF device, which is expected due to the increased area and volume as well as the less effective scupper region. The chi-squared value is very low, which indicates a very good fit. This set of chi-squared values also suffers from some degeneracy, however, shown in Figure 39.

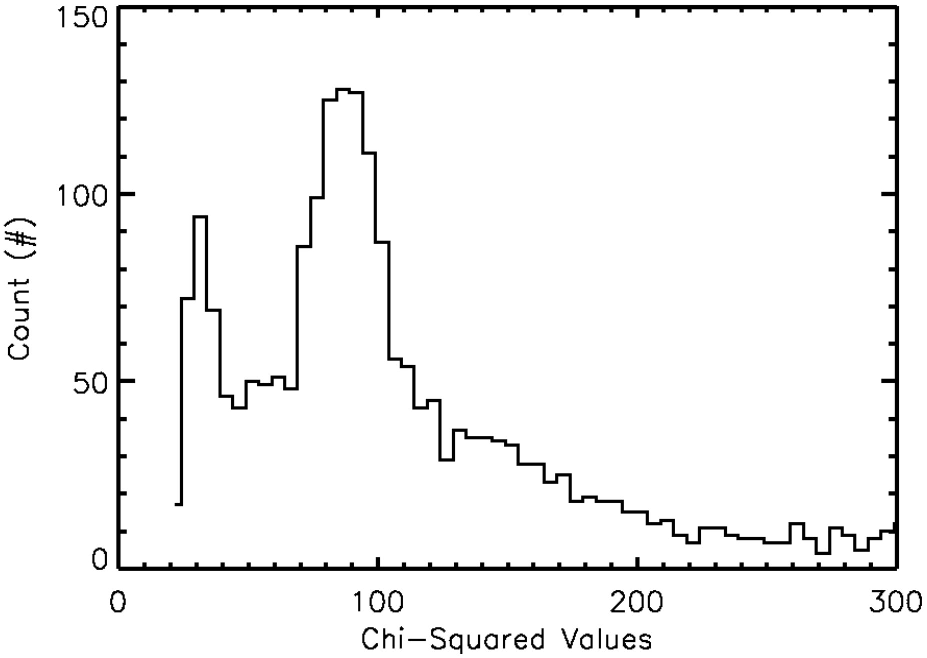


Figure 39 – Histogram of Chi-Squared Results for MFF Simulation Matching

The congregation of points does not exhibit a clear pattern (e.g., the degeneracy associated with the LFF chi-squared results), and so it is more probable that the degeneracy stems from multiple combinations of inputs yielding the same result. Simulating inputs with larger resolution would result in less degenerate chi-squared values. Another solution would be to add a secondary fit constraint beyond that of the DCR value for each quench time, like a first bin residual metric or even histogram comparisons.

The results for matching the HFF data are shown in Figure 40 and Table 7. t_a and t_q were set according to the measurement conditions. The avalanche initiation probability was set at the same value again for consistency. Because the model could not extract the thermal DCR from the experimental data, the input values centered on a best guess.

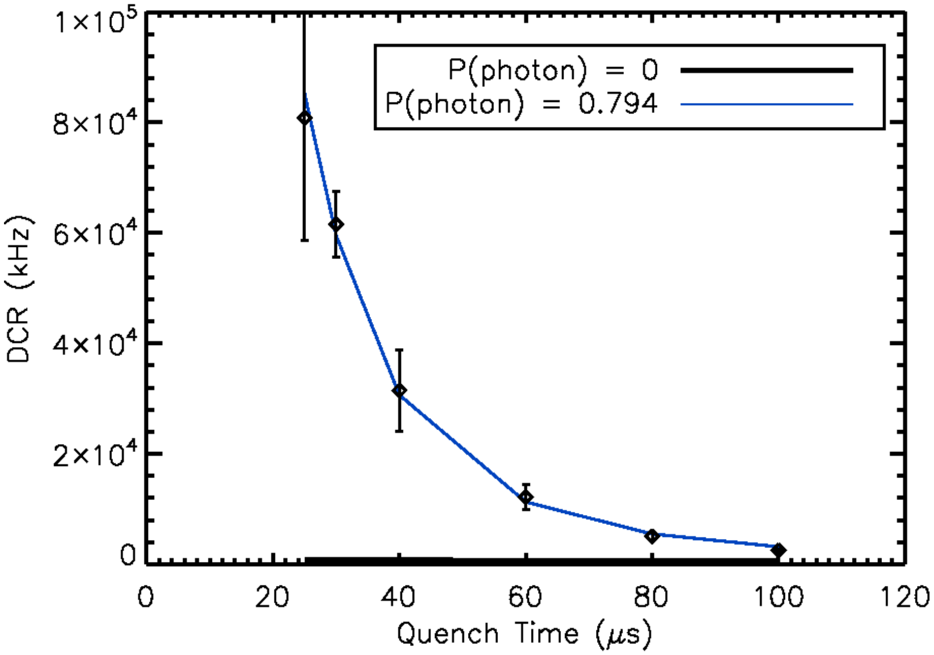


Figure 40 – Simulated Results: DCR vs Quench Time for HFF Device

Table 7: HFF Simulation Match Input Values

Parameter	**Value**
Avalanche Initiation Probability	0.8
Thermal DCR	390 kHz
Number of Filled Traps	250
Detrapping Lifetime	60 μs
Carrier Lifetime	80 μs
P_{phot}	0.794
χ^2 (normalized data)	0.063

Having no benchmark for thermal DCR from the measured data, the simulated thermal DCR is reasonable given the increase in area of the device along with expected increases in contributions from tunneling effects around the edge of the diode. The number of filled traps is also much larger for the HFF device according to the simulation, and the detrapping lifetime is reasonable compared to the lifetimes dictated by the LFF and MFF simulations. The carrier lifetime is significantly longer in this device when compared to the other two, but it is still reasonable for a silicon device. As expected, P_{phot} is higher than the MFF device, due to the increased area of the absorption and multiplication regions as well as the very small scupper region associated with the HFF device architecture. The chi-squared value is small, though it is the largest of the three simulation matches. This simulation set also suffers from some degeneracy due to the fine resolution of the variable input values, shown in Figure 41.

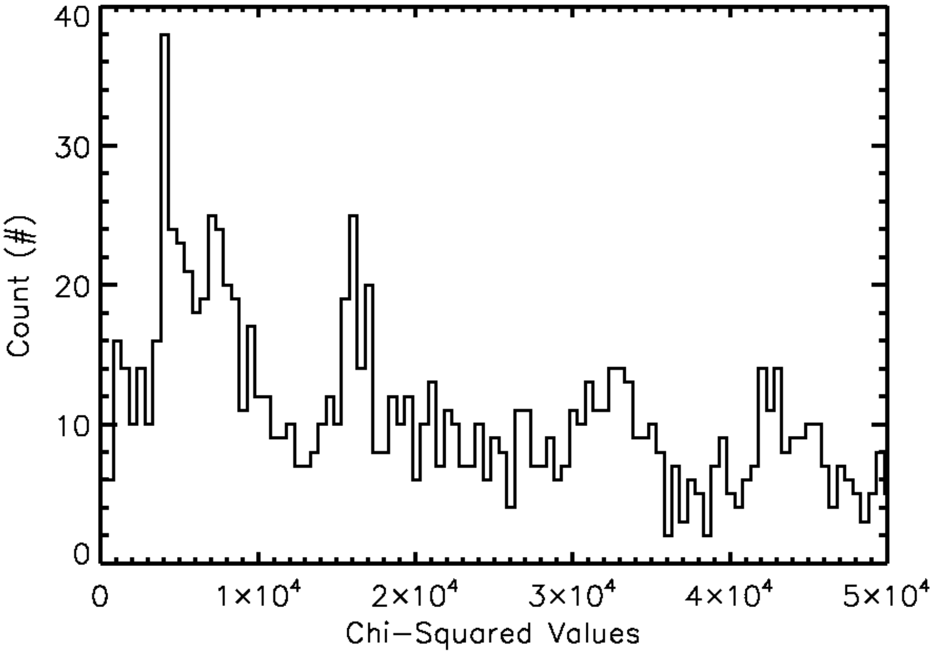


Figure 41 – Histogram of Chi-Squared Results for HFF Simulation Matching

Like the MFF degeneracy, this congregation of small chi-squared values is due to multiple combinations of inputs yielding similar fits.

5.3.4 *FIRST BIN RESIDUAL RESULTS*

IDL calculates the first bin residual by fitting the histogram data (excluding the first bin) and projecting the number of counts that should have accumulated in the first bin due to thermal dark carrier and afterpulsing contributions, based on the extrapolated function. The expected value of counts for the bin is subtracted from the actual number of counts in the bin. A positive value is the excess contribution from self-retriggering carriers (with some noise due to the resolution hampered by the length of the arm pulse). Due to the increase in volume and the decrease in effectiveness and size of the scupper region, the HFF device should have a higher first bin residual over all values of t_q as well as have a greater dependence on t_q. As the quench time decreases, the probability that the photon-generated carrier has not yet recombined increases exponentially. Conversely, the LFF device should have very low or negligible (with respect to the noise) contributions from the self-retriggering mechanism because of the scuppering region and the significantly smaller active area into which the photon-generated carriers must travel. Again, it is logical that the MFF device would be a midpoint between the dependency characteristics of the HFF and LFF devices.

5.3.4.1 LFF

Figure 42 shows the first bin residual results for the LFF device on a linear-linear plot, with the percent of avalanches caused by self-retriggering on the left y-axis and the numerical output on the right.

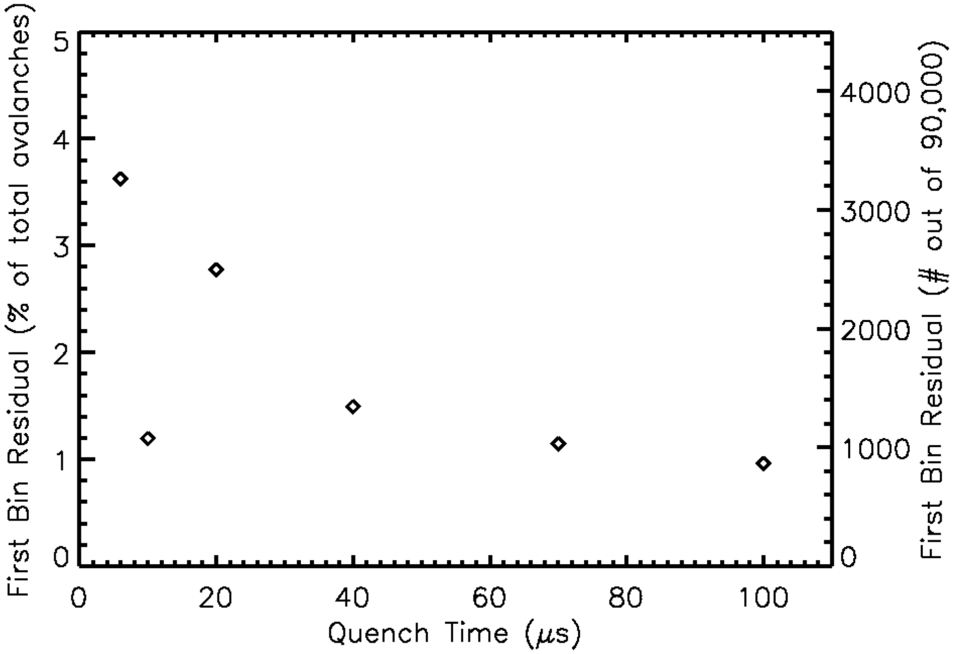


Figure 42 – Experimental Results: First Bin Residuals for LFF Device

No error bars are present in the plot because all available data sets were combined to form one large histogram (which is valid due to the memoryless characteristic of Poisson processes), yielding only one solution for each quench time value. This is to ensure the most accurate fit possible, since a small variation in the fit can translate to a large variation in the first bin residual calculation. Aside from one errant point at 10 µs, the data shows a consistent downward trend at longer quench times. This is expected because the self-retriggering contribution is governed by the carrier lifetime: carriers have a higher probability of being free (as opposed to having recombined) after shorter dead times. The shallow upward trend mimics that seen in the DCR data. However, with the highest measured contribution at roughly 3.6 %, the self-retriggering contribution for this device architecture is very low. This matches expectations based on the simulated fit to the LFF data in Figure 36.

5.3.4.2 MFF

Figure 43 shows the first bin residual results for the MFF device on a log-linear plot, with the percentage value on the left y-axis and the numerical output on the right.

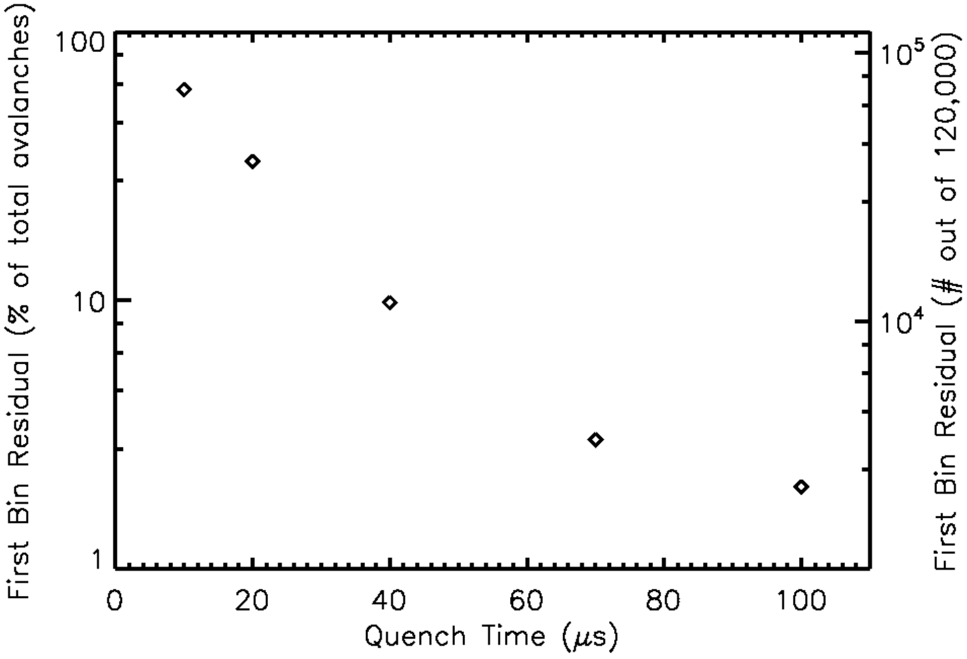


Figure 43 – Experimental Results: First Bin Residuals for MFF Device

The trend in Figure 43 shows a dramatic increase in counts for shorter quench times. This trend is much steeper than the LFF trend shown in Figure 42, and mimics the overall MFF DCR trend. The correlation of increased DCR and increased self-retriggering contribution is reasonable since all the counts represented in the first bin residual plot contribute a “zero” time stamp to the mean arrival time. The higher overall contribution (roughly 61 % as measured) is expected as well due to the less effective scupper region in the MFF internal architecture.

5.3.4.3 HFF

Figure 44 shows the first bin residual results for the HFF device on a linear-linear plot, with the percentage value on the left y-axis and the numerical output on the right.

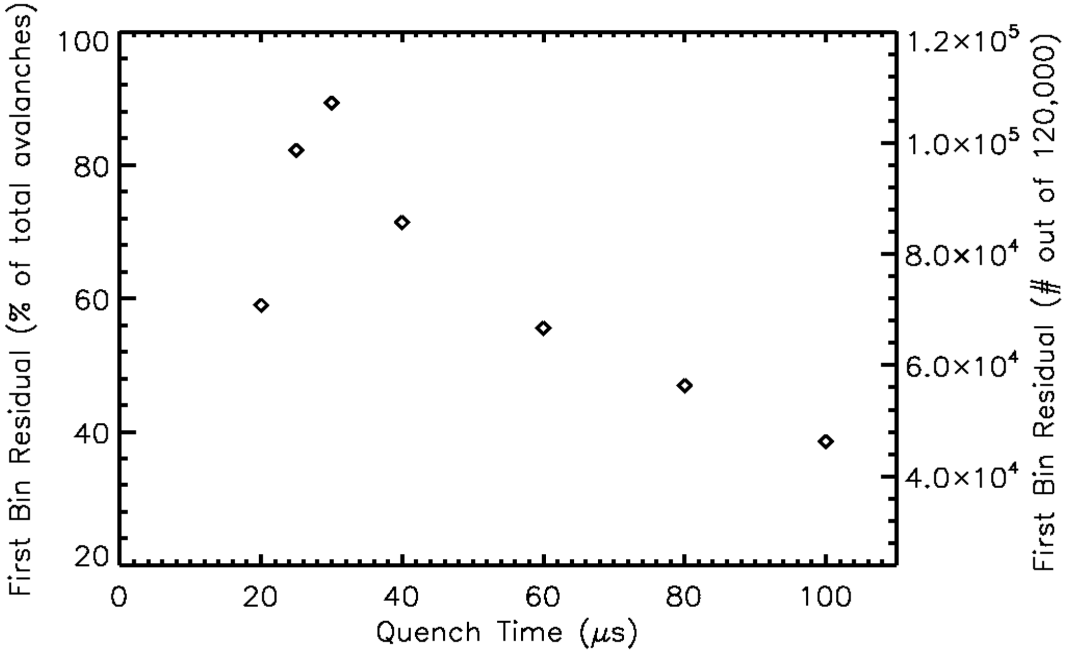


Figure 44 – Experimental Results: First Bin Residuals for HFF Device

As in the DCR measurements, there is a lack of steep increasing slope at shorter quench times (as seen in the MFF device), but there is also an unexpected decrease at the shortest quench time values. This plot represents the self-retriggering contribution (additional avalanches that would not have otherwise been initiated), and not the actual number of carriers generated – the calculation represents an “exclusive or” function. Because of the high number of traps present in the device (according to the simulation), their relatively short detrapping lifetimes, and the relatively long carrier lifetime, the self-retriggering mechanism is no longer dominant as the quench time decreases past 30 µs.

As expected, the overall contribution from self-retriggering is higher for the HFF device due to a very limited scupper region and larger multiplication region volume. The somewhat shallow trend in the first bin residual plot (as compared to the MFF device) mimics the slope seen in the HFF DCR plot (Figure 35).

5.3.4.4 Simulation Matching

Using the best-fit case of inputs from the DCR vs quench time simulation matching, a customized IDL routine extracted data from the simulated arrival time histograms, in the same way as the experimental data, to calculate the first bin residual (self-retriggering carrier contribution). For all device results, each data point represents a calculation based on the same number of samples (90,000 for the LFF and 120,000 for the MFF and HFF) as the measured data. Note that the simulated data matching was not optimized with respect to the first bin residual calculation. Figure 45 shows the simulated and experimental results for the LFF device on a log-linear plot, with the percent of avalanches caused by self-retriggering on the left y-axis and the numerical output on the right.

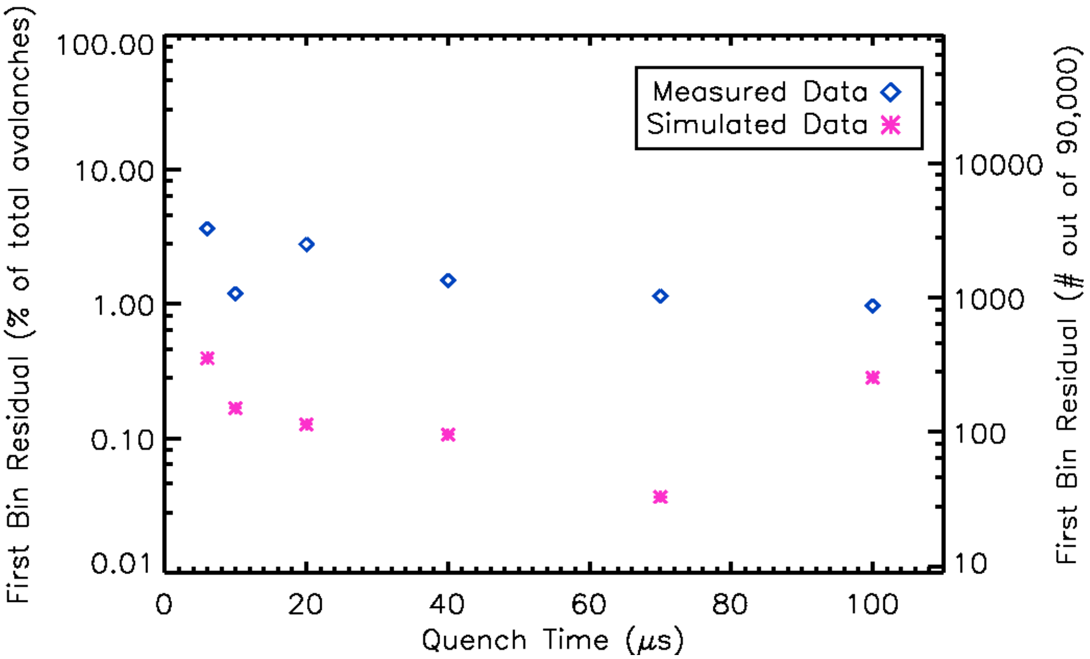


Figure 45 – Simulated vs Experimental Results: First Bin Residuals for LFF Device

The calculations for both simulated and measured data show roughly the same shallow trend, though they differ by about an order of magnitude. This is because the simulation ignores the first bin in its matching criteria, and the discrepancy illustrates that while multiple combinations of inputs will give the same mean DCR, not as many will match both the mean DCR and the first bin residual data. It is also worth noting that the simulation is very precise and has no noise in arrival times, no delay between avalanche initiation and avalanche detection, and no noise associated with circuit logic (small variations in actual values of t_a and t_q). The measured data has all of these noise sources, which leads to error associated with the first bin. Overall, the gradual decrease in first bin residual as quench time increases, as well as the low contribution across all quench times, is expected due to the robust scupper region and small area and volume associated with the LFF device architecture.

Figure 46 shows the simulated and experimental results for the MFF device on a log-linear plot, with the percentage value on the left y-axis and the numerical output on the right.

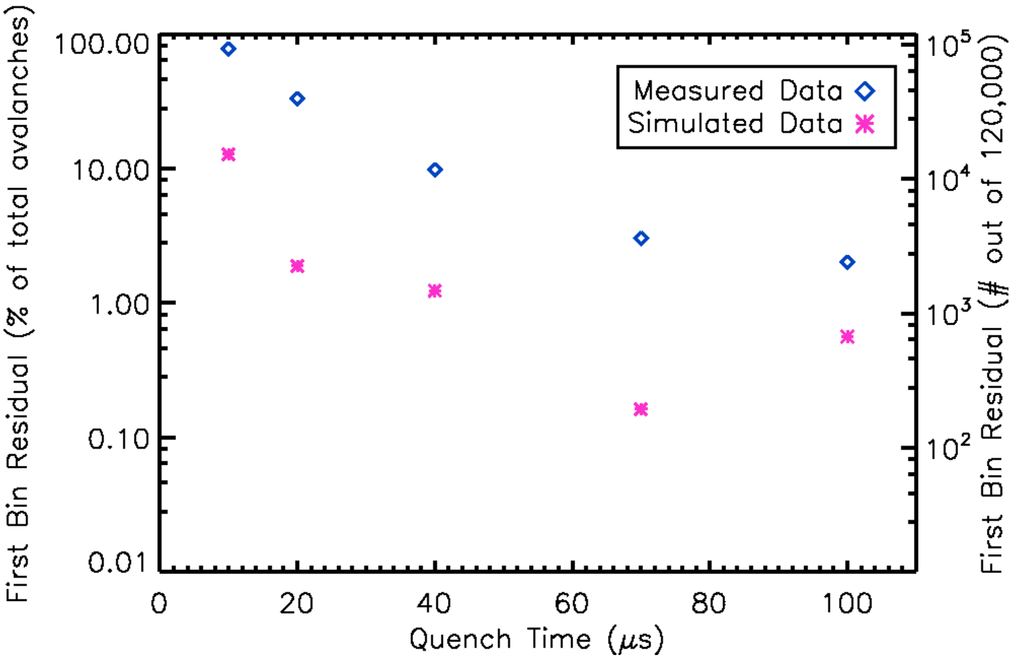


Figure 46 – Simulated vs Experimental Results: First Bin Residuals for MFF Device

Again, the simulated data shows a trend roughly one order of magnitude less than the measured data, yet with the same overall relationship to quench time. The magnitude difference could be due to the noise associated with the experimental data, as posited based on the LFF first bin residual results. The steeper trending associated with the MFF device is present in the simulated data, which makes sense based on the simulation's best-fit input values (high number of traps and high P_{phot}).

Figure 47 shows the simulated and experimental results for the HFF device on a linear-linear plot, with the percentage value on the left y-axis and the numerical output on the right.

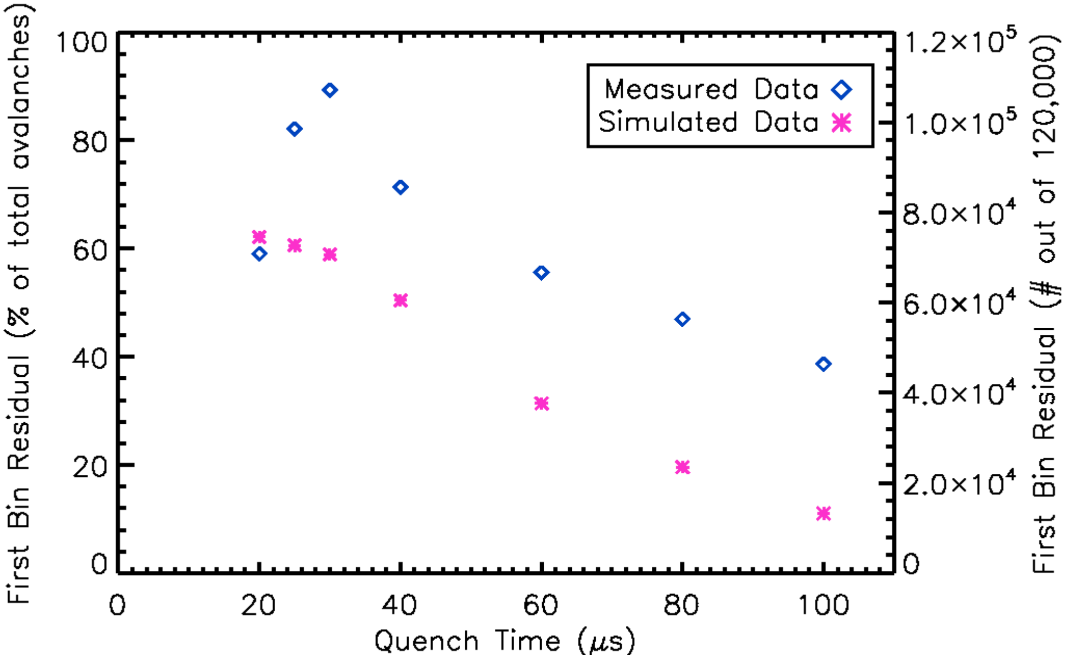


Figure 47 – Simulated vs Experimental Results: First Bin Residuals for HFF Device

The overall match of the fit is not very good, even though the DCR vs quench time trend was a good fit with these inputs. The magnitude is more correct than the results for either the LFF or MFF devices. The trend is reasonable for the device, but even though the data seems to reach a limit at shorter quench times, the simulated data for this particular set of inputs does not mimic the downward trend seen in the measured data (when afterpulsing begins to dominate the first bin contribution).

A similarly good fit (with slightly different input values shown in Table 8) within the degenerate congregation of chi-squared values (see Figure 41) does render the downward trend as seen in Figure 48. This shows that the simulation matching could be improved (and the degeneracy of the chi-squared metric decreased) by adding the first bin residual criteria to the matching evaluation.

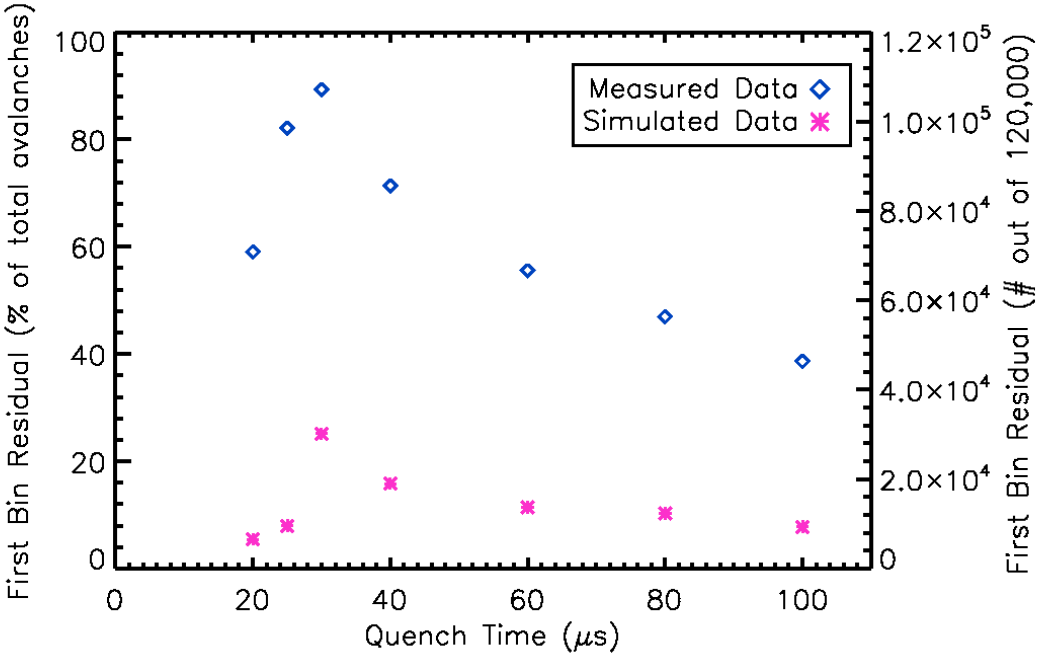


Figure 48 – Alternate Simulated vs Experimental Results: First Bin Residuals for HFF Device

Table 8: Alternate HFF Simulation Match Input Values

Parameter	**Value**
Avalanche Initiation Probability	0.8
Thermal DCR	42.5e4 kHz
Number of Filled Traps	275
Detrapping Lifetime	50 μs
Carrier Lifetime	80 μs
P_{phot}	0.248
χ^2 (normalized data)	0.041

6 CONCLUSIONS

Overall, the results matched expectations in both performance for individual devices and the relative performance between devices of different fill factors. DCR vs quench time curves demonstrated the previously observed trends measured at MIT Lincoln Labs. The first bin residual method of characterization for the self-retriggering mechanism also matched expectations based on a physics-based understanding of the phenomenon. Simulation-based matching yielded reasonable results for specific device parameters, though refinement of the best-fit methodology is necessary.

6.1 DCR VS QUENCH TIME

Figure 49 shows the DCR vs quench time trends for the LFF, MFF, and HFF devices on a log-linear plot.

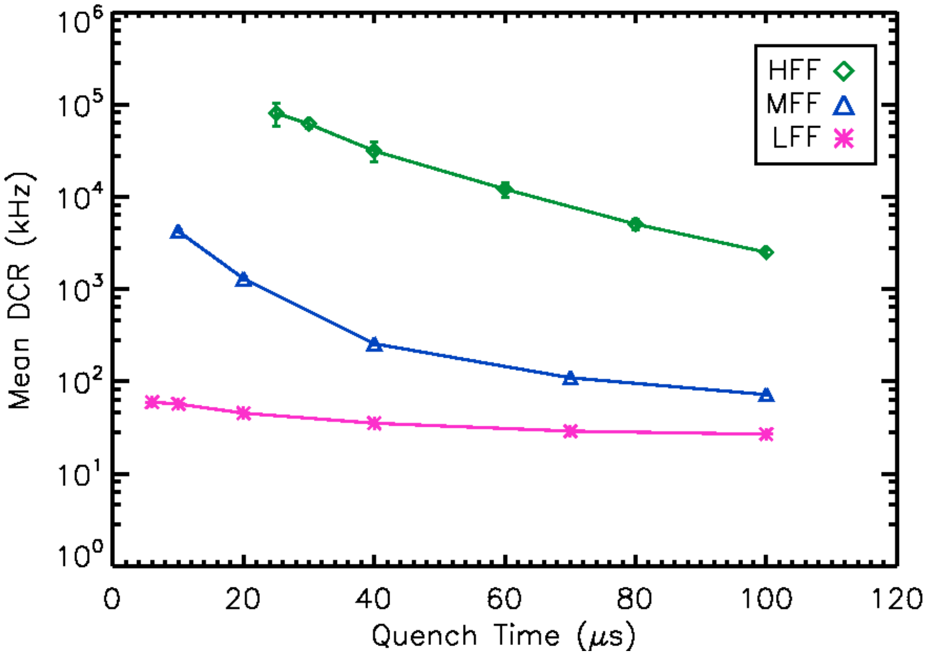


Figure 49 – Overall Experimental Results for DCR vs Quench Time

When viewed on the same set of axes, the devices are unique in both magnitude and trend in DCR as a function of quench time. The lines shown are not fit lines; they only serve to clarify the plot.

As expected, the LFF device has a very gradual upward trend at shorter quench times and has the lowest overall DCR measurements of the three devices. This indicates that the LFF device has the lowest thermal DCR and the lowest contributions from afterpulsing and self-retriggering.

The MFF device shows a markedly steeper upward trend at shorter quench times as well as a higher base thermal DCR. As expected, this indicates that with larger fill factor comes more thermal carrier generation, afterpulsing, and self-retriggering by virtue of the larger area and volume of the device's active area.

The HFF device displays the highest base thermal DCR as well as a steep increase in DCR at all quench time measurements, indicating that the 100 µs maximum is not yet sufficient to quell the non-steady-rate contributing processes. Afterpulsing looks to be a significant contributor, but the self-retriggering mechanism is also an obvious contributor due to the high first bin residuals recorded. The lack of measurements at shorter quench time values are due to loss of resolution of arrival times (the vast majority of recorded times fall within the first bin). In order to make measurements at the lower end of the AQB's capabilities for the HFF device, t_a should be shortened, though not so much that it compromises the stability of the AQB circuit. Varying the length of the arm pulse is a worthwhile set of experiments to do in the future to further characterize the devices.

From the DCR vs quench time relationships seen in Figure 49, the HFF device clearly has the poorest performance of the three devices tested. This is due in part to the increase in volume that leads to an increase in thermal DCR, which is unavoidable, though a decrease in operating temperature improves this effect. The time-dependent processes of afterpulsing and self-retriggering also plague the device, however. Increased precautions in processing and architectural design changes may mitigate afterpulsing. Limiting the magnitude of the avalanches, thus reducing the number of photons released during avalanche, could lessen self-retriggering. Redesigning the device to bolster the effectiveness of the scupper region in the higher fill-factor devices is also a viable improvement, since it is evident that it greatly improves the DCR performance of the APD.

6.2 FIRST BIN RESIDUALS

Figure 50 shows the first bin residual data as measured for all three devices on a log-linear plot. The increase in first bin contributions from self-retriggering carriers at shorter quench times is evident across the three devices.

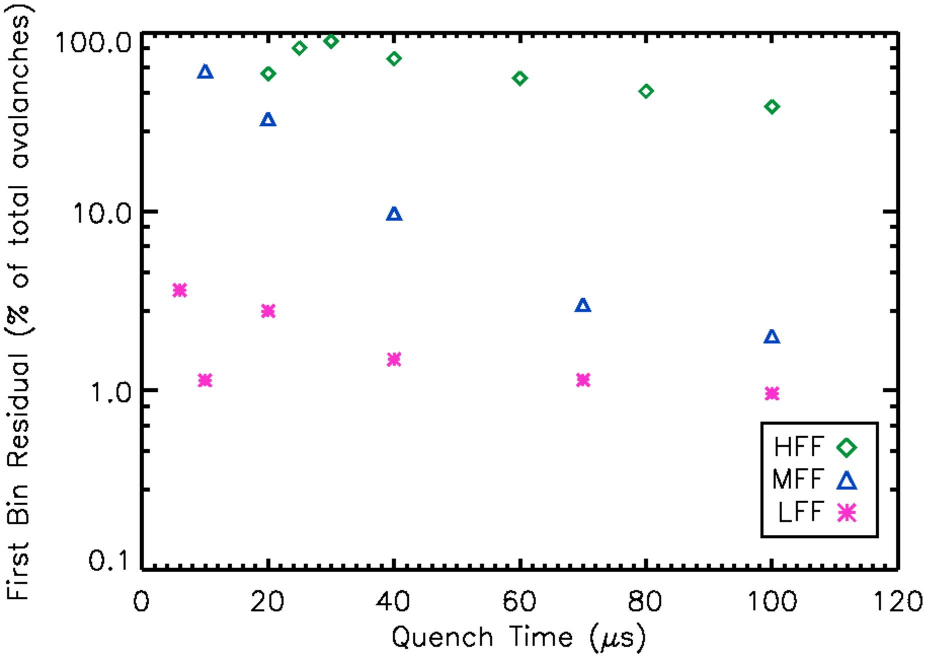


Figure 50 – Overall Experimental Results for First Bin Residuals

The LFF device shows a small magnitude with limited slope across the quench times, which indicates that self-retriggering carriers are not major contributors to the measured DCR. The first bin residual plotted is out of 90,000 possible counts, averaging about a 2 % contribution overall.

The MFF device shows an increase in both magnitude and slope of the first bin residual as a function of quench time. The percentage of counts attributed to self-retriggering (those counts that occur immediately and would not have occurred statistically based on the afterpulsing and thermal DCR trends) is higher as well, with the highest contribution reaching 61 %.

The HFF device shows a different trend when compared to the other two devices. It has a shallower slope than the MFF device, and at the shortest quench times, it exhibits a downward trend. At these short quench times, afterpulsing dominates the DCR of the device. Since the first bin residual is only a measurement of the avalanches that occur due to self-retriggering and would not have occurred otherwise (an "exclusive or" function), a downward trend indicates that another mechanism becomes more prominent in the first (immediate avalanche) bin than self-retriggering at short quench times. Based on the fundamental physics that govern the self-retriggering carriers, their numbers must increase at shorter quench times. However, it appears that the number of carriers released from traps increases at a higher rate.

Based on the analysis of the trends in Figure 50, the MFF device may have the most to gain from reduction of the self-retriggering contribution. Though the HFF device has a large contribution from the mechanism, the afterpulsing seems to be very severe, which may rule out the use of the device even if the self-retriggering contribution were completely removed.

6.3 SIMULATION

Overall, the simulation performed well in matching the measured DCR data for the three devices. By ranging multiple inputs over reasonable values, the simulation compared its output for each unique combination to the DCR values calculated from the measured data using a chi-squared test. The results were somewhat degenerate, with multiple combinations of inputs yielding the same mean DCR values.

The first bin residual simulation results did not match the measured data as well as the DCR portion because that output was not included in the chi-squared best-fit evaluation. While the results for the first two devices were roughly one order of magnitude smaller than the measured contribution, the HFF simulation inputs yielded results much closer to the measured data. Upon further investigation, a fit with a nearly identical chi-squared value modeled the trend very well, though it lacked in magnitude. The first bin residual simulated data for the alternate fit did confirm a downward trend in the first bin residual for the HFF device at shorter quench times.

This differentiation led to the conclusion that, while different combinations of inputs may yield the same mean DCR values, results with very similar chi-squared values have significantly different first bin residual results. Based on this observation, it is reasonable to assume that including a metric of evaluation regarding the first bin analysis for each device in combination with the mean DCR calculation will significantly decrease the degeneracy of the fit model.

7 CONTRIBUTIONS

The author would like to acknowledge those who aided in the definition, development, and realization of this thesis. Without the support of these people, this thesis would not have been as complete.

Thanks to Dr. Don Figer, Director of the Center for Detectors and the student's advisor, for aiding in the inception of this project and for giving valuable guidance and insightful suggestions throughout the project's duration.

Thanks to Dr. Brian Aull, of MIT Lincoln Labs, for allowing the student to visit, learn, and participate in the outstanding research conducted in his lab.

Thanks to Dr. Karl Hirschman of RIT's Microelectronic Engineering Department and Dr. Zoran Ninkov of RIT's Center for Imaging Science for serving on the student's thesis committee and for helping the student to focus the efforts of this thesis.

Thanks to Dr. Joong Lee of RIT's Center for Detectors for sharing his expertise in statistical modeling and for his crucial assistance regarding the derivation of certain physics equations.

Thanks to Christine Trombley, graduate student in the Astrophysics PhD program at RIT, for her input on simulation approaches, statistics, and for many thought-provoking discussions regarding fundamental physics and methodology.

Thanks to John Breese, B.S. / M.S. student in the Computer Engineering program at RIT, for his invaluable help in interface coding and communication between IDL and the Agilent oscilloscope used for the experiments in this thesis.

Thanks to Don Stauffer, engineer at the Center for Detectors, for his help in circuit analysis and modeling.

Thanks to Tom Montagliano, engineer at the Center for Detectors, for his help and input regarding practical testing setup and use of characterization meters.

Thanks to Brandon Hanold, engineer at the Center for Detectors, John Frye, graduate student in the Computer Engineering master's program at RIT, and Chris Maloney, graduate student in the

Microsystems Engineering PhD program at RIT, for their unique insights into the progress and evolution of this thesis, especially pertaining to statistical simulations and their integrity with respect to the physical processes they attempt to represent.

8 FUTURE WORK

This thesis project could progress, given additional time and resources, in a few areas. Supplementary testing scenarios would further distinguish the various dark count sources, leading to better overall characterization of the devices. Aside from additional measurements, changes to the simulation methodology and fit criteria will also lead to more unique results. Possible improvements to the simulation portion of this thesis include the addition of a physics-based simulation (such as with Silvaco software) as well as expansion of the existing simulation to streamline the input process and incorporate more output constraints.

Though the testing described in this thesis is valuable, the measurements did not utilize the full range of AQB settings. For example, lower quench voltage levels should lead to a reduction in time-dependent dark counts but should not affect the thermal DCR, leading to more resolution between the different dark count contributors. Another possible variation is the arm voltage. Increasing the overbias on the APD increases the avalanche initiation probability and therefore leads to a higher DCR. The increase in overall DCR at higher overbias values will give insight into the tradeoff between higher photon detection efficiency and higher detector noise. In addition, shorter arm times lead to less build-up in the first bin (and shorter lengths of time where avalanches are artificially sustained), allowing better separation of self-retriggering counts from thermal and afterpulsing counts.

Beyond the settings pertaining to the AQB circuit, external device conditions can be altered. Temperature-varied measurements affect different dark count sources differently. Thermal DCR, for example, will decrease with decreasing temperature because the carriers' movement relies on thermal energy. Afterpulsing would become more prominent, however, and the self-retriggering mechanism should also become more prominent as the carrier lifetime lengthens at lower temperatures.

Another worthwhile improvement to this project would be the inclusion of a physics-based simulation. This would allow a more direct extraction of diode capacitance and electric field patterns, and may provide a better estimate for some parameter inputs, e.g. P_{phot}, based on the device architecture.

There is also room for improvement in the existing simulation setup and fit evaluation. As discussed above, the chi-squared test only evaluates matched pairs of the simulated and measured DCR results. The first bin analysis is not included in the best-fit metric, but the first bin residual values vary greatly between sets of input parameters that yield equivalent chi-squared values. This suggests that the degeneracy evidenced in the chi-squared value distribution will decrease if the simulation incorporates a second constraint on its best fit analysis. Another way to improve the simulation may be to add noise sources (timing jitter, APD voltage decay due to leakage, etc.) to the simulation. The resulting output would more closely resemble the raw measured data. This would require characterization of the actual noise sources present in the AQB measurement system.

While continuing to develop the testing and evaluation methodology, newer iterations of the APD architecture could be evaluated for improvements and the effect of targeted changes. This additional testing will give valuable feedback to the device designers and lead to further improvements in design and performance.

REFERENCES

Ando, Hiroaki, and Hiroshi Kanbe. "Effect of Avalanche Build-Up Time on Avalanche Photodiode Sensitivity." *IEEE Journal of Quantum Electronics*, 1985: 251-255.

Ardt, Jan Van. *Lidar over RIT's campus.* 2008. http://ipler.cis.rit.edu/node/24 (accessed July 5, 2011).

Aull, Brian F., Daniel R. Schuette, Robert K. Reich, and Robert L. Johnson. "Adaptive optics wavefront sensors based on photon-counting detector arrays." *Proceedings of SPIE* v7736 (2010).

Blazej, Josef, Ivan Prochazka, Karel Hamal, Bruno Sopko, and Dominik Chren. "Gallium-based avalanche photodiode optical crosstalk." *Nuclear Instruments & Methods in Physics Research*, 2006: 239-241.

Craven-Bartle, Thomas V., Reinhold J. Dorn, and James W. Beletic. "Computer simulation comparison of CCDs and APDs for curvature wavefront sensing." *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* 4007 (July 2000): 444-451.

Grein, M. E., et al. "Efficient communication at telecom wavelengths using wavelength conversion and silicon photon-counting detectors." *Proceedings of SPIE* v6709, 670910-1 (2010).

Gullikson, E. M., E. Gramsch, and M. Szawlowski. "Large-area avalanche photodiodes for the detection of soft x rays." *Applied Optics*, 1995: 4662-4668.

Jaeger, Richard C., and Travis N. Blalock. *Microelectronic Circuit Design.* New York City: McGraw-Hill, 2004.

Kay, Steven M. *Intuitive Probability and Random Processes using MATLAB.* New York City, NY: Springer, 2005.

Kindt, W.J., and K.J. de Langen. "Integrated readout electronics for Geiger mode avalanche photodiodes." *Solid-State Circuits Conference*, 1998: 216-219.

McIntyre, Robert J. "On the Avalanche Initiation Probability of Avalanche Diodes Above the Breakdown Voltage." *IEEE Transactions on Electron Devices* Vol. ED-20, no. No. 7 (July 1973): 637-641.

PerkinElmer, Inc. "Avalanche photodiode: A User Guide." *Perkin Elmer.* 2006. http://www.optoelectronics.perkinelmer.com (accessed September 21, 2009).

Pierret, Robert F. *Semiconductor Device Fundamentals.* Reading, Mass: Addison-Wesley Publishing Company, 1996.

Renker, D. "Geiger-mode avalanche photodiodes, history, properties, and problems." *Nuclear Instruments & Methods in Physics Research* Section A, no. 567 (2006): 48-56.

Shack, R. B., and B. C. Platt. "Production and use of lenticular Hartmann screen." *Journal of the Optical Society of America* v61, no. 5 (1971): 656.

Squillante, Michael, et al. "Recent Advances in Very Large Area Avalanche Photodiodes." *Proceedings of SPIE* v5071 (2003): 405-410.

Van Zeghbroeck, B. *Chapter 4: p-n Junctions.* 2007. http://ecee.colorado.edu/~bart/book/book/chapter4/ch4_2.htm (accessed September 21, 2009).

Yuan, Ping, et al. "32 x 32 Geiger-mode LADAR Cameras." *Proceedings of SPIE* v7648, 76840C-1 (2010).

APPENDIX A: List of Acronyms and Definitions

Acronym	Definition
(GM) APD	(**G**eiger-**M**ode) **A**valanche **P**hoto**D**iode
AO	**A**daptive **O**ptics
AQB	**A**ctive **Q**uench **B**oard
CCD	**C**harge-**C**oupled **D**evice
CMOS	**C**ompensated **M**etal-**O**xide-**S**emiconductor
DCR	**D**ark **C**ount **R**ate
HFF	**H**igh **F**ill **F**actor
LADAR	**LA**ser **D**etection **A**nd **R**anging
LFF	**L**ow **F**ill **F**actor
MFF	**M**edium **F**ill **F**actor
PDE	**P**hoton **D**etection **E**fficiency
PMT	**P**hoto**M**ultiplier **T**ube
SNR	**S**ignal to **N**oise **R**atio

APPENDIX B: GM APD P-I-N Diode Capacitance Derivation

To derive a more correct expression for the GM APD multiplication region diode capacitance, the process must begin with the fundamentals. Any characteristic equation for a specific shape's capacitance begins with the basic relationship shown in Eq. 29.

$$\boldsymbol{C = \frac{q}{V}} \qquad \text{Eq. 29}$$

In order to characterize the expected capacitance, the derivation must start with Gauss' Law (Eq. 30), which describes the charge enclosed inside of a Gaussian surface (the numerator in Eq. 29). Such a surface must enclose a three-dimensional space, and the flux of an electromagnetic field through the surface describes the interaction with any element inside of it. The Gaussian surface used in this case, based on the shape of the depletion region, is shown in Figure 51.

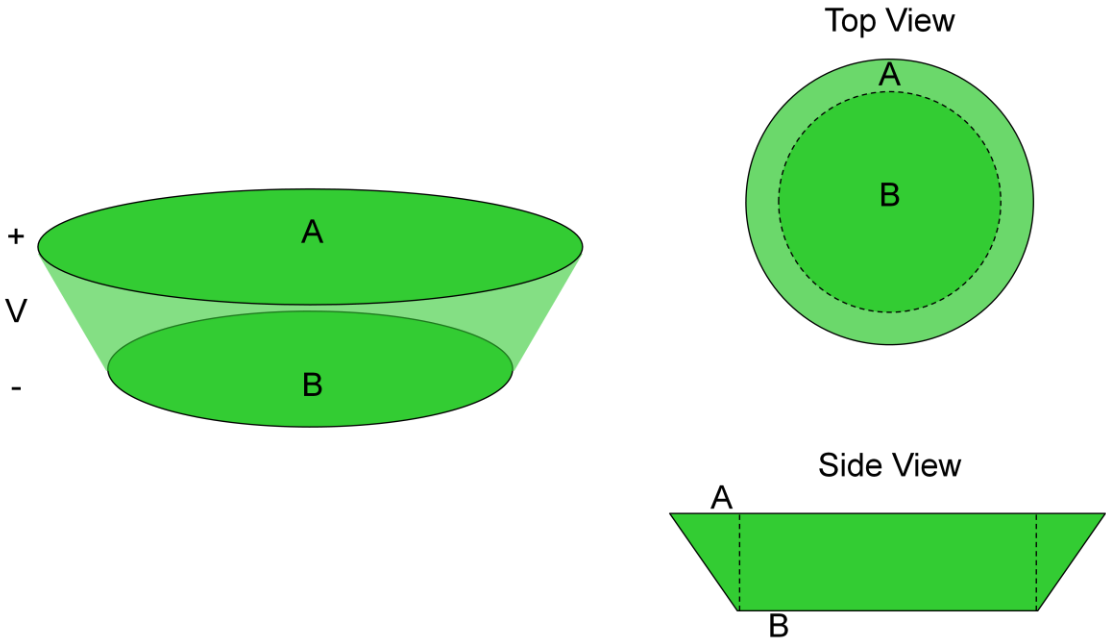


Figure 51 – Gaussian Surface for Diode Capacitance Calculations

Plate A represents the positive node of the diode and plate B the negative node, which is the case when the diode is reverse-biased. This delineation is important because it determines the originating and terminating points of the field lines inside the structure. For the LFF device, plate A has a significantly larger diameter than plate B. As the fill factor increases, these two plates become more similar in size (as the scupper region shrinks). The parallel plate (non-infinite) assumption is appropriate in this case because the multiplication region has the structure of a p-i-n diode, though its operation is as a GM APD multiplier.

Gauss' Law (Eq. 30) states that the electric charge enclosed in a Gaussian surface is equal to the permittivity of the capacitive material multiplied by the integral of the dot product of the electric field at any point and the corresponding normal vector to the cross-sectional area.

$$\boldsymbol{q = \varepsilon_0 \varepsilon_r \oint \vec{E} \cdot d\vec{A}} \qquad \text{Eq. 30}$$

In order to use Gauss' Law, the electric field must first be characterized as a function of location inside the Gaussian surface. Since the shape is symmetrical about the center point along a vertical axis, cylindrical coordinates are ideal. The electric field at any one point in space is equal to the sum of all the electric force contributions from the charged surfaces of the capacitor (the two disks in this case). The magnitude of the electric field can be derived using Coulomb's Law for a point charge, and is shown in Eq. 31. Note that the vectors will have multidimensional components, and so then must their sum.

$$\boldsymbol{\vec{E} = \frac{q}{4\pi\varepsilon_0\varepsilon_r r^2}\hat{r}} \qquad \text{Eq. 31}$$

r is the distance between the evaluation point and the contributing point. An integral function efficiently sums all the forces at a single point. To simplify the integration, the equation uses the distance r', which represents a trigonometric calculation that uses the known parameters of the Gaussian surface. Figure 52 shows both top and perspective views of the geometries utilized to derive the function of the electric field at any point. The geometry takes advantage of symmetry when the coordinate system centers at the point in question, rotating the plate so that one of the axes falls on a line of symmetry on the disk.

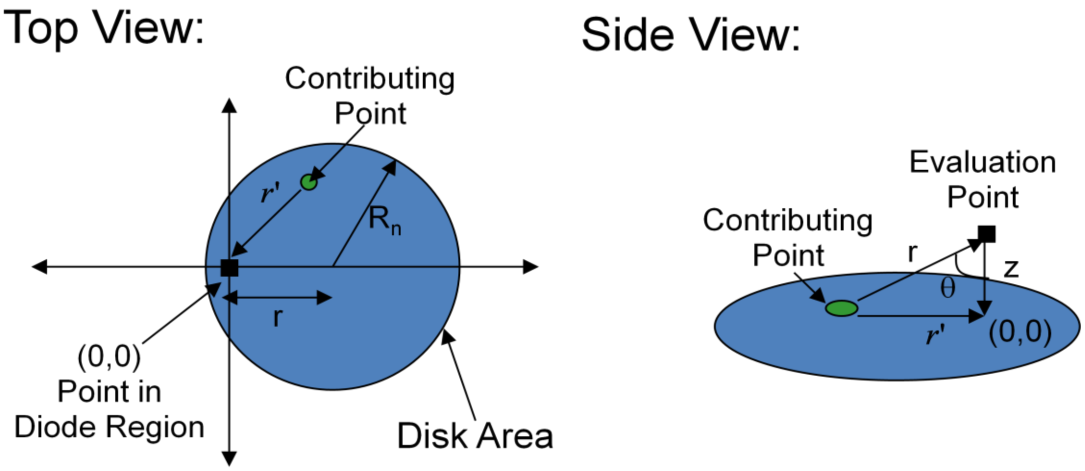


Figure 52 – Frame of Reference for Electric Field Contribution Function Calculation

r is the straight-line distance between the evaluation point and the center of the contributing disk, R_n is the radius of the contributing disk, r’ is the radial distance between the contributing point and the evaluation point in the diode region, and z is the vertical distance between the evaluation point and the contributing disk. The function of the total electric field at any evaluation point can be expressed in terms of r’ and known geometric constants according to Eq. 32 for the case of a single disk.

$$\vec{E}(z,r,\emptyset) = \frac{\rho_{s-n}}{4\pi\varepsilon_0\varepsilon_r}\int_{\emptyset=0}^{2\pi}\int_{r'=0}^{r_{n-max}}\frac{r'}{(r'^2+z^2)^{3/2}}dr'd\emptyset\left(z\cos\emptyset\,\widehat{r'}+r'\hat{z}\right) \qquad \text{Eq. 32}$$

ρ_{s-n} is the surface charge density on plate “n,” and r_{n-max} is the maximum radial distance between the evaluation point and any point on the contributing disk, given by Eq. 33.

$$r_{n-max} = r\cos\emptyset + \sqrt{{R_n}^2 - r^2 + r\cos^2\emptyset} \qquad \text{Eq. 33}$$

By superposition, the electric field function can be derived for each plate separately and then added together (keeping track of vector dimensionality) to arrive at an expression for the total contribution at any point in the space inside the Gaussian Surface. Since the plates are oppositely charged, but also oppositely oriented in space from the perspective of any evaluation point, the resulting effect is the addition of the contributions in the vertical (z) direction. The only difference between the two parts of the sum will be the upper limit to r', since the disks are of unequal size, and the charge concentration (since they must have equal total accumulated charge but are of unequal size).

Now that an expression for the electric field at any point inside this Gaussian surface exists, Gauss' Law (Eq. 30) will be revisited. Since the normal vector to the cross-sectional area is always purely in the direction of z, the dot product between that vector and any other will be equal to the z-direction component of the other vector. A dot product is the projection of one vector onto another. Since the cross-sectional area normal vector exists in a singular dimension (the z-direction), the electric field argument reduces to its z-direction component. Eq. 34 shows the resulting expression.

$$\boldsymbol{q = \varepsilon_0 \varepsilon_r \int_{z=0}^{d} \left(\int_{r=0}^{R(z)} \int_{\emptyset=0}^{2\pi} E_{\hat{z}}(z, r, \emptyset) d\emptyset dr \right) A(z) dz} \quad \text{Eq. 34}$$

d is the separation between the two disks, R(z) is the height-dependent radius of the cross-sectional area of the Gaussian surface, and A(z) is the height-dependent cross-sectional area. Since the radius (and therefore the cross-sectional area) is not constant throughout the depth of the capacitor, the solution must also integrate over height. The expressions for R and A are shown in Eq. 35 and Eq. 36.

$$\boldsymbol{R(z) = R_1 + z\left(\frac{R_2 - R_1}{d}\right)} \quad \text{Eq. 35}$$

$$\boldsymbol{A(z) = \pi\left(R_1 + z\left(\frac{R_2 - R_1}{d}\right)\right)^2} \quad \text{Eq. 36}$$

These equations complete the components necessary for an expression describing the total enclosed charge in the Gaussian surface. The next step is to determine the potential across the depth of the capacitor (the denominator if Eq. 29) according to Eq. 37.

$$\boldsymbol{V = -\int_{s} \vec{E} \cdot d\vec{s}} \qquad \text{Eq. 37}$$

To simplify the equation, a path $\boldsymbol{d\vec{s}}$ is chosen to be parallel to the electric field at all times, following the field lines. As such, the $\boldsymbol{d\vec{s}}$ vector will always be parallel and opposite to the electric field, and the dot product of the two is the negative product of the magnitude of the electric field and the path length. With this assumption, Eq. 38 expresses the potential difference across the diode as a function of position in space.

$$\boldsymbol{V = \int_{r=0}^{R_1} \left(\int_{z=0}^{d} \int_{\emptyset=0}^{2\pi} \left|\vec{E}(z,r,\emptyset)\right| d\emptyset dz \right) s(r) dr} \qquad \text{Eq. 38}$$

The limits for integration over the radius-dependant path length go from zero to R$_1$, which is the maximum point of origin for a field line in the capacitive region. Returning to Eq. 29 and substituting in the expressions derived in Eq. 34 and Eq. 38, the full expression for the diode capacitance can be expressed as in Eq. 39.

$$\boldsymbol{C = \frac{\varepsilon_0 \varepsilon_r \int_{z=0}^{d} \left(\int_{r=0}^{R(z)} \int_{\emptyset=0}^{2\pi} E_{\hat{z}}(z,r,\emptyset) d\emptyset dr \right) A(z) dz}{\int_{r=0}^{R_1} \left(\int_{z=0}^{d} \int_{\emptyset=0}^{2\pi} \left|\vec{E}(z,r,\emptyset)\right| d\emptyset dz \right) s(r) dr}} \qquad \text{Eq. 39}$$

This equation is correct, but there is difficulty in the calculation due to the s(r) term in the denominator. That term represents the path length from the positive plate to the negative plate along the field line originating at point r along the radius of the disk (the lines are symmetrical with respect to the angle around the disk). This is difficult to calculate in practice, and so another approach is discussed where some assumptions and approximations simplify the calculation.

Recall the side view of the Gaussian surface from Figure 51. The shape can be re-defined as a center cylinder of radius R$_B$ (the radius of the lower plate) surrounded by another area characterized by the triangular vertical cross-section of the area outside the dashed lines. The two areas are now capacitors in parallel (neglecting any fringing or edge effects). Capacitors in parallel add magnitudes linearly, so the following calculation addresses the two capacitors independently and adds them together in the final step.

The capacitance of the internal cylinder, assuming that the cross-sectional area remains constant throughout the depth of the device and that the field lines are perpendicular it, is a simple calculation. Recall that when these assumptions hold true, the result is the expression in Eq. 22. The second capacitive region is more difficult to characterize, but it is possible given a few assumptions. To begin, consider a very small portion of the outer capacitive area as shown in Figure 53.

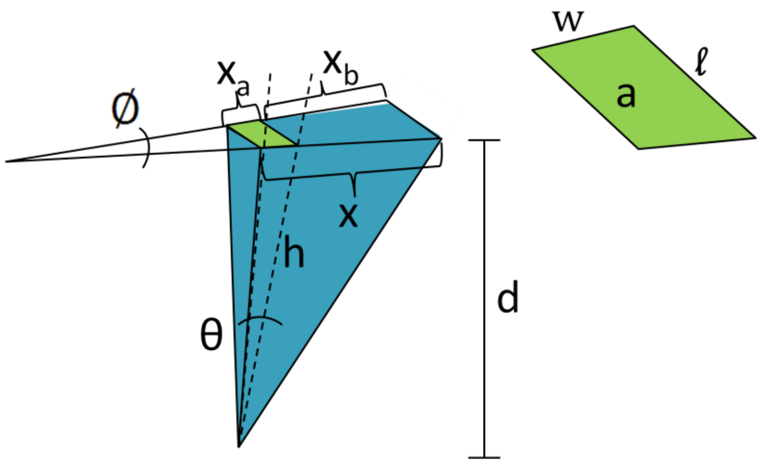


Figure 53 – Derivative Section of the Outer Capacitive Area

This shape represents a very small slice of the outer capacitive area, and is the basis for integration over the whole capacitor section. The pop out section in green is the integrating section that represents the cross-sectional area based on input values. $\emptyset$ is the angle between two lines originating at the origin (the center of the Gaussian surface where $r = 0$), and θ is the angle from the dotted line in the side view seen in Figure 51 to the vector that corresponds to the straight-line distance between the edge of plate B and any radial point on the outer ring of plate A. x_a and x_b change dependent on θ, h is the straight-line distance between the incremental area and the outer edge of plate B, and x and d are geometric constants of the Gaussian surface. The quantities w and ℓ can be defined in terms of the integrating variables ($\emptyset$ and θ) using arc length calculations and trigonometric relationships. The expressions derived for a and h using these integrating variables are shown in Eq. 40 and Eq. 41.

$$\boldsymbol{a = \emptyset(R_B + d\tan\theta)d\tan\theta} \qquad \text{Eq. 40}$$

$$\boldsymbol{h = \frac{d}{\cos\theta}} \qquad \text{Eq. 41}$$

Each small increment of ∅ and θ represent a tiny capacitor, and so they can be added together by integration since they are parallel to each other. For very small values of ∅ and θ, the cross-sectional area down through the surface defined by “a” is assumed to be constant, and the field lines are assumed to be perpendicular to the cross-sectional area normal vector for small areas. So in this case, based on Eq. 29 and the new solutions for Gauss’ Law (Eq. 30) and Eq. 37 derived with new assumptions, the expression for the diode’s capacitance is as shown in Eq. 42.

$$\boldsymbol{C = \frac{\varepsilon_0\varepsilon_r\left(\pi {R_B}^2\right)}{d} + \varepsilon_0\varepsilon_r \int_{\emptyset=0}^{2\pi}\int_{\theta=0}^{\tan^{-1}\left(\frac{R_A-R_B}{d}\right)} \frac{a(\theta,\emptyset)}{h(\theta)}\, d\theta d\emptyset} \qquad \text{Eq. 42}$$

The upper limit for integration of θ is the maximum angle based on the fixed geometry of the Gaussian surface. This expression is actually a slight over-estimation of the capacitance based on the assumption that the field lines are parallel to the vector along h. In reality they bend, which diminishes the dot product of the electric field with the cross-sectional area normal vector.

APPENDIX C: C-V Measurements

Capacitance-Voltage Measurements for Large-Area Diode

Frequency = 2.0 kHz Doping profile:
Auto circuit mode Area = 0.02 cm^2
Slow meas. Speed Cox =1.0 F
Low test signal level
Auto range on

Table 9: CV Measurements for Large-Area Diode

Voltage (V)	Capacitance (pF)	Q	Doping (cm^{-3})	Depth (um)
0.00554010	551.000	100.000	1.11307e+016	0.376021
0.246772	506.000	100.000	1.27285e+016	0.409462
0.495524	474.000	100.000	1.48844e+016	0.437105
0.746524	451.000	100.000	1.70327e+016	0.459396
0.994354	433.000	100.000	1.85440e+016	0.478494
1.24333	418.000	200.000	2.05703e+016	0.495665
1.49528	406.000	200.000	2.19913e+016	0.510315
1.74408	395.000	200.000	2.46452e+016	0.524526
1.99516	387.000	100.000	2.71947e+016	0.535369
2.24314	379.000	200.000	2.56085e+016	0.546670
2.49510	371.000	100.000	2.77300e+016	0.558458
2.74338	365.000	100.000	3.05035e+016	0.567638
2.99501	359.000	100.000	3.17345e+016	0.577125
3.24186	354.000	100.000	3.33185e+016	0.585276
3.49373	349.000	200.000	3.57315e+016	0.593661
3.74143	345.000	200.000	3.42842e+016	0.600544
3.99381	340.000	100.000	3.30687e+016	0.609376
4.24132	336.000	100.000	3.56614e+016	0.616630
4.49319	332.000	100.000	3.95618e+016	0.624060
4.74051	329.000	100.000	3.81629e+016	0.629750
4.99228	325.000	100.000	3.72860e+016	0.637501
5.24276	322.000	200.000	4.20898e+016	0.643440
5.49453	319.000	100.000	4.07066e+016	0.649492
5.74253	316.000	200.000	3.95007e+016	0.655658
5.99323	313.000	200.000	4.64015e+016	0.661942
6.24168	311.000	300.000	4.53477e+016	0.666199
6.49441	308.000	100.000	4.44200e+016	0.672688
6.74219	306.000	100.000	4.30158e+016	0.677084
6.99374	303.000	300.000	4.21228e+016	0.683788
7.24082	301.000	0.000000	5.13270e+016	0.688332
7.49362	299.000	0.000000	5.04827e+016	0.692936

7.74253	297.000	0.000000	4.94105e+016	0.697602
7.99450	295.000	0.000000	3.82029e+016	0.702332
8.24136	292.000	0.000000	3.74431e+016	0.709547
8.49322	290.000	0.000000	4.58631e+016	0.714441
8.74095	288.000	0.000000	4.49657e+016	0.719402
8.99314	286.000	0.000000	5.93637e+016	0.724433
9.24097	285.000	0.000000	5.81528e+016	0.726975
9.49321	283.000	0.000000	4.26393e+016	0.732112
9.74080	281.000	0.000000	4.19926e+016	0.737323
9.99593	279.000	0.000000	5.25406e+016	0.742609
10.1959	278.000	0.000000	5.68375e+016	0.745280
10.4946	276.000	0.000000	5.59385e+016	0.750681
10.6951	275.000	0.000000	5.50418e+016	0.753410
10.9934	273.000	0.000000	3.56847e+016	0.758930
11.1972	271.000	0.000000	3.52100e+016	0.764531
11.4963	269.000	0.000000	1.10550e+017	0.770215
11.6966	269.000	0.000000	7.14080e+016	0.770215
11.9950	266.000	0.000000	4.06261e+016	0.778902
12.1949	265.000	0.000000	4.93053e+016	0.781841
12.4945	263.000	0.000000	4.87531e+016	0.787786
12.6974	262.000	0.000000	6.34821e+016	0.790793
13.0003	261.000	0.000000	3.63175e+016	0.793823
13.2010	259.000	0.000000	3.03983e+016	0.799953
13.5008	257.000	0.000000	4.51261e+016	0.806178
13.7000	256.000	0.000000	5.81711e+016	0.809327
13.9980	255.000	100.000	5.76849e+016	0.812501
14.1991	254.000	0.000000	4.36351e+016	0.815700
14.5000	252.000	0.000000	4.28207e+016	0.822174
14.7012	251.000	100.000	5.52402e+016	0.825449
14.9989	250.000	0.000000	3.18764e+016	0.828751
15.1992	248.000	0.000000	3.16601e+016	0.835435
15.4984	247.000	0.000000	3.08236e+016	0.838817
15.6993	245.000	0.000000	3.06189e+016	0.845665
15.9979	244.000	0.000000	5.06505e+016	0.849130
16.1996	243.000	0.000000	3.78194e+016	0.852625
16.4946	241.000	100.000	3.73848e+016	0.859701
16.6997	240.000	0.000000	4.89702e+016	0.863283
16.9958	239.000	0.000000	4.71808e+016	0.866895
17.1957	238.000	0.000000	4.69603e+016	0.870537
17.4960	237.000	100.000	2.73884e+016	0.874210

APPENDIX D: Raw Measured Data

Raw Data for DCR vs Quench Time and First Bin Residuals Measurements

Table 10: DCR Values - LFF

Quench Time (µs)	**Mean DCR (kHz)**		
6	64.8	58.6	55.8
10	59.9	56.4	54.7
20	45.5	45.0	45.7
40	35.6	34.6	35.8
70	29.7	28.3	29.0
100	25.8	26.1	28.5

Table 11: DCR Values - MFF

Quench Time (µs)	**Mean DCR (MHz)**			
10	4.23	4.38	4.20	4.19
20	1.38	1.28	1.27	1.27
40	0.266	0.254	0.256	0.250
70	0.109	0.110	0.111	0.108
100	0.0722	0.0733	0.0715	0.0716

Table 12: DCR Values - HFF

Quench Time (µs)	**Mean DCR (MHz)**			
25	103.0	55.7	95.7	69.1
30	67.9	58.4	64.9	54.8
40	37.4	25.3	38.3	24.9
60	15.5	11.2	11.4	10.4
80	59.9	51.8	44.4	46.6
100	27.2	26.2	23.5	23.6

Table 13: First Bin Residual Values - LFF

Quench Time (µs)	**First Bin Residual (#)**		
6	463	606	-199
10	688	652	470
20	457	458	-65
40	385	807	-439
70	610	612	559
100	608	524	382

Table 14: First Bin Residual Values - MFF

Quench Time (µs)	**First Bin Residual (#)**			
10	21788	23012	18431	21635
20	9457	9763	9134	9388
40	3004	3169	2908	2723
70	1508	1607	1660	1384
100	1011	1089	1086	1032

Table 15: First Bin Residual Values - HFF

Quench Time (µs)	**First Bin Residual (#)**			
25	-	-	-	-
30	26585	23960	22464	26266
40	22151	20632	19503	20828
60	15961	15279	17199	15824
80	15344	14929	14231	14893
100	12242	11798	12379	11921

APPENDIX E: Sensitivity Analysis

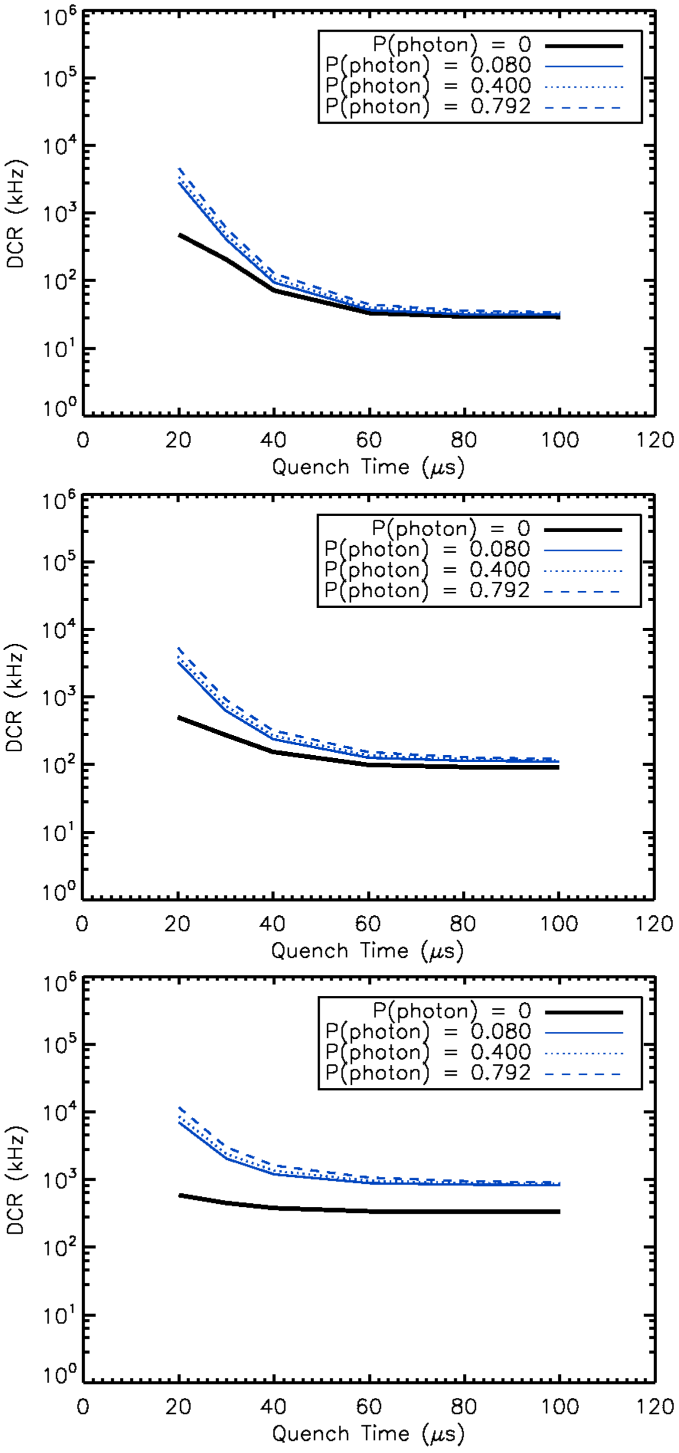
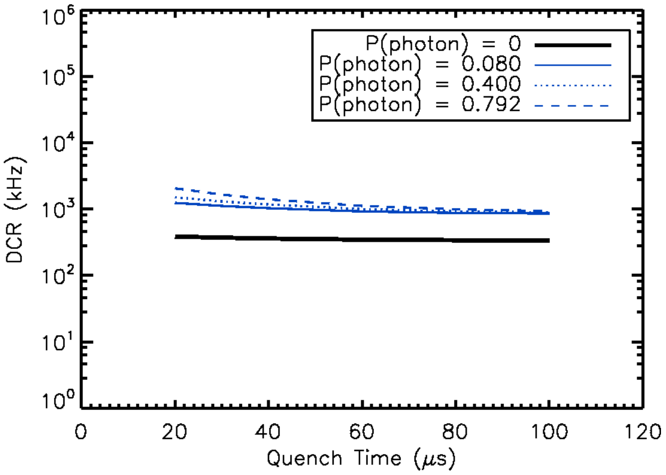
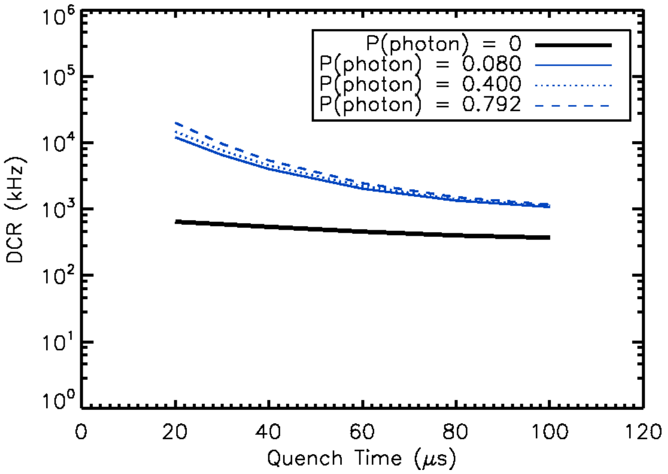


Figure 54 – Variations on DCR vs Quench Time with respect to Base Thermal DCR; 30 kHz (top), 100 kHz (middle), 500 kHz (bottom)

Table 16: Input Values for Sensitivity Analysis – Variations on Base Thermal DCR

Parameter	Value
Avalanche Initiation Probability	0.8
Thermal DCR	30 kHz, 100 kHz, 500 kHz
Number of Filled Traps	100
Detrapping Lifetime	10 µs
Carrier Lifetime	100 µs





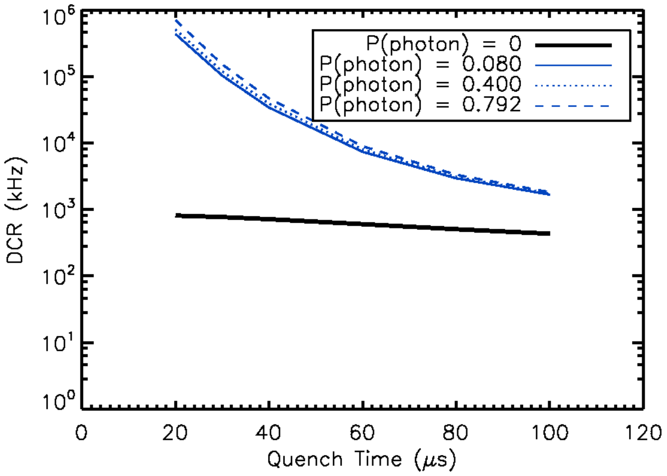
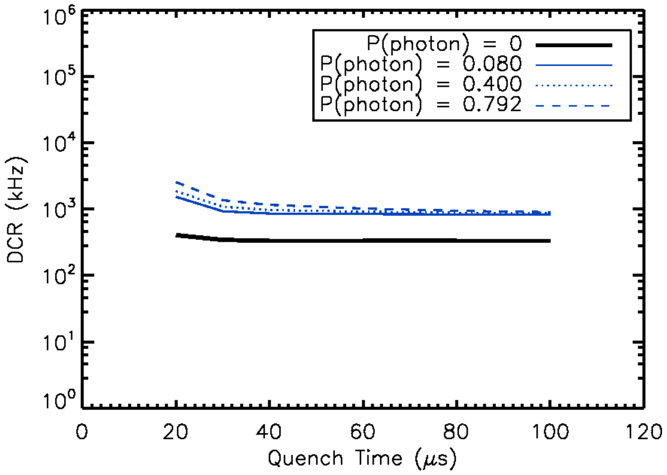


Figure 55 – Variations on DCR vs Quench Time with respect to Number of Filled Traps; 10 (top), 100 (middle), 300 (bottom)

Table 17: Input Values for Sensitivity Analysis – Variations on Number of Filled Traps

Parameter	Value
Avalanche Initiation Probability	0.8
Thermal DCR	500 kHz
Number of Filled Traps	10, 100, 300
Detrapping Lifetime	30 μs
Carrier Lifetime	100 μs



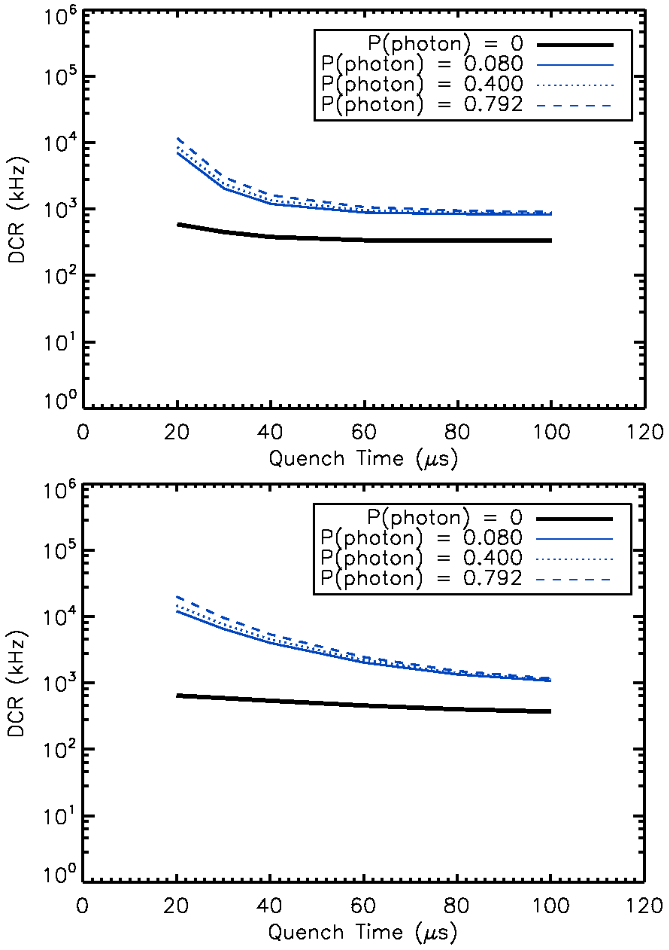


Figure 56 – Variations on DCR vs Quench Time with respect to Detrapping Lifetime; 5 µs (top), 10 µs (middle), 30 µs (bottom)

Table 18: Input Values for Sensitivity Analysis – Variations on Detrapping Lifetime

Parameter	**Value**
Avalanche Initiation Probability	0.8
Thermal DCR	500 kHz
Number of Filled Traps	100
Detrapping Lifetime	5 µs , 10 µs , 30 µs
Carrier Lifetime	100 µs

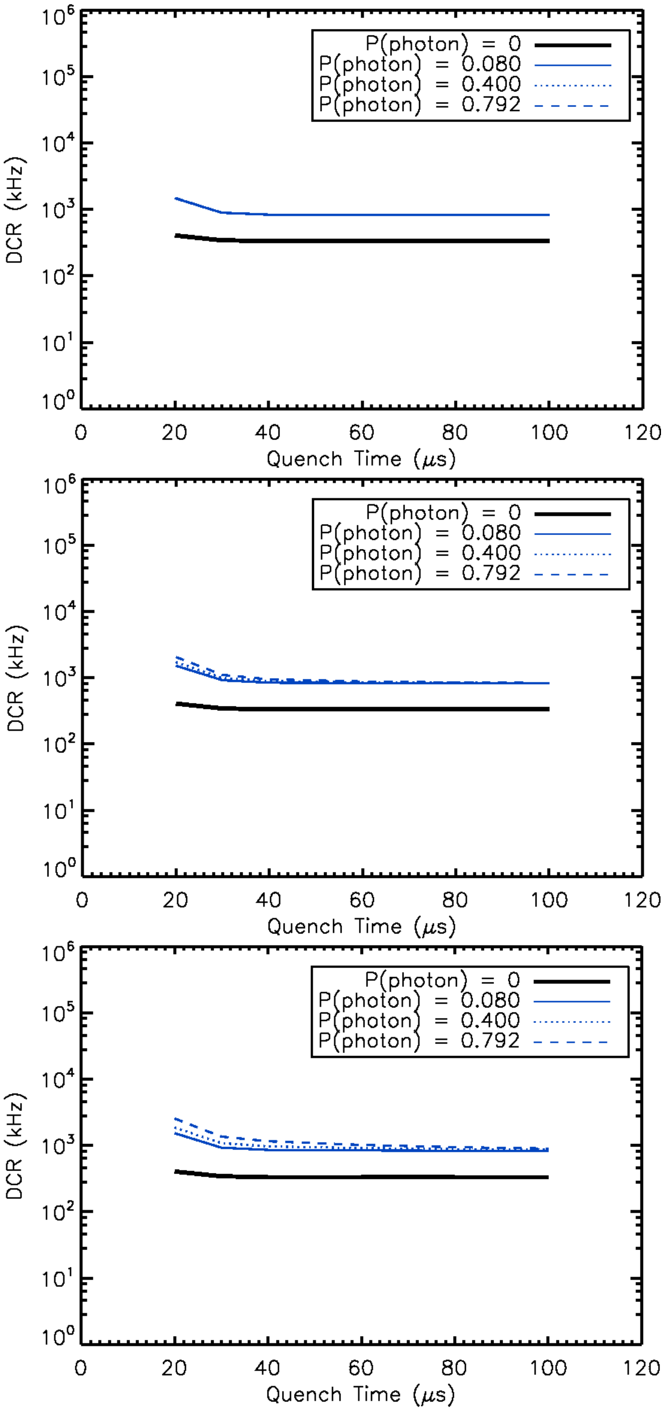


Figure 57 – Variations on DCR vs Quench Time with respect to Carrier Lifetime;10 µs (top), 50 µs (middle), 100 µs (bottom)

Table 19: Input Values for Sensitivity Analysis – Variations on Carrier Lifetime

Parameter	**Value**
Avalanche Initiation Probability	0.8
Thermal DCR	500 kHz
Number of Filled Traps	100
Detrapping Lifetime	50 μs
Carrier Lifetime	10 μs, 50 μs, 100 μs

Note that this appendix is not a comprehensive representation of the trends sampled in all possible combinations. For example, if the detrapping lifetime is very short, the number of traps will have less effect on the mean DCR. There are many higher-order interactions between the inputs, but the plots shown were chosen to demonstrate the trends in a variable space that lent itself to reasonable trend resolution.